Running head: IRT modeling of forced choice

Item response modeling of forced-choice questionnaires

Anna Brown
SHL Group

Alberto Maydeu-Olivares
University of Barcelona

Abstract

Multidimensional forced-choice formats can significantly reduce the impact of numerous response biases typically associated with rating scales. However, if scored with classical methodology these questionnaires produce ipsative data, which leads to distorted scale relationships and makes comparisons between individuals problematic. This research demonstrates how Item Response Theory (IRT) modeling may be applied to overcome these problems. A multidimensional IRT model based on Thurstone's framework for comparative data is introduced, which is suitable for use with any forced-choice questionnaire composed of items fitting the dominance response model, with any number of measured traits, and any block sizes (i.e. pairs, triplets, quads etc.). Thurstonian IRT models are normal ogive models with structured factor loadings, structured uniquenesses, and structured local dependencies. These models can be straightforwardly estimated using structural equation modeling (SEM) software Mplus. A number of simulation studies are performed to investigate how latent traits are recovered under various forced-choice designs, and to provide guidelines for optimal questionnaire design. An empirical application is given to illustrate how the model may be applied in practice. It is concluded that when the recommended design guidelines are met, scores estimated from forced-choice questionnaires with the proposed methodology reproduce the latent traits well.

Keywords: forced-choice format, forced-choice questionnaires, ipsative data, comparative judgment, multidimensional IRT

Item response modeling of forced-choice questionnaires

The most popular way of presenting questionnaire items is through rating scales (Likert scales), where participants are asked to rate a statement using some given categories (for example, ranging from "strongly disagree" to "strongly agree", or from "never" to "always", etc.). It is well-known that such format (*single-stimulus* format) can lead to various response biases, for instance because participants do not interpret the rating categories in the same way (Friedman & Amoo, 1999). One way of dealing with some simpler biases is modeling them post-completion (e.g. Maydeu-Olivares & Coffman, 2006), another way is to present questionnaire items in a comparative or *forced-choice* fashion.

Typical multidimensional forced-choice format (MFC) questionnaires consist of blocks of two or more statements from different dimensions, for example:

a. I manage to relax easily

b. I am careful over detail

c. I enjoy working with others

d. I set high personal standards

Instead of evaluating each statement in relation to a rating scale, respondents have to choose between statements according to the extent these statements describe their preferences or behavior. When there are 2 statements in a block, respondents are simply asked to select the statement that better describes them. For blocks of 3, 4 or more statements, respondents may be asked to rank-order the statements, or to select one statement which is "most like me" and one which is "least like me" (i.e., to provide a partial ranking).

Because it is impossible to endorse every item, the forced-choice format eliminates uniform biases such as acquiescence responding (Cheung & Chan, 2002), and can increase operational validity by reducing "halo" effects (Bartram, 2007). However, there are serious problems with the way the forced-choice questionnaires have been scored traditionally.

Typically, rank orders of items in a block are reversed and then added to their respective scales to make up a scale score. Therefore *relative* positions of items are treated as *absolute*. Not only is this model inadequate for describing the responses to forced-choice items (Meade, 2004); it also results in *ipsative* scale scores. The term *ipsative* (from the Latin *ipse*: he, himself) was first used by R.B. Cattell to name a type of scale where a score on one attribute is relative to scores on other attributes for this individual. It is easy to see that regardless of choices made, the same number of points is distributed between items in a block, and the total score on the questionnaire is therefore the same (constant) for everyone. Therefore, although trait scores will vary from individual to individual, their sum will remain constrained, making it impossible to score above or below average on *all* scales. Ipsative scores present a problem for score interpretation, and for almost every conventional type of psychometric analysis (for a discussion see Baron, 1996).

Given the potential advantages of the forced-choice format, and the disadvantages of the scoring methods associated with it, an alternative approach to modeling the comparative responses arising from forced-choice items is needed. Thurstone's (1927, 1931) model provides a powerful framework for modeling comparative data such as paired comparisons and rankings. Although typically used as a model to scale stimuli (items), Maydeu-Olivares and Brown (2010) showed how Thurstonian models can be used also to scale individuals. Indeed, when used in a respondent-centered formulation, Thurstonian models are Item Response Theory (IRT) models allowing estimation of the individual trait scores. The aim of this paper is to extend the application of the Thurstonian IRT modeling from single ranking tasks measuring one dimension described in Maydeu-Olivares and Brown (2010) to multiple ranking tasks, thus providing a modeling framework for forced-choice questionnaires measuring multiple dimensions. In what follows, we refer to "personality traits", or "traits", but of course the same approach applies to forced-choice questionnaires measuring

motivation, attitudes etc.

This article is structured into five sections. In the first section, we describe how to code responses to forced-choice items using binary outcome variables. Section two describes how to apply the Thurstonian factor model (see Maydeu-Olivares & Böckenholt, 2005) to forced-choice questionnaires. This model relates the binary outcomes obtained from the comparative responses to the unobserved utilities of items (first-order factors), and the utilities in turn are related to a set of underlying personality traits (second-order factors). Thus, the Thurstonian factor model is a second-order factor model with binary indicators. In section three, we introduce the Thurstonian IRT model. This is simply a Thurstonian factor model reparameterized as a first order factor model. As we shall discuss, the reparameterization is needed to obtain latent trait estimates from forced-choice tests. Simulation studies are presented in section four to illustrate the model properties, as well as item parameter recovery and latent trait recovery. Section five includes a real-data application, where a forced-choice test measuring the Big Five is estimated and participants are scored using the Thurstonian IRT model. The popular statistical modeling program Mplus (Muthén & Muthén, 1998-2007) is used in all analyses presented in this paper. We conclude with a discussion of the main points of the work presented here.

Binary Coding Of Forced-Choice Response Data

This section describes how to code responses to forced-choice blocks using binary outcome variables, one for each pairwise comparison between the items within a block. This is the standard procedure to code comparative data (see Maydeu-Olivares & Böckenholt, 2005), but here we apply it specifically to forced-choice questionnaire blocks.

In a forced-choice block, a respondent is asked to assign ranks to n items according to the extent the items describe the respondent's personality. For instance, for $n = 4$ items {A, B, C, D}, the respondent has to assign ranking positions – numbers from 1 (most preferred)

to 4 (least preferred).

	Ranking
Item A	_
Item B	_
Item C	_
Item D	_

Alternatively, the respondent might be asked to indicate only two items: one item that most accurately describes their personality, and one item that describes it least accurately. It is easy to see that this format type provides an incomplete ranking, because it only assigns the first and the last ranks.

	Most like me	Least like me
Item A	□	□
Item B	□	□
Item C	□	□
Item D	□	□

Any ranking of n items can be coded equivalently using $\tilde{n} = \frac{n(n-1)}{2}$ binary outcome variables. In a block of 2 items {A, B}, there is only one comparison to be made between items A and B. In a block of 3 items {A, B, C}, there are 3 pairwise comparisons: between items A and B, between A and C, and between B and C. In a block of 4 items {A, B, C, D}, there are 6 comparisons to be made between items: A is compared with B, C and D; B is compared with C and D; and C is compared with D.

In each pair, either the first item is preferred to the second, or otherwise. Thus, observed responses to the pairwise comparisons can be coded as *binary outcomes*:

$$y_{_l} = \begin{cases} 1 & \text{if} \quad \text{item } i \text{ is preferred over item } k \\ 0 & \text{if} \quad \text{item } k \text{ is preferred over item } i \end{cases} \tag{1}$$

where l indicates the pair $\{i,k\}$. For example, the ordering {A, D, B, C} can be coded as:

Ranking				Binary Outcomes					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
1	3	4	2	1	1	1	1	0	0

In the case of partial rankings, such as ones observed using the "most like me" – "least like me" format when $n > 3$, the information for some binary outcomes is missing by design. For instance, when items are presented in blocks of $n = 4$ items the outcome of the comparison between the two items that are not selected either as "most" or "least" is unknown. Following the previous example, the resulting partial ranking can be coded as follows:

Partial ranking				Binary Outcomes					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
most		least		1	1	1	1	.	0

One consequence of dealing with blocks of statements (each of which is a ranking task) is that responses made within one block are always *transitive.* For example, if the respondent rank-orders A above B, and B above C, it automatically follows that A is ranked before C and therefore the outcome of for the {A,C} pair can be deducted from pairs {A,B} and {A,C}. It then follows that only $n!$ different binary patterns may be observed for a block of n items.

Applying A Thurstonian Factor Model To Forced-Choice Questionnaires

Response Model For Rankings

Arguably, the best known model for describing comparative choices, such as ones made in forced-choice blocks, is Thurstone's Law of Comparative Judgment. Although Thurstone (1927) focused initially on paired comparisons, he recognized later (Thurstone, 1931) that many other types of choice data, including rankings, could be modeled in a similar way. He argued that in a comparative task, 1) each item elicits a utility as a result of

a *discriminal process*; 2) respondents choose the item with the largest utility value at the moment of comparison; and 3) the utility is an unobserved (continuous) variable and is normally distributed in the population of respondents.

According to Thurstone's model, each of the n items to be ranked elicits a utility. We shall denote by t_i the latent utility associated with item i. Therefore, there are exactly n such latent variables when modeling n items. A respondent prefers item i over item k if his/her latent utility for item i is larger than for item k, and consequently ranks item i before item k. Otherwise, he/she ranks item k before item i. The former outcome is coded as "1" and the latter as "0". That is,

$$y_l = \begin{cases} 1 & \text{if} \quad t_i \geq t_k \\ 0 & \text{if} \quad t_i < t_k \end{cases}, \tag{2}$$

where the equality sign is arbitrary as the latent utilities are assumed to be continuous and thus by definition two latent variables can never take on exactly the same value.

The response process can be alternatively described by computing differences between the latent utilities. Let

$$y_l^* = t_i - t_k \tag{3}$$

be a continuous variable that represents the difference between utilities of items i and k. Because t_i. and t_k are not observed, y_l^* is also unobserved. Then, the relationship between the observed comparative response y_l and the latent comparative response y_l^* is

$$y_l = \begin{cases} 1 & \text{if} \quad y_l^* \geq 0 \\ 0 & \text{if} \quad y_l^* < 0 \end{cases}. \tag{4}$$

Note that the difference of utilities determines the preference response, i.e. there is no error term in Equation (3). This is because here we consider ranking tasks, for which responses are transitive (Maydeu-Olivares & Bockenholt, 2005).

It is convenient to present the response model in matrix form. Let $\mathbf{t}$ be the $n \times 1$ vector of latent utilities and $\mathbf{y}^*$ be the $\tilde{n} \times 1$ vector of latent difference responses, where $\tilde{n} = \frac{n(n-1)}{2}$. Then we can write the set of $\tilde{n}$ equations (3) as

$$\mathbf{y}^* = \mathbf{A}\,\mathbf{t}\,, \tag{5}$$

where $\mathbf{A}$ is a $\tilde{n} \times n$ design matrix. Each column of $\mathbf{A}$ corresponds to one of the n items, and each row of $\mathbf{A}$ corresponds to one of the $\tilde{n}$ pair-wise comparisons. For example, when $n = 2$, $\mathbf{A} = \begin{pmatrix} 1 & -1 \end{pmatrix}$, whereas when $n = 3$, and $n = 4$

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

respectively. For instance, in the design matrix for $n = 3$ items, each column corresponds to one of the 3 items {A, B, C}. Rows represent 3 possible pair-wise comparisons. Row 1 corresponds to the comparison between A and B, and row 3 to the comparison between B and C.

Moving from one forced-choice block to multiple blocks, we let p be the number of blocks, n the number of items per block, and the total number of items therefore is $p \times n = m$. In this case, the design matrix will consist of m columns corresponding to all items in the questionnaire, and $p \times \tilde{n}$ rows corresponding to the $\tilde{n}$ pair-wise comparisons made in all of p blocks. The design matrix $\mathbf{A}$ is then partitioned in correspondence to the blocks. For instance, for a questionnaire with $p = 3$ blocks of $n = 3$ items in each block (9 items in total), the design matrix $\mathbf{A}$ is:

$$\mathbf{A} = \left(\begin{array}{ccc|ccc|ccc} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{array}\right).$$

The Thurstonian Factor Model

Because questionnaire items are designed to measure some psychological constructs (personality traits, motivation factors, attitudes etc.), a set of d common factors (latent traits) is introduced into the model. It is then assumed that the latent utilities $\mathbf{t}$ are a linear function of the traits, that is,

$$\mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,. \tag{6}$$

In Equation (6), $\boldsymbol{\mu}_t$ contains m means of the latent utilities $\mathbf{t}$, $\boldsymbol{\Lambda}$ is an $m \times d$ matrix of factor loadings, $\boldsymbol{\eta}$ is an d-dimensional vector of common factors (latent traits in IRT terminology), and $\boldsymbol{\varepsilon}$ is an m-dimensional vector of unique factors. We will assume that every item measures one trait only, i.e., $\boldsymbol{\Lambda}$. is an independent clusters solution. As in a standard factor analytic model, the latent traits are uncorrelated with the unique factors and that their means are zero. The latent traits are freely correlated (their covariance matrix is $\boldsymbol{\Phi}$) but their variances are fixed to one for identification. The unique factors are uncorrelated, so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal. In addition, it is assumed that latent traits and unique factors are normally distributed.

This linear model describes a dominance response process. As such, it is suitable for modeling items written such that the relationship between the utility of an item and the latent trait the item measures *monotonically* increases for positively keyed items or

monotonically decreases for negatively keyed items. Examples of positively and negatively keyed dominance items are: "I keep my paperwork in order" and "I struggle to organize my paperwork", respectively. For any two respondents, the utility for the first item will be higher for the individual whose score on Conscientiousness is higher, and this will be reversed for the second item. This is in contrast to *ideal-point* models (Coombs, 1964), where the utility for an item has a peak at a certain level of the latent trait, and decreases in either direction of the latent trait from that point. An example of an item for which an ideal-point model would be more suitable is: "My attention to detail is about average". Clearly, utility for this item would be high for respondents with an average score on Conscientiousness, and be lower for respondents with very high or very low scores.

Dominance items are by far more prevalent in existing personality questionnaires, either using the Likert scale or the forced-choice formats (Stark, Chernyshenko, Drasgow & Williams, 2006). Given the absence of an adequate forced-choice model using this popular type of items, we chose to deal only with the dominance items in this paper. There are alternative IRT approaches to creating and scoring forced-choice questionnaires relying on an ideal-point response process (McCloy, Heggestad & Reeve, 2005; Stark, Chernyshenko & Drasgow, 2005), and we refer readers to those papers for modeling forced-choice tests with ideal-point items. Models throughout this paper assume that only dominance items are used.

To illustrate how binary outcomes, their underlying utilities and traits are modeled, in Figure 1 a Thurstonian factor model is depicted for a very short forced-choice questionnaire. The questionnaire measures $d = 3$ correlated traits; each trait is measured by 3 items. The nine questionnaire items ($m = 9$) are presented in triplets (blocks of $n = 3$ items) so that there are no two items within a block measuring the same trait. There are $p = 3$ such blocks in this simple example. Trait 1 is measured by items 1, 4, and 7; trait 2 is measured by items 2, 5, and 8; and trait 3 is measured by items 3, 6, and 9. Respondents are

asked to rank-order the items within each block. The resulting rankings are transformed into 3 binary outcomes per block (9 outcomes in total), which are modeled as differences of underlying utilities using Equation (5). Because each binary outcome is the result of comparing two items, it depends on two latent utilities. Utilities, in turn, are functions of the 3 personality traits. The 9 binary outcomes are measured without error (because responses to ranking blocks are transitive). However, the 9 utilities have uniqueness parameters to be estimated, i.e. the variance of the unique factors in each item's utility.

Insert Figure 1 about here

Thurstonian IRT Model For Forced-Choice Questionnaires

In this section we show how the Thurstonian factor model, which is a second-order factor model for binary data with some special features, can be equivalently expressed as a first-order model, again, with some special features. To distinguish both models, we refer to the first-order model as to the Thurstonian IRT model. We provide the item characteristic and information function for the latter and discuss item parameter estimation, latent trait estimation, and reliability estimation.

Reparameterized Model

There are several reasons for reparameterizing the Thurstonian factor model for forced-choice presented above as a first-order model. First, in psychometric testing applications the first-order factors (the latent utilities) are not of interest. Rather, interest lies in estimating the second-order factors (the latent traits). Second, the use of the Thurstonian IRT model instead of the Thurstonian factor model speeds up computations considerably in the case of large models. Third, and most importantly, since the residual error variances of the latent response variables $\mathbf{y}^*$ are zero in the Thurstonian factor model (see Figure 1), latent trait estimates cannot be computed (see Maydeu-Olivares, 1999;

Maydeu-Olivares & Brown, 2010). When the model is reparameterized as a first order model, the residual error variances of the latent response variables are no longer zero (see Figure 2), enabling latent trait estimation. In addition, the reparameterization provides some valuable insights into the characteristics of the model, and facilitates the formulation of such important descriptors of any IRT model as item characteristic functions and information functions.

The reparameterization involves writing the second-order factor model obtained from Equations (5) and (6)

$$\mathbf{y}^* = \mathbf{A}\left(\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\right) = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon}\,, \tag{7}$$

as the first-order model

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \breve{\boldsymbol{\Lambda}}\boldsymbol{\eta} + \breve{\boldsymbol{\varepsilon}}\,. \tag{8}$$

The reparameterized model (8) involves

a) a structured $(p \times \tilde{n}) \times d$ matrix of factor loadings

$$\breve{\boldsymbol{\Lambda}} = \mathbf{A}\boldsymbol{\Lambda}\,, \tag{9}$$

b) a structured $(p \times \tilde{n}) \times (p \times \tilde{n})$ covariance matrix of the unique pairwise errors $\breve{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon}$ with $\mathrm{cov}\left(\breve{\boldsymbol{\varepsilon}}\right) = \breve{\boldsymbol{\Psi}}^2$, where

$$\breve{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}'\,, \tag{10}$$

c) an unrestricted $(p \times \tilde{n}) \times 1$ vector of thresholds

$$\boldsymbol{\gamma} = -\mathbf{A}\boldsymbol{\mu}_t\,. \tag{11}$$

That is, we do not impose the restriction (11) on $\boldsymbol{\gamma}$. This is because in IRT applications the means $\boldsymbol{\mu}_t$ of the latent utilities are not of interest. We will therefore be estimating an unrestricted vector of thresholds $\boldsymbol{\gamma}$ leading to a considerably less constrained model.

To illustrate the structure imposed by the model on the matrices $\breve{\boldsymbol{\Lambda}}$ and $\breve{\boldsymbol{\Psi}}^2$ we

return to our previous example of a very short forced-choice questionnaire measuring $d = 3$ latent traits with $p = 3$ blocks of $n = 3$ items. For this example,

$$\breve{\mathbf{\Lambda}} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \hline \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \hline \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \end{pmatrix}, \tag{12}$$

whereas

$$\breve{\mathbf{\Psi}}^2 = \left(\begin{array}{ccc:ccc:ccc} \psi_1^2+\psi_2^2 & & & & & & & & \\ \psi_1^2 & \psi_1^2+\psi_3^2 & & & & & & & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2+\psi_3^2 & & & & & & \\ \hdashline 0 & 0 & 0 & \psi_4^2+\psi_5^2 & & & & & \\ 0 & 0 & 0 & \psi_4^2 & \psi_4^2+\psi_6^2 & & & & \\ 0 & 0 & 0 & -\psi_5^2 & \psi_6^2 & \psi_5^2+\psi_6^2 & & & \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & \psi_7^2+\psi_8^2 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_7^2 & \psi_7^2+\psi_9^2 & \\ 0 & 0 & 0 & 0 & 0 & 0 & -\psi_8^2 & \psi_9^2 & \psi_8^2+\psi_9^2 \end{array}\right) \tag{13}$$

It can be seen in (12) that pairs involving the same item share a factor loading of the same magnitude on the respective trait. As for $\breve{\mathbf{\Psi}}^2$, it is not a block diagonal matrix: unique errors of pairs related to the same item within a block are correlated. The pattern of within-block covariances does not depend on the number of latent traits, but on the number of items in the block. Importantly, not all uniquenesses are inter-correlated, but only those related to the same items. This becomes apparent in blocks of 4 or more items, where, for example, unique error of binary outcome {A, B} is not correlated with the error of outcome {C, D}.

Figure 2 shows the reparameterized model for our example with 3 traits and 9 binary outcomes. There are no latent utilities, and the traits are linked directly to the latent

response variables underlying the binary outcomes. Because, by construction, each binary outcome is the result of comparing two items from different dimensions and because each item is assumed to measure only one trait, the model implies that each binary outcome depends only on two traits. This is true regardless of the number of items per block, the number of blocks, or the number of latent traits involved in any given forced-choice test.

Insert Figure 2 about here

It can be seen that there are 9 binary outcomes in Figure 2, each depending on two traits; therefore 18 factor loadings are involved. However, 9 constraints are imposed on these factor loadings. For example, the loading involving binary outcome {i2, i3} on trait 2 is constrained to be equal to the loading of outcome {i1, i2} on trait 2, but with the signs reversed. These two loadings are of opposite signs because item 2 is the first element of the directional pair {i2, i3} whereas it is the second element of the directional pair {i1, i2}.

Furthermore, the residual errors of the latent response variables $\mathbf{y}^*$ are structured. The residual error variance associated with a binary outcome equals the sum of residual error variances of utilities of the two items involved in the pair. The residual errors of latent response variables involving the same item are correlated. For instance, there are correlated errors between latent response variables {i1, i2} and {i1, i3} because these are pairs obtained by comparing item 1 to other items in the block. Both of these outcomes will be influenced by the unique factors of the utility of item 1, sharing common variance that is not accounted for by the latent trait.

To summarize, for multidimensional forced-choice questionnaires measuring d traits using p blocks of n items each, the model presented here involves d first-order common factors (the latent traits) and $p \times \tilde{n}$ binary outcomes, and each binary outcome depends on two traits. In contrast, when expressed as a second-order Thurstonian factor model, it

involves $m = p \times n$ first-order factors (the utilities) and d second-order factors (the latent traits).

Identification Of Thurstonian IRT Models For Forced-Choice Questionnaires

The reparameterized IRT model is algebraically equivalent to the Thurstonian factor model, thus yielding the same number of parameters, and requiring exactly the same identification constraints. For a single ranking task, Maydeu-Olivares and Böckenholt (2005) suggested the following constraints to identify the model: (a) fixing all factor loadings involving (arbitrarily) the last item to 0 ($\lambda_{ni} = 0$ for all $i = 1, \ldots, d$); and (b) fixing the unique variance of the last item to 1, $\psi_n^2 = 1$. These identification constraints are needed to set the scale origin for factor loadings and for the uniquenesses because of the comparative nature of the data. To set the scale for the latent traits, their variances are simply set equal to one.

In the case of a multidimensional model involving several blocks of items each measuring a single trait (i.e. forced-choice questionnaire model), the identification constraints are simpler than in the case of a single block. The model is identified simply by imposing a constraint among the uniquenesses within each block. Arbitrarily, we suggest fixing the uniqueness of the last item in each block to 1. For example, to identify the Thurstonian IRT model depicted in Figure 2, we impose the following constraints $\psi_3^2 = 1$, $\psi_6^2 = 1$, and $\psi_9^2 = 1$.

This general identification rule is valid in all but two special cases: a) when $n = 2$ and $d > 2$ (i.e., items presented in pairs measuring more than 2 traits), and b) when $d = n = 2$ (only two traits are measured using pairs of items). In Case a), no item uniqueness ψ_i^2 can be identified. They can be set equal to 0.5, so that $\breve{\psi}_l^2 = \psi_i^2 + \psi_k^2 = 1$. Case a) is discussed in more detail in Appendix A. Regarding Case b), all item uniquenesses need to be fixed as in the case above. In addition, each binary outcome will depend on both traits involved, the factor loading matrix contains no zero elements, and the model is essentially an

exploratory factor model. To avoid the indeterminacy problem in this case, it is sufficient to fix the 2 factor loadings of the first pair. For a model with 3 or more traits, no such constraints are needed because there are sufficient numbers of zero elements in each column and row of the factor loading matrix.

Item Characteristic Function

Since the latent traits $\boldsymbol{\eta}$ and the unique factors $\boldsymbol{\varepsilon}$ are normally distributed, the latent response variables $\mathbf{y}^*$ are also normally distributed, and the item characteristic function (ICF) is that of a normal ogive model with some special features. Indeed, it follows from (4) and (8) that the conditional probability of preferring item i over item k is

$$\Pr\left(y_l = 1 \middle| \boldsymbol{\eta}\right) = \Phi\left(\frac{-\gamma_l + \breve{\boldsymbol{\lambda}}_l{}'\boldsymbol{\eta}}{\sqrt{\breve{\psi}_l^2}}\right), \tag{14}$$

where $\Phi\left(x\right)$ denotes the cumulative standard normal distribution function evaluated at x, γ_l is the threshold for binary outcome y_l, $\breve{\boldsymbol{\lambda}}_l{}'$ is the $1 \times d$ vector of factor loadings, and $\breve{\psi}_l^2 = \psi_i^2 + \psi_k^2$ is the uniqueness of the latent response variable y_l^*. Because we assume that each item only measures one trait ($\boldsymbol{\Lambda}$ is an independent clusters solution), each binary outcome only depends on two traits. As a result, the ICF for the binary outcome variable y_l, which is the result of comparing item i measuring trait η_a and item k measuring trait η_b, is

$$\Pr\left(y_l = 1 \middle| \eta_a, \eta_b\right) = \Phi\left(\frac{-\gamma_l + \lambda_i\eta_a - \lambda_k\eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right). \tag{15}$$

Equation (15) describes the item characteristic function using a threshold/loading parameterization. This is simply a standard two-dimensional normal ogive IRT model for binary data except that (a) factor loadings are structured so that every binary outcome y_l involving the same item will share the same factor loading, (b) uniquenesses of latent response variables are structured so that they equal the sum of uniquenesses of the 2 items

involved, and (c) the item characteristic functions are not independent (local independence conditional on the latent traits does not hold). Rather, there are patterned covariances among the errors $\breve{\varepsilon}$ – see Equation (13).

Now, letting

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_k^2}}, \qquad \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_k^2}} \quad , \qquad \beta_k = \frac{\lambda_k}{\sqrt{\psi_i^2 + \psi_k^2}}, \tag{16}$$

the item characteristic function (15) can be written in an intercept/slope form as

$$\Pr\left(y_l = 1 \middle| \eta_a, \eta_b\right) = \Phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right). \tag{17}$$

Information functions can be more easily expressed as a function of intercepts ($\boldsymbol{\alpha}$) and slopes ($\boldsymbol{\beta}$) than as a function of thresholds ($\boldsymbol{\gamma}$), factor loadings ($\boldsymbol{\lambda}$), and uniquenesses ($\boldsymbol{\psi}^2$), and we do so in this paper. However, the reader must bear in mind that the intercepts and slopes are not mathematically independent parameters (except in the case of $n = 2$). Rather, they are functions of the smaller set of thresholds, factor loadings, and uniquenesses parameters.

Estimation Of Thurstonian IRT Models For Forced-Choice Questionnaires

IRT models are most often estimated using full information maximum likelihood (FIML). For models describing forced-choice questionnaires such estimation is not feasible due to the presence of local dependencies (when $n > 2$). However, the Thurstonian IRT model described here can be straightforwardly estimated using limited information methods. First, the sample thresholds and tetrachoric correlations are estimated. Then, the model parameters are estimated from the first stage estimates by unweighted least squares (ULS) or diagonally weighted least squares (DWLS). In practice, differences between using ULS or DWLS in the second stage of the estimation procedure are negligible (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009). All models in this paper are estimated with M*plus* using the DWLS estimator with mean corrected Satorra-Bentler goodness-of-fit tests. Note that

this estimation procedure is denoted as WLSM estimation in M*plus*.

When the number of items per block is larger than 2, a correction to degrees of freedom is needed when testing model fit. This is because for a ranking block there are $r = n(n-1)(n-2)/6$ redundancies among the thresholds and tetrachoric correlations estimated from the binary outcome variables (Maydeu-Olivares, 1999). For instance, there is $r = 1$ redundancy in every block of 3 items, and there are $r = 4$ redundancies in every block of 4 items. With p ranking blocks in the questionnaire, the number of redundancies is $p \times r$. Thus, when $n > 2$, one needs to subtract $p \times r$ from the degrees of freedom given by the modeling program to obtain the correct p-value for the test of exact fit. Goodness-of-fit indices involving degrees of freedom in their formula, such as the RMSEA, also need to be recomputed using the correct number of degrees of freedom. When $n = 2$, no degrees of freedom adjustment is needed, the p-value and RMSEA printed by the program are correct.

Latent Trait Estimation

Once the IRT model parameters have been estimated, scores on the latent traits for individuals can be estimated using their pattern of binary outcome responses. There are three popular procedures for latent trait estimation: maximum likelihood (ML), expected a posteriori (EAP), and maximum a posteriori (MAP) estimation (Embretson & Reise, 2000). Our focus will be on the MAP estimator, which maximizes the mode of the posterior distribution of the latent traits, as it is the method implemented in M*plus*, the software used throughout this paper. The posterior distribution is obtained by multiplying the joint likelihood of the binary outcome responses by the density of the population distribution, which is standard multivariate normal in the model considered here. The MAP estimator exists for all response patterns, is more efficient than the ML estimator when a small number of items is involved (and in personality questionnaires the number of items per trait is generally small), but is known to produce estimates biased towards the population mean (see

Embretson & Reise, 2000, p. 174).

To evaluate the joint likelihood of the binary outcomes pattern, it is assumed that the binary outcomes are independent given the latent traits. We know, however, that in Thurstonian IRT models structured dependencies exist between the error terms within blocks of 3 or more items. Effects of ignoring these dependencies on the latent trait estimates have been shown to be negligible in applications involving a single ranking task (Maydeu-Olivares & Brown, 2010), and they are likely to be even smaller in forced-choice questionnaires where blocks are smaller and there are fewer local dependencies per item. Throughout this paper we will use the simplifying assumption that the item characteristic functions for the binary outcomes are locally independent. Note that this simplifying assumption is only employed for latent trait estimation, not for item parameter estimation.

Information Functions and Reliability Estimation

In IRT, unlike in classical scoring, the precision of measurement depends on the latent traits and therefore is not the same for all respondents. The precision of measurement is provided by the test information function $\mathcal{I}\left(\boldsymbol{\eta}\right)$, which is computed from item information functions $\mathcal{I}_l\left(\boldsymbol{\eta}\right)$. Recall that in forced-choice questionnaires, the "items" referred to when describing item information are the binary outcomes of pairwise comparisons between the questionnaire items.

The item information function is computed in a manner similar to its one-dimensional IRT counterpart, except that since each binary outcome depends on two dimensions, the direction of the information must be also considered (Reckase, 2009; Ackerman, 2005). Let $P_l\left(\boldsymbol{\eta}\right) = \Pr\left(y_l = 1 \middle| \eta_a, \eta_b\right)$, $\boldsymbol{\alpha}$ be a vector of angles to all d axes that defines the direction from a point $\boldsymbol{\eta}$, and $\nabla_{\boldsymbol{\alpha}} P_l\left(\boldsymbol{\eta}\right)$ be the gradient (directional derivative) in direction $\boldsymbol{\alpha}$, which is given by (Reckase, 2009):

$$\nabla_{\alpha} P_l(\boldsymbol{\eta}) = \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_1}\cos\alpha_1 + \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_2}\cos\alpha_2 + \ldots + \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_d}\cos\alpha_d . \tag{18}$$

Then, the definition of item information in the multidimensional case is generalized to accommodate the change in slope with direction taken from a point in the latent trait space:

$$\mathcal{I}_{l}^{\alpha}(\boldsymbol{\eta}) = \frac{\left[\nabla_{\alpha} P_l(\boldsymbol{\eta})\right]^2}{P_l(\boldsymbol{\eta})\left[1 - P_l(\boldsymbol{\eta})\right]}. \tag{19}$$

Because each binary outcome depends on 2 latent traits, in the above expression directional derivatives for all but the 2 relevant dimensions will be 0. Therefore, for each binary outcome we consider two directions of information: one along the axis η_a, and another along the axis η_b. When computing the information in direction η_a, the angle to η_a is 0° and therefore $\cos(\alpha_a) = 1$, whereas the angle to η_b is determined by the correlation between η_a and η_b so that $\cos(\alpha_b) = \text{corr}(\eta_a, \eta_b)$ – see Bock (1975).

Using the intercept/slope parameterization of Equation (17), the directional derivatives with respect to η_a and η_b are simply

$$\frac{\partial P_l(\eta_a, \eta_b)}{\partial \eta_a} = \beta_i \phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right) \quad \text{and} \quad \frac{\partial P_l(\eta_a, \eta_b)}{\partial \eta_b} = -\beta_k \phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right), \tag{20}$$

where $\phi(z)$ denotes a standard normal density function evaluated at z (McDonald, 1999, p. 284). It follows from (19) and (20) that the information provided by one binary outcome about traits η_a and η_b are, respectively:

$$\mathcal{I}_{l}^{a}(\eta_a, \eta_b) = \frac{\left[\beta_i - \beta_k \text{corr}(\eta_a, \eta_b)\right]^2 \left[\phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right)\right]^2}{P_l(\eta_a, \eta_b)\left[1 - P_l(\eta_a, \eta_b)\right]}, \tag{21}$$

$$\mathcal{I}_{l}^{b}(\eta_a, \eta_b) = \frac{\left[-\beta_k + \beta_i \text{corr}(\eta_a, \eta_b)\right]^2 \left[\phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right)\right]^2}{P_l(\eta_a, \eta_b)\left[1 - P_l(\eta_a, \eta_b)\right]}. \tag{22}$$

Equations (21) and (22) describe the Item Information Surfaces (IIS). Figures 3(a) and 3(b) display the Item Characteristic Surface (ICS) for a binary outcome, along with its IISs.

Insert Figure 3 about here

Now, Equations (21) and (22) show that for binary outcomes involving uncorrelated traits, only the derivative in the direction of the trait itself contributes to the information. However, for binary outcomes involving correlated traits, derivatives in directions of both traits involved will contribute. For positively keyed items, binary outcomes involving *positively* correlated traits will provide *less information* than if the traits were orthogonal. And, for positively keyed items, binary outcomes involving *negatively* correlated traits will provide *more information* than if the traits were orthogonal. These properties, as we will see, have important implications for test design.

Assuming local independence, the total information about trait η_a is a sum of all information functions from binary outcomes independently contributing to the measurement of this trait:

$$\mathcal{I}^a\left(\boldsymbol{\eta}\right) = \sum_l \mathcal{I}_l^a\left(\boldsymbol{\eta}\right). \tag{23}$$

All the above applies to IRT scores estimated by the maximum likelihood method (ML). When Bayes MAP estimation of the latent traits is used, the information given by the prior distribution is added to the ML test information yielding the posterior test information $\mathcal{I}_P\left(\boldsymbol{\eta}\right)$ (see Du Toit, 2003). In the Thurstonian IRT model, since the latent traits are assumed to be normally distributed,

$$\mathcal{I}_P^a\left(\boldsymbol{\eta}\right) = \mathcal{I}^a\left(\boldsymbol{\eta}\right) - \frac{\partial^2 \ln\left(\phi\left(\boldsymbol{\eta}\right)\right)}{\partial^2 \eta_a} = \mathcal{I}^a\left(\boldsymbol{\eta}\right) + \varpi_a^a, \tag{24}$$

where ϖ_a^a is the diagonal element of the inverted latent trait covariance matrix $\boldsymbol{\Phi}^{-1}$ related to the dimension of interest, η_a (see Appendix B for a proof). The standard error of the MAP-estimated score $\hat{\eta}_a$ is the reciprocal of the square root of the posterior test information

(in direction of the trait η_a),

$$SE\left(\hat{\boldsymbol{\eta}}_a\right) = \frac{1}{\sqrt{\mathcal{I}_P^a\left(\boldsymbol{\eta}\right)}}. \tag{25}$$

The precision of measurement in IRT, as we can see, is indeed a function of the latent trait and therefore varies for each respondent. Nevertheless, providing a summary index of the precision of measurement can be useful, particularly for comparison with classical test statistics, and also for predicting expected levels of recovery of the true latent trait. After the trait scores have been estimated for a sample, these scores are used as empirical values at which the test information function is evaluated, and the standard errors are computed. The reliability index based on the estimated scores for a sample is referred to as *empirical* reliability (Du Toit, 2003), and is obtained by computing the observed score variance and the error variance for the sample. Importantly, estimates of empirical reliability depend on the method by which scores were computed.

When IRT scores are obtained by the MAP method, the posterior test information is evaluated at the point MAP estimates $(\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_d)$ for each respondent j in a sample of size N, and the squared standard errors are computed as the reciprocal of the test information. To compute the sample error variance (related to the precision of measurement of trait η_a), the squared standard errors (reciprocals of the posterior test information) are averaged across the sample

$$\bar{\sigma}^2_{error}\left(\hat{\boldsymbol{\eta}}\right) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\mathcal{I}_P^a\left(\hat{\boldsymbol{\eta}}_j\right)}. \tag{26}$$

Since the observed score variance is known for the sample (it is simply the variance of the MAP scores), the true score can be computed as the observed score variance minus error variance. Therefore, the empirical reliability for the MAP estimated scores is computed as follows (Du Toit, 2003):

$$\rho = \frac{\sigma^2 - \overline{\sigma}^2_{error}}{\sigma^2}\,, \tag{27}$$

where σ^2 is estimated using the sample variance of the estimated MAP scores, and $\overline{\sigma}^2_{error}$ is estimated using (26). Finally, the correlation between the true latent trait and the estimated latent trait can be estimated as follows:

$$\text{corr}\left(\eta_a, \hat{\eta}_a\right) = \sqrt{\rho}\,. \tag{28}$$

It is important to emphasize that because the classical concept of test reliability has no direct correspondence in IRT, any estimate of reliability obtained from the test information is only an approximation. Strictly speaking, the reliability will vary for different levels of the latent trait. Reliability estimates would be more accurate and more descriptive of the sample as a whole when the test information function is relatively uniform.

Simulation Studies

In this section we report a number of simulation studies performed to investigate how well item parameters and latent trait scores can be recovered. This is necessary to provide benchmarks to which similar real-world applications can be compared. First, we consider an extremely simplified questionnaire with 2 traits measured by item pairs. This low-dimensionality example provides an opportunity to look at the graphical illustrations of ICS and test information functions. Most importantly, it provides a benchmark for the precision of the latent trait estimation when no local dependencies exist. Then, a more realistic model will be considered measuring five traits. For this model, we manipulate the block size, i.e. blocks of 2, 3 and 4 items will be considered.

Simulation 1. A forced-choice questionnaire measuring 2 traits using item pairs

The purpose of this simulation study is to show the empirical behavior of the multidimensional Thurstonian IRT model with the smallest number of traits, 2. For example, one can think of measuring global personality factors, such as 'Dynamism' and

'Social Propriety' also referred to as 'Getting Ahead' and 'Getting Along' (Hogan, 1983). Alternatively, any narrow traits can also be measured in this fashion.

Twelve conditions were examined in this simulation study by crossing the following factors: (a) keyed direction of items (all positively worded in one condition, or a mixture of positively and negatively worded items in the other condition – items are combined in pairs so that half of the comparisons are between items keyed in the same direction, and half are between items keyed in the opposite directions); (b) items per trait (12 or 24); and (c) correlation among traits (0, 0.5, or -0.5). Items were presented in pairs, so there are $p \times \tilde{n} =$ 12 or 24 binary outcome variables in the short and long questionnaires, respectively. 1000 replications were obtained for each condition. Sample size was 1000 observations in all conditions.

Table 1 shows the true item intercepts, factor loadings and uniquenesses for conditions with 12 items per trait (uniquenesses are fixed when only 2 traits are involved). The true item parameters were drawn from a uniform distribution: between 0.65 and 0.95 for factor loadings, and between -1 and 1 for thresholds. True uniquenesses were specified to be $\psi_i^2 = 1 - \lambda_i^2$. Such parameters would be typical for standardized continuous utilities of good items measuring a common factor. For the 24 items per trait, these item parameters are simply duplicated.

To estimate these models, we fixed the loadings for the first two items as well as all uniquenesses (see the discussion above on identification). These parameters were fixed to their true values to compute the bias of the estimates more easily. For completeness, Table 1 also includes the model parameters in the intercept/slope parameterization.

Insert Tables 1 and 2 about here

Item parameter recovery. For all conditions investigated, Table 2 provides the

average relative bias across the estimated thresholds and factor loadings, as well as the average relative bias of their standard errors. For the inter-trait correlations, since some of the conditions involve true values of 0, we provide instead the mean and standard deviation of the parameter estimates, and the mean of the estimated standard errors. It can be seen that item parameters and their SEs are very accurately estimated when positively and negatively keyed items are intermixed in a questionnaire. Thresholds are estimated with a relative bias of about 1%, and factor loadings' relative bias is smaller than 3%. Also, their SEs are estimated with a relative bias smaller than 1%, and smaller than 3%, respectively. Similar results are obtained for all values of correlations among the traits.

When only positively keyed items are employed, parameter estimates are unacceptable for some conditions, using a cutoff of 10% for relative bias. Particularly, standard errors are unacceptable for all conditions with positive items (although they are considerably better in the conditions involving the long test). In fact, the number of converged replicas, also reported in Table 2, immediately reveals that there are estimation problems when all the items are positively keyed.

Latent trait score recovery. We used the first replication in each condition to evaluate the trait recovery. MAP scores for each latent trait were obtained using M*plus* and they were correlated with the true latent trait scores. The square of this correlation provides us an estimate of the actual reliability of the MAP scores for each latent trait – see (28). These are presented in Table 3 along with the empirical reliability estimates computed using (27).

The estimated reliabilities are just an averaged result; a more accurate view of the error of measurement obtained is procured by examining the MAP information function. Figure 4 provides the MAP test information functions computed in direction of Trait 1 for the short and the long questionnaires with uncorrelated traits and positively and negatively

keyed items.

Insert Table 3 and Figures 4 and 5 about here

We see in Table 3 that the actual reliabilities obtained (or equivalently, the latent trait recoveries, which are the square root of the actual reliabilities) depend clearly on the three factors experimentally manipulated in this simulation study. For positively keyed items, none of the latent trait recoveries (and hence actual reliabilities) obtained are acceptable. Indeed, correlations between MAP scores and true scores are around 0.35 when the traits correlate positively, around 0.6 when they are uncorrelated, and around 0.79 when they are negatively correlated. Also, the effect of test length on latent trait recovery is small.

When positively and negatively keyed items are combined in blocks, latent trait recovery is much more accurate. It is around 0.87 for the short questionnaire, and around 0.92 for the long questionnaire. Figure 5 explains this difference. It shows plots of the true trait scores versus the estimated MAP scores for the short and the long questionnaires. There are clear floor and ceiling effects in the case of the short questionnaire, but these effects disappear in the case of the longer questionnaire. Converting these correlations into the actual reliabilities shown in Table 3 reveals that in this case 12 item-pairs provide reliability levels that are considered just acceptable for a personality questionnaire, and 24 item-pairs provide very good reliability indeed.

In applications, true latent traits are not known, and one needs to resort to the empirical reliability estimate to infer latent trait recovery. Hence, it is of interest to compare the empirical reliabilities and the actual test reliabilities shown in Table 3. It can be seen that empirical reliabilities are fairly accurate for the long test, but they underestimate the true reliability by about 0.07 or 10% in the short test. This is most likely due to the variance of the observed score being low –see equation (27), which is typical when the MAP estimator

is used with a small number of items (it is biased towards the population mean).

Goodness-of-fit tests. The estimation method used also yields a goodness-of-fit test of the model to the estimated tetrachoric correlations. Results for these tests are shown in Table 4. Given that model parameters are poorly estimated when only positively keyed items are employed, it is not surprising that goodness-of-fit tests in this case are off as well. The test statistic retains the model more often than it should. The results for the conditions where positively and negatively keyed items are combined in blocks are, not surprisingly, much better. For small models, the test statistic maintains its nominal rates. For the larger models, it tends to over-reject the model, although very slightly. Interestingly, the condition with 12 pairs and negatively correlated traits behaves very much like the conditions involving 24 pairs.

Insert Table 4 about here

Discussion. To understand why the results obtained when all the items are positively keyed are so poor, we turn to Figure 3a, which depicts the ICS for the binary outcome involving two positively keyed items from this example. We see in this figure that the change in the surface's slope depends on the direction in the trait space. The slope is high in the direction taken from an angle of about 45° towards the positive end of the first trait (η_1) and towards the negative end of the second trait ($-\eta_2$). It means this binary outcome contributes a sizeable amount of information to the trait difference score ($\eta_1 - \eta_2$). Therefore pairs where one has to chose between two positively keyed items will highlight *differences* in the two latent traits. At the same time, the ICS appears essentially flat in the direction taken from an angle of about 45° towards the positive ends of both traits. The same pair of items would provide virtually no information on the sum score ($\eta_1 + \eta_2$) of the two latent traits. Therefore, this binary outcome provides information on the *relative* position of the two

underlying trait scores, but not on their *absolute* locations. When all binary outcomes provide information on the relative position of the traits but no binary outcome provides information towards their absolute location, trait recovery score is poor, and item parameter recovery is poor as well. This problem is aggravated even further when the measured traits are positively related to each other. This is because the information provided by the binary outcome is lower in this situation, as can be seen from equations (21) and (22). On the other hand, in the case of negatively correlated traits, binary outcomes provide more information.

In contrast, pairs including items keyed in opposite directions add information about the traits' sum. Thus, in the conditions marked as +/- in Tables 2 to 4 in which some pairs consist of positively keyed items, and some pairs of items keyed in opposite directions we obtain information about the traits' difference and their sum, thus being able to locate both traits. This is why latent trait recovery and item parameter recovery is so much better in these conditions.

We have to conclude that when measuring 2 traits, the forced-choice design with items keyed in the same direction is not recommended. When traits are negatively correlated, the recovery of scores is better but still falls short of acceptable levels. In contrast, latent trait estimation can be precise when both positively and negatively keyed items are combined in the same blocks. Relationships between the traits do not affect the effectiveness of the IRT score estimates in this case. We suggest combining positive and negative items (making positive-positive item pairs, positive-negative item pairs, and negative-positive item pairs) in order to locate the absolute trait scores in all applications with 2 dimensions. Combining negative items with negative should be avoided because it provides the same information as positive items, but can be confusing for respondents. Finally, when combining positively and negatively keyed items, as few as 12 item-pairs can be used to obtain reliability levels of around 0.75. If higher precision of measurement is

required, more item pairs should be used. Also, on the basis of this example, we tentatively conclude that the empirical reliability estimates give fairly accurate results, more so for longer questionnaires.

Simulation study 2. A forced-choice questionnaire measuring 5 traits using blocks of different sizes

The main purpose of this simulation is to investigate the effect of using different block sizes. Six conditions were investigated by crossing (a) keyed direction of items (all positively worded, or both positively and negatively worded so that there are equal numbers of outcomes of each type); and (b) block size (2, 3, or 4 items per block). For each condition 1000 replications were obtained. Sample size was 1000 observations in all conditions. Twelve items were used per dimension. The same values used in the previous simulation (shown in Table 1) were also used here. The true correlations among the latent traits were set to values reported for the Big Five factors measured in the NEO PI-R (Costa & McCrae, 1992): -0.21, 0, -0.25, -0.53, 0.40, 0, 0.27, 0, 0, 0.24 for traits 1 and 2, traits 1 and 3, etc.

The number of traits in this example allows combining various numbers of items in each block, still keeping the pure multidimensional forced-choice design. We will investigate the three most popular forced-choice formats: blocks of 2 items (pairs), blocks of 3 items (triplets), and blocks of 4 items (quads). For each of these formats, a questionnaire was designed where no items from the same dimension were presented in the same block, using all 12 items per trait, 60 items in total. The questionnaire with *pairs* consisted of $60/2 = 30$ blocks, the questionnaire designed with *triplets* consisted of $60/3 = 20$ blocks, and the questionnaire designed with *quads* consisted of $60/4 = 15$ blocks.

Design 1: Blocks of 2 items (pairs). In the first questionnaire design with blocks of $n = 2$ items, we measure $d = 5$ traits with $m = 60$ items (12 items per trait), and the number of blocks is $p = 30$. Each block produces $\tilde{n} = 1$ binary outcome, therefore the total number

of binary outcomes is $p \times \tilde{n} = 30$, and each trait is measured by 12 binary outcomes. To identify this model, all uniquenesses have to be fixed, but no equality constraints on factor loadings are required (see the section on identification of Thurstonian IRT models above). The degrees of freedom do not need to be adjusted as there are no redundancies in blocks of 2 items.

The model estimation proceeded successfully for 954 replications when positive items only were used; and for all 1000 replications when both positive and negative items were combined in blocks. Both versions yielded correct empirical rejection rates for the chi-square tests (see Table 5 for goodness-of-fit statistics results). Item parameters and trait correlations were estimated accurately (see Table 6 for parameter estimation statistics). The correlations between traits were positively biased by about 12% for the model with all positive items, but for the model with both positive and negatively keyed items they were recovered to a very high degree of accuracy. In the questionnaire with all items being positively keyed the standard errors of correlations were negatively biased by about 30%, the item loadings' SEs were negatively biased by about 20% and the item thresholds' SEs were negatively biased by about 10%. In the questionnaire with positive and negatively keyed items standard errors had negligible bias.

Insert Tables 5 and 6 about here

We consider the first replication to evaluate how well the true scores were recovered in this example. The true scores and MAP scores correlated on average at 0.822 for the questionnaire with all positive items, and at 0.889 for the questionnaire combining both positive and negative items. When these correlations are converted into estimates of reliability using Equation (28), they yield reliabilities just below 0.7 for the positively keyed items design, and at around 0.79 for the positive/negative item design (all reliabilities are

reported in Table 7).

Insert Table 7 about here

The test information functions and the average squared errors were also computed for this replication, and turned into the reliability estimates using (27). Comparing these empirical reliability estimates presented in Table 7 to the actual reliabilities, we can see that for both designs the information method slightly underestimates the reliability, on average by about 5%. This is likely due to the relatively small number of binary outcomes per trait (12), leading to a substantial "compression" of the MAP score, and consequently small observed score variance.

We conclude that in a forced-choice application with 5 traits, the design with 30 positively keyed item-pairs would fall slightly short of the measurement precision that is typically required. However, the questionnaire can be sufficiently precise when both positive and negative items are combined in blocks. In this design with pairs we also note that only 12 binary outcomes per trait are produced. Increasing the number of binary outcomes should lead to a higher measurement precision. This can be achieved in 2 ways: by increasing the number of items per trait, or simply changing the questionnaire format to blocks of 3 or 4 items, drawing them from the same item pool. Next we turn to the design using the same 60 items combined in blocks of 3.

Design 2: Blocks of 3 items (triplets). In this case, the questionnaire consists of the same $m = 60$ items used in the previous design but presented in $p = 20$ triplets ($n = 3$). The items are arranged into triplets so that all 10 permutations of 3 out of 5 traits are equally represented. This makes each subset of 3 traits appear exactly 2 times in the questionnaire. Each block produces $\tilde{n} = 3$ binary outcomes, therefore the total number of binary outcomes in this model is $p \times \tilde{n} = 60$, and each trait is measured by 24 binary outcomes.

To identify this model, one item's uniqueness per block has to be fixed, but no constraints on factor loadings are required. The degrees of freedom in this case need to be adjusted because there are 20 redundancies in 20 blocks of 3 items (1 redundancy per block).

The estimation proceeded successfully for both versions (with positively keyed items only and with positively and negatively keyed items) for all replications. Empirical rejection rates for the chi-square, however, are much higher than expected (see Table 5); therefore models of this kind will be rejected more often than they should based on the test of exact fit. All item parameters were estimated very accurately, with negligible bias (see Table 6). The correlations between traits were positively biased by about 10% for the questionnaire with positive items, but they were recovered very accurately for the questionnaire with positive/negative items. In the questionnaire with all positive items the standard errors of correlations were negatively biased by about 15%, and the SE of item loadings were negatively biased by about 10%.

We consider the first replication to evaluate how well the true scores were recovered for both versions of the questionnaire. It has been shown that MAP estimation using the simplifying assumption of local independence provides very accurate results even when this assumption is violated in blocks of 3 or more items (Maydeu-Olivares & Brown, 2010). In the current design, the true scores and MAP scores correlated on average at 0.863 for the questionnaire with all positive items, and at 0.929 for the questionnaire combining positive and negative items. Converting these correlations into estimates of reliability using Equation (28), we obtain reliabilities of about 0.75 for the positive items design, and of about 0.86 for the positive/negative items design (see Table 7). The test information functions and the average squared errors were turned into the reliability estimates using (27), yielding figures of about 0.87 for the positively keyed items design, and of about 0.88 for the positive/negative items design. We can see that the information method is very accurate in

estimating the reliabilities for the questionnaire with positive and negative items, despite ignoring the correlated uniquenesses in this triplet forced-choice design, and making a simplifying assumption of local independence. The very minor over-estimation of about 2% is totally acceptable in practice.

However, the information method overestimates the reliability by about 17% for the design with positive items only. This is because positively keyed items on their own, as was explained above, are good at recovering the differences between the traits but not their sums, and therefore have limits in recovering the traits' absolute locations. There is a clear improvement in the trait recovery compared to the example with 2 traits; however, this improvement is due to the increased number of traits (we will expand this point in the discussion) and not to the increased number of binary outcomes. Adding binary outcomes of comparisons between positively keyed items is unlikely to improve the trait recovery further, as we will see in the design with quads.

We conclude that in a forced-choice application with 5 traits, 20 triplets can provide sufficient measurement precision. Particularly, the questionnaire combining both positive and negative items within blocks provides very good levels of measurement accuracy.

Design 3: Blocks of 4 items (quads). Our next design consists of $p = 15$ blocks of $n = 4$ items (quads), using the same 60 items as in the previous examples. The items are arranged into quads so that all 5 permutations of 4 out of 5 traits are equally represented. This makes each subset of 4 traits appear exactly 3 times in the questionnaire. Each block produces $\tilde{n} = 6$ binary outcomes, therefore the total number of binary outcomes in this model is $p \times \tilde{n} = 90$, and each trait is measured by 36 binary outcomes. Note that here we assume that full rankings are performed in each block (not the "most"-"least" incomplete ranking), and therefore there is no missing data involved.

To identify this model, one item's uniqueness per block has to be fixed, but no

constraints on factor loadings are required. The degrees of freedom in this case need to be adjusted because there are 60 redundancies in 15 blocks of 4 items (4 redundancies per block). The estimation proceeded successfully for both positive and positive/negative questionnaire versions for all 1000 replications. Similarly to the model with triplets, empirical rejection rates for the chi-square are much higher than expected (see Table 5). All item parameters were estimated very accurately (see Table 6). The trait correlations were positively biased by about 10% for the questionnaire with positive items, but they were recovered very accurately for the questionnaire with positive/negative items. In the questionnaire with all positive items the standard errors of correlations were negatively biased by about 14%, and the SE of item loadings were negatively biased by about 10%.

Again, we consider the first replication to evaluate how well the true scores were recovered. While the trait recovery has not improved compared to the triplet design for the questionnaire with all positive items (the true scores and MAP scores correlated on average at 0.87), it has improved even further to the impressive average of 0.94 for the questionnaire combining both positive and negative items. Converting these correlations into estimates of reliability using Equation (28), we obtain reliabilities of about 0.75 for the positively keyed items design, and of about 0.89 for the positive/negative item design (see Table 7). The empirical reliability estimates were 0.91 for the positively keyed items design, and 0.92 for the positive/negative item design. Thus, the information method is again accurate in estimating the empirical reliabilities for the questionnaire with positive and negative items, despite ignoring correlated errors in this forced-choice design with quads, and making a simplifying assumption of local independence. The information method makes a very minor over-estimation of around 3%, which would be considered acceptable in practice.

We conclude that in a forced-choice application with 5 traits, the design with 15 quads can provide sufficient measurement precision, particularly for the questionnaire

combining both positive and negative items within blocks.

An empirical application: A Big Five questionnaire constructed from IPIP items

To create a real forced-choice questionnaire, we used one of the designs described in the simulated Big Five example above as a template. Items were drawn from the International Personality Item Pool (IPIP), more specifically from its subset of 100 items measuring the Big Five factor markers (Goldberg, 1992). Note that constructs measured by this questionnaire are not the same as in the NEO-PIR, and therefore correlations between the five traits are expected to be different from those used in the simulation study with 5 traits. We selected 60 items so that 12 items would measure each of the 5 marker traits. We chose the triplet design from the simulation study above, with 8 positively and 4 negatively keyed items per trait combined in a way that equal number of pairwise comparisons occur between items keyed in the same direction and items keyed in opposite directions.

Each block of the questionnaire was presented in 2 formats. First, participants rated the 3 items using a 5-point rating scale suggested by Goldberg (1992), ranging from "very accurate" to "very inaccurate". This single-stimulus presentation was immediately followed by the forced-choice presentation, where the participants were asked to select one "most like me" item, and one "least like me" out of the same block of 3 items. Two formats were used in order to compare trait scores as estimated from the single-stimulus and forced-choice formats.

Four-hundred-and-thirty-eight volunteers from the UK completed the questionnaire online in return for a feedback report. Out of 433 participants who provided demographic information, 48.4% were male and 51.6% were female. Age ranged from 16 to 59 years with a mean of 33.3 and a standard deviation of 10.37 years. The largest ethnic group was white (64%) followed by Asian (18%) and Black (6.6%). Most participants were employed (55%), 23% were students and 14% were unemployed.

First, the single-stimulus version of the questionnaire was analyzed. We fitted the multidimensional version of the normal ogive graded response model (Samejima, 1969) to the item responses for all 5 traits simultaneously, using the ULS estimation in M*plus*. The five latent traits were allowed to correlate freely. The model fit was relatively poor with chi-square 3621.59 on 1700 degrees of freedom ($p < 0.001$), RMSEA = 0.051. Fitting the model one scale at a time revealed that the scale Openness had its items loading on 2 dimensions (namely imagination, and preference for complex and abstract material). The scale Conscientiousness had 2 items with highly similar content (preference for order) that shared common variance not explained by the main factor. Other scales were broadly one-dimensional and showed good fit indices when tested on their own. However, we chose to proceed with the Big Five model without any modifications to estimate the model parameters and compute the MAP scores for individuals. The estimated correlations between the five traits are given in Table 8 (above the diagonal).

Next, the forced-choice questionnaire was analyzed. After coding the forced-choice rankings as binary outcomes, the 5-dimensional IRT model with freely correlated latent traits was fitted to these data in M*plus,* also using the ULS estimation. One item's uniqueness per block was fixed for identification. The forced-choice model yielded a better fit than the single-stimulus model: a chi-square of 2106.06 on 1640 degrees of freedom, RMSEA = 0.025 (degrees of freedom and RMSEA are corrected for the number of redundancies in the model, 20). The estimated correlations between the five dimensions in this model are given in Table 8 (below the diagonal). It can be seen that these correlations are very similar to the trait correlations estimated from the single-stimulus data for all but one correlation. The correlation between traits Agreeableness and Openness is higher for the single-stimulus version (0.41) than for the forced-choice version (0.15).

Insert Tables 8 and 9 about here

The MAP estimated trait scores for individuals based on single-stimulus and forced-choice responses correlated strongly, with correlations ranging from 0.69 for Agreeableness to 0.82 for Extraversion (see Table 9). Scale empirical reliability estimates for the forced-choice data were computed based on the IRT information method described above. Reliability estimates for the single-stimulus data were also computed using equations (26) and (27). The reliability estimates ranged from 0.775 to 0.844 for the single-stimulus data, and from 0.601 to 0.766 for the forced-choice data (see Table 9). It can be seen that the rank-order of scales in terms of their reliability is the same for both formats, however, the reliabilities are lower by about 0.1 for the forced-choice format. Clearly, responses to 60 items using the ordinal 5-point scale provided more information than 60 binary outcomes of rankings.

Also, the reliability estimates in this application are lower than those obtained in the simulation study with 5 traits and the same triplet design. This is due to generally lower item loadings found in this application than those used in the simulation. For most items, standardized factor loadings found in the single-stimulus version of the IPIP Big Five questionnaire were between 0.5 and 0.7, whereas they were between 0.65 and 0.95 in the simulated examples. The nature of the broad marker traits in this application meant that the factor loadings were lower than would be typically found in a questionnaire with more narrowly defined traits.

Discussion

In this paper we introduced a Thurstonian IRT model suitable for modeling responses to multidimensional forced-choice questionnaires with dominance items. The model proposed here is an IRT formulation of the Thurstonian second-order factor model for comparative data introduced in Maydeu-Olivares and Böckenholt (2005) applied to the problem at hand. The model enables straightforward estimation of individual trait scores and test information functions. The proposed estimation method is extraordinarily fast and capable of handling

models of any size. As such, we have successfully managed to estimate models with 32 traits. Provided the model is not too large (15 traits each measured by at most 10 items appears to be the limit with current computing capabilities), standard errors and fit indices can be obtained as well.

Simulation studies were performed to investigate the performance of the model across a variety of forced-choice designs. The simulation studies show that the model parameters (trait correlations, factor loadings, thresholds, and uniquenesses) are recovered very accurately from the binary outcomes in all reasonable designs. Some designs are simply not recommended, such as an unrealistic design involving exactly two traits using positively keyed items only. The poor results obtained in these designs do not reflect the limitations of the model or estimation method employed, but rather, the limitations of the forced-choice format.

The simulation studies also provide important information about the assessment of model fit. The chi-square statistic provide reasonable empirical rejection rates in all models where item parameters are accurately estimated, provided the model is not too large. In models with over 1000 degrees of freedom the chi-square statistic grossly underestimates the degree of model fit even though item parameters and their standard errors are very accurately estimated. For instance, around 27% of models would be empirically rejected in the five factor designs with triplets, and around 37% of models with quads, where only 5% should be rejected.

The designs used in the simulation studies were chosen to answer important questions about strengths and limitations of forced-choice questionnaires with dominance items. Despite many discussions in the literature, many of these questions remain controversial as the evidence is largely based on inadequate scoring schemes that assign consecutive integers to the subjects' responses leading to ipsative scores on the measured traits. In addition, much

of past research is based on specific questionnaires with very different properties that preclude meaningful generalization. Results of the simulation studies in this paper have important implications on how forced-choice tests should be designed and used in the future. We will discuss the most important points here.

Perhaps the most interesting and much debated question is whether scores based on relative forced-choice responses can resemble the absolute trait scores. Our research shows that the true trait scores can be recovered to a high degree of accuracy under certain conditions. Certainly more items with higher discriminations will, generally speaking, improve the latent trait recovery, just as it is the case with single-stimulus questionnaires. However, there are additional important factors specific to the forced-choice format. These are: keyed direction of items, the number of measured traits, the trait correlations, and block size. We discuss each of these factors in turn.

Keyed direction of items. When the forced-choice design produces binary outcomes from comparing items from different traits keyed in the same direction, and approximately the same number of binary outcomes from comparing items keyed in opposite directions, the trait recovery can be good with any number of traits, and any trait correlations. This is because items keyed in the same direction contribute to the measurement of the *difference* between 2 trait scores; and items keyed in opposite directions contribute to the measurement of the *sum* of the 2 traits involved. When information on the sums and differences of the traits involved is available, trait scores can be determined accurately because their absolute value can be located in the traits' continuum. However, when only differences between traits are estimated, which is the case in forced-choice designs with items keyed in the same direction, latent trait and item parameter recovery largely depends on the number of traits assessed – and this is the next factor in our discussion.

Alternatively, items that measure the same trait can be used together in blocks to

provide information on the latent trait directly, as the binary outcomes of such comparisons will depend on only one trait. However, the comparative nature of the forced-choice format means that the dominance items measuring the same trait will only provide sizeable amount of information when their factor loadings are very different (Maydeu-Olivares & Brown, 2010), as it is the case with items keyed in opposite directions.

One last comment on using negatively keyed items in forced-choice questionnaires concerns the use of negation. In our experience, responding to forced-choice blocks involving items with negation can be confusing for respondents; therefore straight negation should be avoided and replaced wherever possible with appropriate synonyms.

Number of traits. When the number of traits is large, and traits are not strongly positively correlated overall, any forced-choice designs will reliably locate trait scores provided that sufficient numbers of good quality items are used. That is, it is possible to locate absolute trait scores using only positively keyed items when the number of traits assessed is large. Baron (1998) shows that even questionnaires scored using classical methods leading to ipsative data with all positive items measuring many relatively independent traits (30 or more) correlate strongly with their single-stimulus counterparts. Why it is important that traits are relatively independent will be the next point of our discussion.

Our simulation studies show that when assessing only two traits, positively keyed items on their own cannot recover the absolute latent trait scores. In simulations involving five factors, where positive and negative correlations between traits were balanced bringing the average inter-trait correlation close to 0, the true score recovery was good for designs with positive items only (except when blocks of 2 items were used, where the number of binary outcomes was not sufficient). How do binary outcomes measuring only *differences* between traits provide information on *absolute* trait scores when the number of traits is large?

When only two traits are measured, the information about the first trait is conditional only on the second trait (and vice versa). As we can see from Figures 3b, there is sizeable amount of information for scores that are similar, for example (-2, -2) or (2, 2), but virtually no information for scores that are different, for example (2, -2). There are many combinations of two scores possible that are very different from each other. For instance, assuming normally distributed traits that are uncorrelated with unit variance, trait scores are different by more than 0.5 standard deviations for around 75% of cases. Therefore for most combinations of latent scores, the test information provided by such a questionnaire will be very low.

There are much fewer ways in which 5 trait scores can be different from each other. Again, assuming uncorrelated normally distributed traits with unit variances, only 3% of the cases can be expected to have differences greater than 0.5 standard deviations between all five trait scores. Because the test information about one trait will be conditional on the other 4 traits in the five-dimensional case, and because it is more likely that at least one of those traits will be similar to the target trait, it is more likely that the information on the target trait will be higher overall.

Extending this logic further, for 30 independent traits there is less than 0.03% chance that all trait scores will be different by 0.5 standard deviations or above. In this case the information about one trait is conditional on 29 other traits, and because many of them will be similar to the target trait the information will be high for most combinations of scores.

Correlations between traits. For a given set of item parameter values, comparing items keyed in the same direction is more effective when the traits are uncorrelated than when they are positively correlated, and it is even more effective when they are negatively correlated. This is apparent from the information functions provided by equations (21) and (22). For binary outcomes of comparisons between items measuring uncorrelated traits, only

the focus trait contributes to the information. However, for pairs involving correlated traits, the other trait involved will also contribute to the information. It will increase the information if correlated negatively with the target trait, and reduce it if correlated positively.

The inter-trait correlations have a major impact on the effectiveness of any forced-choice questionnaire with positively keyed items. Given the same number of traits, the lower the average correlation between them the better the true scores are recovered. For example, in the simulation study with 5 traits the average off-diagonal trait correlation was 0. In the design with positive items only, reversing the first scale, which negatively correlated with the rest (imagine turning Neuroticism into Emotional Stability in the context of the Big Five), would turn the average correlation positive and significantly worsen the trait recovery.

Block size. By using blocks of different sizes in the five-factor-model simulation studies we show that the same items provide more information by simply combining them in larger blocks. This is because, given the same number of items, the number of binary outcomes will increase when the block size increases. For example, 60 items will produce only 30 binary outcomes when put in blocks of 2, but produce 60 binary outcomes when put in blocks of 3, and produce 90 binary outcomes when put in blocks of 4. In other words, using larger blocks is attractive because it saves producing and trialing new items, which can be time consuming and expensive. However, increasing block size increases respondents' cognitive load as there are $\tilde{n} = \frac{n(n-1)}{2}$ binary comparisons to be performed in a block of n items. In practice, blocks of 4 items are probably the upper limit for forced-choice tests.

To summarize, adhering to the above recommendations (i.e. balancing the number of traits and their correlations, the direction of items, the number of items and the block size) is important for the quality and usefulness of the resulting questionnaire. Provided these factors have been taken into account, most personality items can be used in forced-choice

questionnaires. Thousands of dominance personality items have been written and translated to different languages over years. We have shown how these simple items can be used effectively.

The usefulness of the IRT model proposed here was illustrated in an empirical study involving a questionnaire measuring the broad Big Five markers, using both ratings and rankings (forced-choice blocks). In applications, results from single-stimulus questionnaires (i.e., using ratings) are often used as a benchmark to compare with results obtained from forced-choice questionnaires. However, doing so assumes that that no systematic biases affect the ratings. Yet, extant research (e.g. Van Herk, Poortinga & Verhallen, 2004; Bartram, 2007) shows that different types of biases can be present when responses are obtained using ratings. The fact that we obtained a better model fit in the forced-choice version of the Big Five questionnaire than in the single-stimulus version might be an indication of some reduction in response biases when a comparative response format is used. Responding in socially desirable manner is often associated with such personality traits as Openness and particularly Agreeableness. It is possible that inflation of responses to these traits by some individuals is responsible for the much higher correlation between Openness and Agreeableness observed in the single-stimulus version of the IPIP questionnaire application.

We also described a method of estimating the empirical test reliability based on computing MAP information and the average error variance for a scored sample. Reliability estimates produced by this information method were compared to reliabilities derived from the correlations between estimated and the true trait scores. In general, our proposed method provides accurate estimates of the reliability coefficients in designs where increasing the number of binary outcomes improves the latent trait estimation accuracy. This was the case in the simulation studies with 2 and 5 traits where positively and negatively items were combined together in blocks. In these studies, the information method provided sufficiently

accurate estimates of test reliability even for triplets and quads, where the local independence does not hold and the simplifying assumption of local independence is made for estimating the test information. Ignoring correlated errors led to a very minor overestimation of reliability – for blocks of 3 the reliability was overestimated by about 2%, and for blocks of 4 by about 3%. The reader must also be aware that for very short questionnaires, the information method might underestimate the reliability when the MAP latent trait estimator is used due to fact that this estimator "compresses" the scores.

In questionnaires using only positively keyed items, as is shown earlier, the accuracy of the latent trait recovery depends heavily on the number of traits assessed. In such questionnaires, after a certain precision of latent trait recovery has been reached, increasing the number of binary outcomes will not improve latent trait recovery further. In this case the information method might overestimate the reliability for blocks of any size. This is because the addition of binary outcomes does not increase information uniformly across the latent trait distribution. Rather, the information function becomes very picked in areas where latent trait scores are very similar to each other, and is almost zero elsewhere. In this situation, the information method to estimate the reliability fails to reflect very varied levels of test information at different trait scores. Thus, the information method of computing reliability is recommended only when the information function is relatively uniform.

Conclusions

The Thurstonian IRT model introduced in this paper describes the decision process of responding to forced-choice personality questionnaires measuring multiple traits. This model can be used with any forced-choice instrument composed of items fitting the dominance response model, with any number of measured traits, and any block sizes (i.e. pairs, triplets, quads etc.). This makes it widely applicable to many existing forced-choice questionnaires such as the Occupational Personality Questionnaire (OPQ; SHL, 2006), the Customer

Contact Styles Questionnaire (CCSQ; SHL, 1997), the Survey of Interpersonal Values (SIV; Gordon, 1976), and useful in designing future questionnaires. The Thurstonian IRT model can be embedded within a familiar SEM framework to be estimated and scored by general-purpose software (we used M*plus* throughout this paper). The model also provides means of estimating reliability for forced-choice questionnaires, which has been problematic under the classical scoring schemes (Tenopyr, 1988; Baron, 1998).

The proposed IRT approach allows using the forced-choice format, which reduces certain response biases, while getting the benefits of standard data analysis techniques that users of single-stimulus questionnaires have enjoyed. The forced-choice format itself, of course, cannot correct faults in test construction, and sometimes might make them even more apparent. As we have shown, creating a forced-choice questionnaire requires consideration of many more factors than a single-stimulus questionnaire. Provided these factors are carefully taken into account, and sufficient work has been put into combining suitable statements together in forced-choice blocks, the format can deliver significant advantages. By removing the peculiar properties of ipsative data, we hope that the theoretical barriers against the use of the forced-choice format will start to fall.

References

Ackerman, T.A. (2005). Multidimensional Item Response Theory Modeling. In A. Maydeu-Olivares & J. J. McArdle. (Eds.). *Contemporary Psychometrics* (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum.

Baron, H. (1996). Strengths and Limitations of Ipsative Measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*, 263-272.

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*, 55-77.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Costa, P.T. & McCrae, R.R. (1992). *NEO-PI-R Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Du Toit, M. (Ed.). (2003). *IRT from SSI*. Chicago: SSI Scientific Software International.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Forero, C.G., Maydeu-Olivares, A. & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26-42.

Gordon, L.V. (1976). *Survey of interpersonal values*. Revised manual. Chicago, IL: Science

Research Associates.

Hogan, R. (1983). A socioanalytic theory of personality. In M.M. Page (Ed.), *Nebraska Symposium on Motivation* (pp. 336-355). Lincoln: University of Nebraska Press.

International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64*, 325-340.

Maydeu-Olivares, A. & Böckenholt, U. (2005). Structural equation modeling of paired comparisons and ranking data. *Psychological Methods*, *10*, 285-304.

Maydeu-Olivares, A. & Brown, A. (in press). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research.*

Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362.

McCloy, R., Heggestad, E., Reeve, C. (2005). A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods, 8*, 222-248.

McDonald, R.P. (1999). *Test theory. A unified approach.* Mahwah, NJ: Lawrence Erlbaum.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organisational Psychology*, 77, 531-552.

Muthén, L.K. & Muthén, B. (1998-2007). *Mplus 5.* Los Angeles, CA: Muthén & Muthén.

Reckase, M. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

SHL. (1997). *Customer Contact: Manual and User's Guide*. Surrey, UK. SHL Group.

SHL. (2006). *OPQ32 Technical Manual*. Surrey, UK. SHL Group.

Stark, S., Chernyshenko, O. & 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.

Stark, S., Chernyshenko, O., Drasgow, F. & Williams, B. (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.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology*, *73*, 749-751.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, *79*, 281-299.

Thurstone, L.L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology*, *14*, 187-201.

Van Herk, H., Poortinga, Y., & Verhallen, T. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology*, 35, 346.

Appendix A. Designs involving blocks of 2 items (pairs)

In designs involving blocks of $n = 2$ items (pairs), there is only one binary outcome per block, and both uniquenesses involved cannot be identified. Without loss of generality, they can be fixed to 0.5. This is equivalent to setting the error variance of the latent response variable to 1.

To illustrate, consider a short test measuring $d = 3$ traits using pairs. Each trait is measured by 4 items. The contrast matrix $\mathbf{A}$ and a typical factor loadings matrix $\mathbf{\Lambda}$ are

$$\mathbf{A} = \left(\begin{array}{cc|cc|c|cc} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ \hline 0 & 0 & 1 & -1 & & 0 & 0 \\ \hline \vdots & & & & \ddots & & \vdots \\ \hline 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{array}\right), \mathbf{\Lambda}' = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & \lambda_6 & 0 & 0 & \lambda_9 & 0 & 0 & \lambda_{12} \\ 0 & \lambda_2 & 0 & \lambda_4 & 0 & 0 & 0 & \lambda_8 & 0 & \lambda_{10} & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & \lambda_5 & 0 & \lambda_7 & 0 & 0 & 0 & \lambda_{11} & 0 \end{pmatrix}.$$

With $\mathbf{\Psi} = .5\,\mathbf{I}$ (for identification), the parameter matrices of the Thurstonian IRT model are $\breve{\mathbf{\Psi}} = \mathbf{A}\mathbf{\Psi}\mathbf{A}' = \mathbf{I}$ and

$$\breve{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ 0 & -\lambda_4 & \lambda_3 \\ -\lambda_6 & 0 & \lambda_5 \\ 0 & -\lambda_8 & \lambda_7 \\ \lambda_9 & -\lambda_{10} & 0 \\ -\lambda_{12} & 0 & \lambda_{11} \end{pmatrix}.$$

As this example illustrates, designs involving pairs are very special because a) item responses are locally independent under the Thurstonian IRT model ($\breve{\mathbf{\Psi}}$ is diagonal), b) the model contains much fewer parameters ($\breve{\mathbf{\Psi}}$ is a fixed matrix), and c) no constraints among the model parameters need to be enforced (each factor loading λ_i only appears once in the factor loading matrix $\breve{\mathbf{\Lambda}}$). Modeling forced choice tests is much easier when items are presented in blocks of 2 (pairs), than in blocks of 3 or more items (triplets, quads, etc.).

Appendix B. Posterior MAP information for a trait in a multidimensional forced-choice questionnaire

The posterior MAP test information in direction a (direction in the factor space that coincides with the trait η_a) is the sum of the ML information and the additional component provided by the prior distribution (Du Toit, 2003):

$$\mathcal{I}_P^a\left(\boldsymbol{\eta}\right) = \mathcal{I}^a\left(\boldsymbol{\eta}\right) - \frac{\partial^2 \ln\left(\phi\left(\boldsymbol{\eta}\right)\right)}{\partial^2 \eta_a}. \tag{29}$$

For the d-variate standard normal distribution with means 0 and the covariance matrix $\boldsymbol{\Phi}$, the density function $\phi(\boldsymbol{\eta})$ is:

$$\phi(\boldsymbol{\eta}) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Phi}\right|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}\right), \quad \text{and}$$

$$\ln\left(\phi(\boldsymbol{\eta})\right) = -\ln\left((2\pi)^{d/2}\left|\boldsymbol{\Phi}\right|^{1/2}\right) - \frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}.$$

Computing the first derivative by η_a of the expression above, we notice that because the first part of the sum does not depend on η_a (it is constant), its derivative is 0. Thus

$$\frac{\partial \ln\left(\phi(\boldsymbol{\eta})\right)}{\partial \eta_a} = \frac{\partial}{\partial \eta_a}\left(-\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}\right) = -\frac{1}{2}\cdot\frac{\partial}{\partial \eta_a}\left(\boldsymbol{\eta}'\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}\right). \tag{30}$$

Let ϖ_i^j be an element of the inverted trait covariance matrix $\boldsymbol{\Phi}^{-1}$ in i^{th} row and j^{th} column:

$$\begin{aligned}&\frac{\partial}{\partial \eta_a}\left(\boldsymbol{\eta}'\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}\right) = \frac{\partial}{\partial \eta_a}\left(\eta_1\left(\sum_{j=1}^{d}\eta_j\varpi_j^1\right) + \ldots + \eta_a\left(\sum_{j=1}^{d}\eta_j\varpi_j^a\right) + \ldots + \eta_d\left(\sum_{j=1}^{d}\eta_j\varpi_j^d\right)\right) = \\ &\eta_1\varpi_a^1 + \ldots + \sum_{j=1,j\neq a}^{d}\eta_j\varpi_j^a + 2\eta_a\varpi_a^a + \ldots + \eta_d\varpi_a^d = 2\sum_{j=1}^{d}\eta_j\varpi_j^a\end{aligned} \tag{31}$$

Now, it follows from (30) and (31) that the second derivative by η_a of $\ln\left(\phi(\boldsymbol{\eta})\right)$ is

$$\frac{\partial^2 \ln\left(\phi(\boldsymbol{\eta})\right)}{\partial^2 \eta_a} = -\frac{1}{2}\cdot\frac{\partial}{\partial \eta_a}\left(2\sum_{j=1}^{d}\eta_j\varpi_j^a\right) = -\frac{1}{2}\cdot 2\frac{\partial}{\partial \eta_a}\left(\sum_{j=1}^{d}\eta_j\varpi_j^a\right) = -\varpi_a^a, \tag{32}$$

and we can substitute the above expression into Equation (29), arriving at Equation (24).

Table 1

True item parameters for the short questionnaire (12 item-pairs) using both positively and negatively keyed items; simulated example with 2 traits.

	threshold / factor loading parameterization							intercept / slope parameterization			
i	μ_i	λ_i	ψ_i^2	k	μ_k	λ_k	ψ_k^2	$l = i,k$	α_l	β_i	β_k
1	-0.44	0.91	0.17	2	-0.1	0.81	0.35	1, 2	-0.47	1.26	1.12
3	-0.77	0.75	0.44	4	0.21	0.73	0.47	3, 4	-1.03	0.79	0.77
5	0.02	0.83	0.31	6	-0.65	0.67	0.55	5, 6	0.72	0.90	0.72
7	0.64	0.94	0.12	8	0.71	0.66	0.57	7, 8	-0.08	1.13	0.79
9	-0.2	0.8	0.36	10	0.69	-0.7	0.51	9, 10	-0.95	0.86	-0.75
11	0.3	-0.72	0.49	12	0.68	0.88	0.23	11, 12	-0.45	-0.85	1.04
13	0.03	0.91	0.17	14	-0.5	-0.79	0.37	13, 14	0.72	1.24	-1.08
15	-0.57	-0.84	0.29	16	-0.57	0.7	0.51	15, 16	0.00	-0.94	0.78
17	0.77	-0.87	0.24	18	0.36	0.79	0.37	17, 18	0.52	-1.11	1.01
19	0.65	0.79	0.38	20	-0.25	-0.7	0.51	19, 20	0.95	0.84	-0.74
21	-0.47	-0.68	0.54	22	-0.62	0.72	0.48	21, 22	0.15	-0.67	0.71
23	-0.21	0.7	0.51	24	0.28	-0.66	0.56	23, 24	-0.47	0.68	-0.64

Notes: The order of traits is alternated in the questionnaire to avoid carry-over effect, so that in odd pairs the first item measures Trait 1 and the second measures Trait 2, and in even pairs this order is reversed.

Table 2

Results for parameter estimates and standard errors in the simulation studies with 2 traits

Keyed direction of items	Number of items per trait	Number of converged replications	Correlation between traits			Loadings' average relative bias		Thresholds' average relative bias	
			True	Estimates mean (SD)	SE mean	Est.	SE	Est.	SE
+	12	868	0	-.104 (.250)	.232	.126	12.17	.161	13.33
+	12	853	0.5	.437 (.184)	.151	.039	6.75	.078	7.67
+	12	815	-0.5	-.560 (.252)	.306	.171	10.99	.184	12.73
+	24	974	0	-.097 (.218)	.194	-.005	.913	.037	1.34
+	24	980	0.5	.438 (.149)	.115	-.035	-.094	.009	.145
+	24	893	-0.5	-.580 (.226)	.258	.016	1.60	.028	.916
+/-	12	1000	0	.039 (.191)	.188	.023	-.029	.010	-.002
+/-	12	999	0.5	.460 (.184)	.166	.028	-.020	.007	.007
+/-	12	1000	-0.5	-.479 (.106)	.106	.011	.000	.007	.001
+/-	24	1000	0	.031 (.171)	.166	.018	-.013	.013	-.008
+/-	24	1000	0.5	.480 (.109)	.103	.017	-.022	.009	-.002
+/-	24	1000	-0.5	-.477 (.099)	.097	.005	-.002	.000	.001

Table 3

Test reliabilities in the simulation studies with 2 traits

Keyed direction of items	Number of items per trait	True trait correlation	Actual reliability		Empirical reliability	
			Trait 1	Trait 2	Trait 1	Trait 2
+	12	0	.385	.386	-	-
+	12	0.5	.171	.107	-	-
+	12	-0.5	. 627	. 568	-	-
+	24	0	.402	.438	-	-
+	24	0.5	.256	.209	-	-
+	24	-0.5	.637	. 621	-	-
+/-	12	0	.760	.740	.691	.674
+/-	12	0.5	.780	.779	.739	.756
+/-	12	-0.5	.747	.725	.665	.645
+/-	24	0	.842	.843	.842	.840
+/-	24	0.5	.851	.859	.871	.877
+/-	24	-0.5	.819	.820	.822	.812

Notes: The actual test reliability is computed as squared correlation between true scores and MAP score estimates; the empirical reliability is calculated using equation (26). For the questionnaires with positive items, the information method is not recommended (see text).

Table 4

Goodness-of-fit results for the simulation studies with 2 traits

Keyed direction of items	Number of items per trait	Correlation between traits	Degrees of freedom	Chi-square mean	Chi-square SD	Chi-square rejection rates			
						.01	.05	.10	.20
+	12	0	43	36.88	7.73	.000	.009	.018	.050
+	12	0.5	43	38.26	8.16	.004	.012	.034	.076
+	12	-0.5	43	36.39	7.91	.000	.005	.016	.043
+	24	0	229	217.84	20.36	.001	.012	.033	.093
+	24	0.5	229	221.93	20.55	.006	.021	.049	.115
+	24	-0.5	229	215.18	20.44	.000	.011	.027	.071
+/-	12	0	43	43.37	9.51	.016	.049	.105	.206
+/-	12	0.5	43	42.66	9.32	.010	.041	.093	.185
+/-	12	-0.5	43	44.00	10.19	.018	.079	.146	.260
+/-	24	0	229	232.09	24.57	.026	.078	.150	.261
+/-	24	0.5	229	233.09	25.25	.032	.074	.137	.261
+/-	24	-0.5	229	232.51	23.01	.027	.071	.134	.240

Table 5

Goodness-of-fit results for the simulation studies with 5 traits

Block size	Keyed direction of items	Number of converged replications	Degrees of freedom	Chi-square mean	Chi-square SD	Empirical rejection rates of chi-square test			
						.01	.05	.10	.20
2	+	954	365	360.38	33.60	.023	.067	.119	.206
	+/-	1000	365	367.44	34.23	.047	.110	.156	.248
3	+	1000	1640*	1669.01	106.36	.154	.272	.334	.418
	+/-	1000	1640*	1677.73	101.00	.155	.276	.354	.447
4	+	1000	3830*	3920.31	192.62	.266	.362	.432	.518
	+/-	1000	3830*	3914.01	196.52	.268	.388	.458	.534

Notes: Number of items per trait is 12 for all designs. (*) Degrees of freedom are adjusted according to the number of redundancies in the model.

Table 6

Average relative bias for parameter estimates and standard errors in the simulation studies with 5 traits

Block size	Keyed direction of items	Correlations		Loadings		Thresholds		Uniquenesses	
		est	SE	est	SE	est	SE	est	SE
2	+	.117	-.306	.011	-.212	.013	.131	fixed	fixed
	+/-	.006	-.012	.020	-.027	.014	-.015	fixed	fixed
3	+	.102	-.153	-.001	-.103	.007	.000	.018	-.024
	+/-	.006	-.018	.015	-.015	.008	-.004	.020	-.022
4	+	.095	-.141	-.004	-.107	.001	-.010	.006	-.026
	+/-	.006	-.015	.011	-.010	.005	-.007	.010	-.023

Notes: Uniquenesses are fixed for the design with pairs to identify the model.

Table 7

Test reliabilities in the simulation studies with 5 traits

Block size	Keyed direction of items	Reliability	Dimension				
			1	2	3	4	5
2	+	actual	.698	.664	.648	.689	.683
		empirical	.717	.603	.576	.707	.663
	+/-	actual	.772	.811	.783	.798	.786
		empirical	.709	.771	.747	.754	.761
3	+	actual	.767	.752	.710	.735	.762
		empirical**	.904	.862	.852	.879	.875
	+/-	actual	.849	.872	.859	.872	.863
		empirical**	.885	.878	.873	.880	.878
4	+	actual	.767	.744	.710	.764	.774
		empirical**	.922	.892	.880	.917	.915
	+/-	actual	.878	.889	.880	.894	.895
		empirical**	.920	.918	.914	.917	.918

Notes: The actual test reliability is computed as squared correlation between true scores and MAP score estimates; the empirical reliability is calculated using equations (27) and (26). (**) For blocks of 3 or 4 items, empirical reliability is computed using a simplifying assumption of local independence.

Table 8

Estimated correlations between the Big Five markers based on the single-stimulus and forced-choice questionnaires in the empirical example, N = 438.

	N	E	O	A	C
Neuroticism (N)	1	-.44 (.04)	-.49 (.04)	-.37 (.05)	-.33 (.05)
Extraversion (E)	-.40 (.06)	1	.52 (.04)	.49 (.04)	.29 (.05)
Openness (O)	-.48 (.07)	.48 (.06)	1	.41 (.05)	.31 (.05)
Agreeableness (A)	-.40 (.08)	.41 (.07)	.15 (.08)	1	.30 (.05)
Conscientiousness (C)	-.30 (.07)	.23 (.07)	.35 (.07)	.31 (.08)	1

Notes: The single-stimulus correlation estimates are above the diagonal, the forced-choice estimates are below the diagonal, the standard errors are in parentheses.

Table 9

Reliabilities and correlations between the single-stimulus and forced-choice Big Five trait scores in the empirical example, N = 438.

	N	E	O	A	C
SS reliability	0.825	0.844	0.824	0.775	0.802
FC reliability	0.704	0.766	0.729	0.601	0.685
corr(SS,FC)	0.804	0.817	0.772	0.692	0.764

Notes: The reliability estimates are computed by the sample-evaluated information method. SS = single-stimulus questionnaire; FC = forced-choice questionnaire.

Figure 1

Thurstonian factor model for a short questionnaire with 3 traits and 3 blocks of 3 items

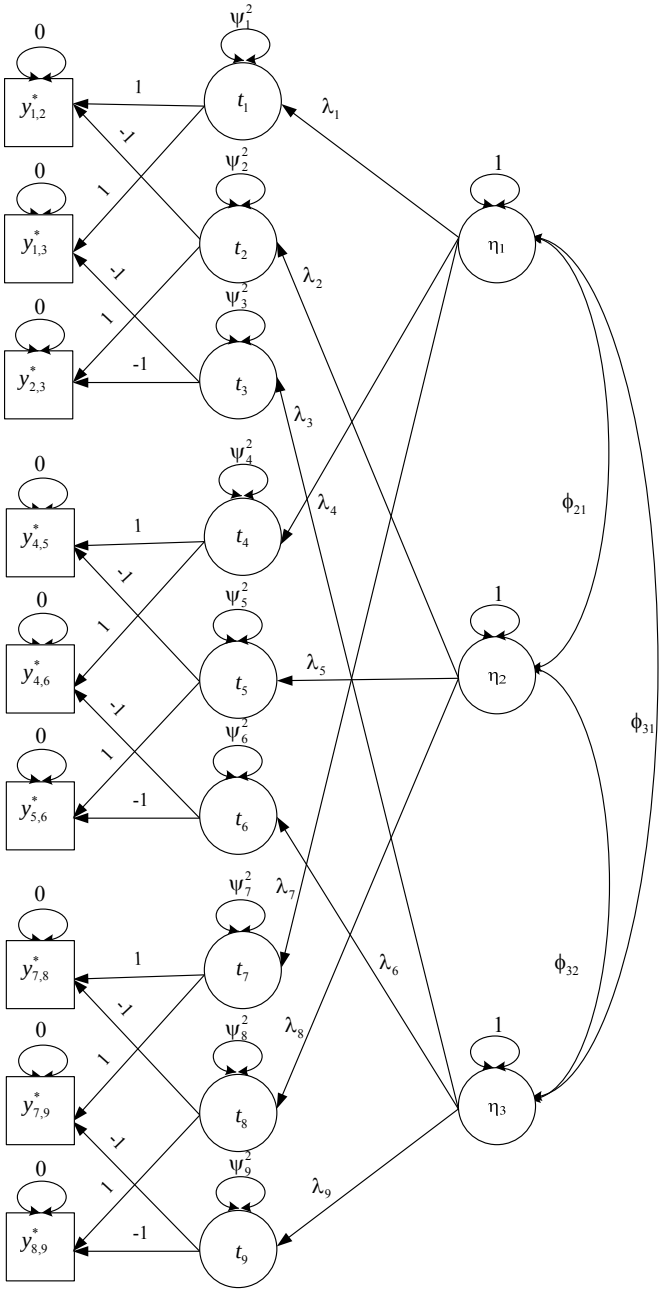


Figure 2

IRT representation of a Thurstonian model for a short questionnaire with 3 traits and 3 blocks of 3 items

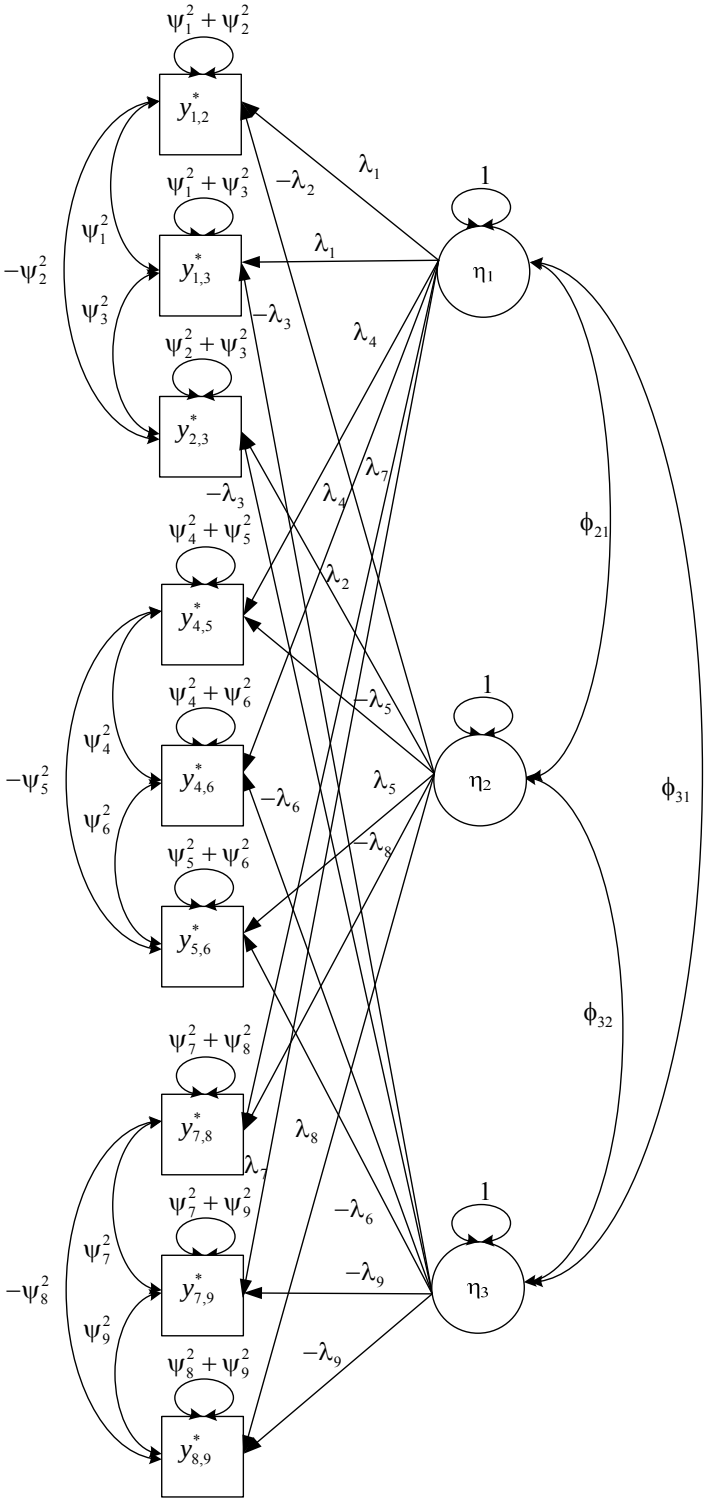
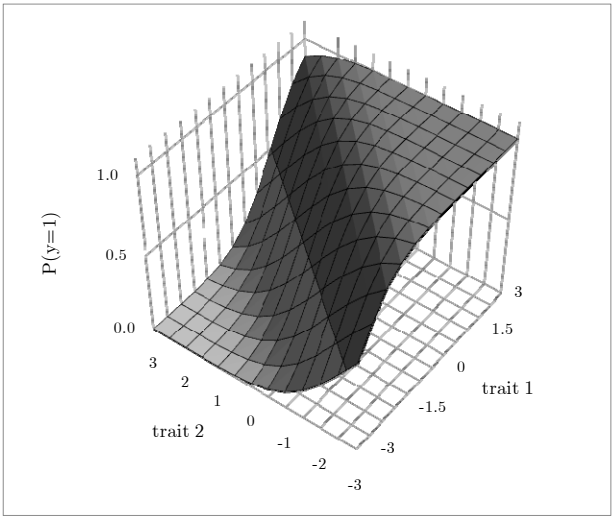


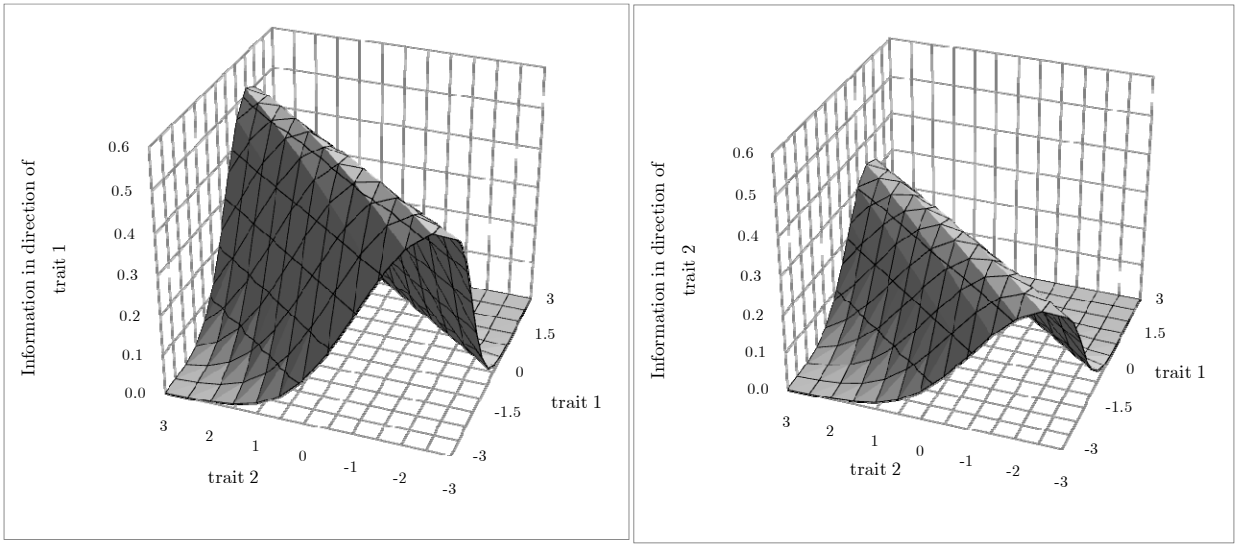
Figure 3

Item Characteristic and Item Information functions for the binary outcome of comparison {i5, i6} for the simulation study with 2 uncorrelated traits.

a. *Item Characteristic Surface (ICS)*



b. *Item Information Surfaces (IIS) in directions of Trait 1 and Trait 2*

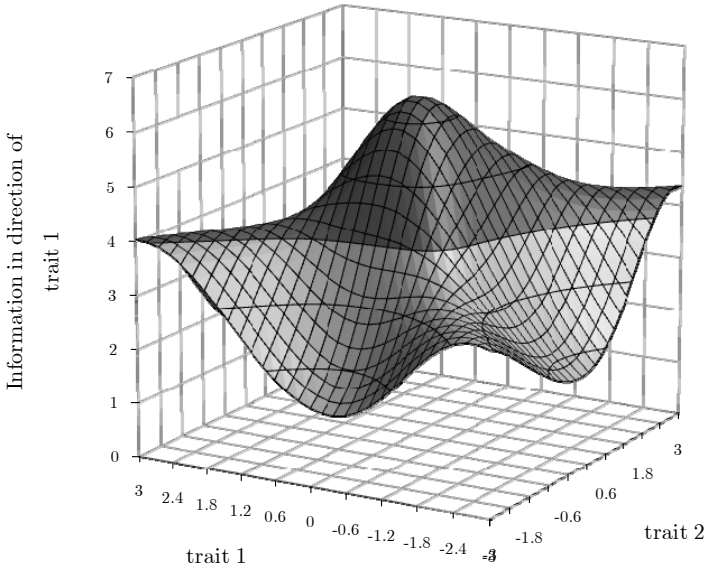


Notes: Item parameters for the binary outcome {i5, i6} in the intercept/slope form: $\alpha = 0.726$; $\beta_1 = 0.902$; $\beta_2 = 0.730$ (see Table 1).

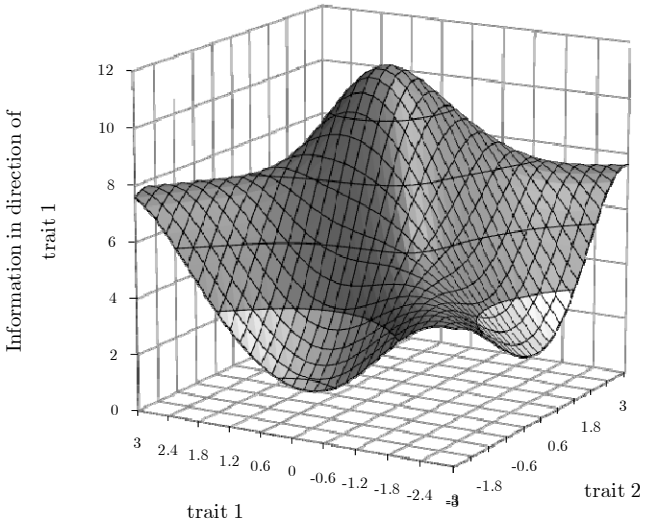
Figure 4

MAP test information function for the simulation study with 2 uncorrelated traits. Information is computed in direction of Trait 1.

a. *Short questionnaire with positively and negatively keyed items*



b. *Long questionnaire with positively and negatively keyed items*

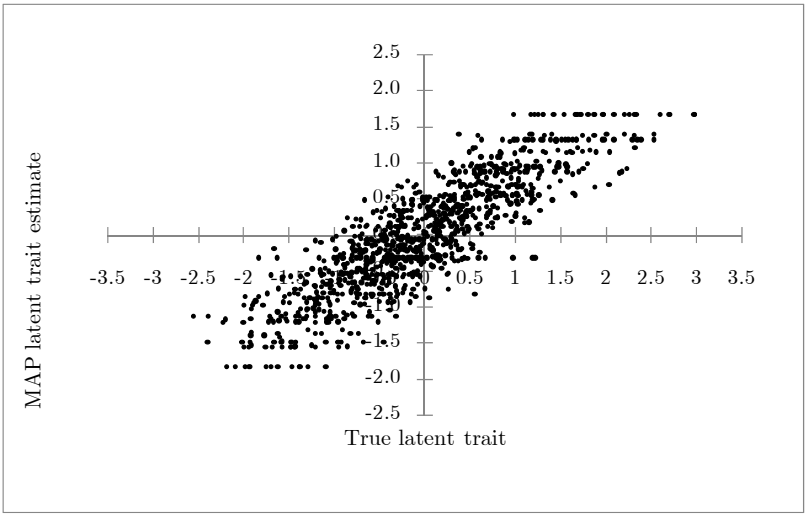


Note: Darker shading on the graphs signifies the information values over 4, corresponding to the test reliability over 0.75.

Figure 5

Scatterplot of MAP estimated trait scores vs. true latent trait scores for the simulation study with 2 uncorrelated traits.

a. *Short questionnaire with positively and negatively keyed items*



b. *Long questionnaire with positively and negatively keyed items*