Kent Academic Repository

Full text document (pdf)

Citation for published version

Maydeu-Olivares, Alberto and Brown, Anna (2010) Item Response Modeling of Paired Comparison and Ranking Data. Multivariate Behavioral Research, 45 (6). pp. 935-974. ISSN 0027-3171.

DOI

https://doi.org/10.1080/00273171.2010.531231

Link to record in KAR

http://kar.kent.ac.uk/29630/

Document Version

Author's Accepted Manuscript

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version.
Users are advised to check **http://kar.kent.ac.uk** for the status of the paper. **Users should always cite the published version of record.**

Enquiries
For any further enquiries regarding the licence status of this document, please contact:
researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at **http://kar.kent.ac.uk/contact.html**

KAR Kent Academic Repository

University of Kent

Item response modeling of paired comparison and ranking data

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona
Anna Brown
SHL Group

This research has been supported by a SMEP dissertation support grant awarded to the second author, and by grants PSI2009-07726 from the Spanish Ministry of Education and SGR 2009 74 from the Autonomous Government of Catalonia awarded to the first author.
Address correspondence to: Alberto Maydeu-Olivares. Faculty of Psychology. University of Barcelona. P. Valle de Hebrón, 171. 08035 Barcelona (Spain). E-mail: amaydeu@ub.edu .

Abstract

The comparative format used in ranking and paired comparisons tasks can significantly reduce the impact of uniform response biases typically associated with rating scales. Thurstone's model provides a powerful framework for modeling comparative data such as paired comparisons and rankings. Although Thurstonian models are generally presented as scaling models, i.e. stimuli-centered models, they can also be used as person-centered models. In this paper, we discuss how Thurstone's model for comparative data can be formulated as Item Response Theory (IRT) models, so that respondents' scores on underlying dimensions can be estimated. Item parameters and latent trait scores can be readily estimated using a widely used statistical modeling program. Simulation studies show that item characteristic curves can be accurately estimated with as few as 200 observations and that latent trait scores can be recovered to a high precision. Empirical examples are given to illustrate how the model may be applied in practice, and to recommend guidelines for designing ranking and paired comparisons tasks in the future.

Keywords: paired comparisons, ranking, preferences, comparative judgment, multidimensional IRT, factor analysis

Item response modeling of paired comparison and ranking data

Presenting items in a single-stimulus fashion, using for instance rating scales often can lead to uniform response biases such as acquiescence and extreme responding (e.g. Van Herk, Poortinga & Verhallen, 2004), or lack of differentiation commonly referred to as 'halo' effects (Murphy, Jako & Anhalt, 1993). One approach to overcome this problem is to model such bias (e.g. Maydeu-Olivares & Coffman, 2006). Another approach is to present test items instead in a comparative, or forced-choice format. This approach can significantly reduce the impact of numerous uniform response biases (Cheung & Chan, 2002). Thurstone's (1927, 1931) model provides a powerful framework for describing the response process to comparative data such as paired comparisons and rankings. Although Thurstonian models are generally presented as scaling models, i.e. stimuli-centered models, they can also be used as person-centered models. For instance, in a ranking task, respondents may be presented with a set of behavioral statements and be asked to order them according to the extent that the statements describe their personality. Or, respondents may be asked to order a set of attitudinal statements according to the extent they represent their own attitudes. In a paired comparison task, pairs of statements are selected from a set of available items, and respondents are instructed to select the item that best describes them from each pair. In these applications, the focus is not on the items under comparison and their relationships, but rather on the individuals' personality traits, attitudes, etc. When used in this fashion, Thurstonian models for comparative data are item response theory (IRT) models (Maydeu-Olivares, 2001). The aim of this paper is to describe the properties and characteristics of Thurstonian models for comparative data as IRT models.

This article is structured into seven sections. In the first section, we describe how to code rankings and paired comparisons using binary outcome variables. This binary coding allows straightforward estimation of models for comparative data using standard statistical

software. Section two describes Thurstonian models for comparative data. In this section we provide the response model for ranking tasks and for paired comparisons tasks. We also describe embedding common factors in these models. Thurtonian factor models are second order normal ogive models with some special features. Section three introduces the Thurstonian IRT model. This is simply a reparameterization of the Thurstonian factor model as a first order model, again with special features. The Thurstonian IRT model provides some valuable insights into the features of Thurstonian models as person-centered models and it enables straightforward estimation of latent trait scores for ranking data, something that is not possible with the Thurstonian factor model. Section four discusses item parameter estimation of Thurstonian models for paired comparisons and rankings. Section five provides a detailed account of the Thurstonian IRT model. In this section we (a) provide the item characteristic function for these models, (b) discuss how to estimate the latent traits. and (b) provide the information function and discuss how to estimate test reliability. Because in today's IRT applications unidimensional models are most often used, in this paper we focus mostly on unidimensional models. Section five reports the results of simulation studies to investigate the accuracy of item parameter estimates and their standard errors, goodness of fit tests, and latent trait scores. The widely used statistical modeling program Mplus (Muthén & Muthén, 2001-2009) is used throughout the paper to estimate the item parameters models and to obtain latent trait scores. Section six includes two applications to illustrate our presentation, one involving ranking data, and one involving paired comparisons data. We conclude with a summary of the main points of this article and a discussion of extensions of the work presented here.

Binary coding of comparative data

This section discusses how to code the observed paired comparison and ranking data in a form suitable for estimating Thurstonian choice models when using standard software

packages for IRT modeling. This section relies heavily on Maydeu-Olivares and Böckenholt (2005).

Paired comparisons

In a paired comparison task, respondents are presented with pairs selected from an item set and are instructed to select the more preferred item from each pair. With n items there are $\tilde{n} = \frac{n(n-1)}{2}$ pairs of items. For instance, $\tilde{n} = 6$ pairs can be constructed with $n = 4$ items. If the $n = 4$ items are labeled {A, B, C, D}, the following pairs can be constructed: {{A,B},{A,C},{A,D},{B,C}, {B,D}, {C,D}}. A presentation of the pairs in this order may result in strong carry-over effects. To control for this effect, it is important to randomize the presentation order of the pairs as well as the order of items within each pair (Bock & Jones, 1968). The observed paired comparison responses can be coded as follows:

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

where l indicates the pair $\{i,k\}$. Thus, we obtain a pattern of $\tilde{n}$ binary responses from each respondent.

Two types of response patterns can be obtained in a paired comparison task, and it is important to distinguish between them. A response pattern consistent with an ordering of the items is called *transitive* pattern, and it is *intransitive* otherwise. As an example of a transitive pattern consider a set of items {A, B, C}. A respondent may choose B when given the pair {A,B}, A when given the pair {A,C}, and B when given the pair {B,C}. These choices are consistent with a {B,A,C} ordering of the items, and the pattern of paired comparisons is said to be transitive. In contrast, an intransitive pattern results when choosing B for the pair {A,B}, A for the pair {A,C}, but C for the pair {B,C}.

Ranking tasks

In a ranking task, all items are presented at once (in a randomized order) and

respondents are asked to either assign ranks or order them. For instance, for the $n = 4$ items {A, B, C, D}, a ranking task consists of assigning ranking positions – numbers from 1 (most preferred) to 4 (least preferred).

Ranking			
A	B	C	D
_	_	_	_

Alternatively, an ordering for the items above is obtained when the ranking positions (1^{st}, 2^{nd}, 3^{rd} and 4^{th}) have to be filled with the given items {A, B, C, D}.

Ordering			
1^{st}	2^{nd}	3^{rd}	4^{th}
_	_	_	_

Any ordering or ranking of n items can be coded equivalently using $\tilde{n}$ paired comparisons. Thus, to continue our example, the ordering {A,D,B,C} (or its equivalent ranking) can be coded using the following paired comparisons:

Ranking				Ordering				Pairwise Outcomes					
A	B	C	D	1^{st}	2^{nd}	3^{rd}	4^{th}	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
1	3	4	2	A	D	B	C	1	1	1	1	0	0

The converse is *not* true because not all paired comparison outcomes can be transformed into rankings or orderings. Intransitive paired comparisons cannot be converted into an ordering of the items. In a paired comparisons task $2^{\tilde{n}}$ binary patterns may be observed but in a ranking task only $n!$ binary patterns may be observed.

In the following, we analyze rankings and orderings after transforming them into binary outcomes. Although both paired comparisons and rankings can be coded using binary outcome variables, we show later that the two data types require slightly different IRT models and that needs to be taken into account in a data analysis.

Thurstonian Models for Ranking and Paired Comparison Data

To model comparative data, such as the data arising from a ranking or paired

comparisons task, Thurstone (1927) proposed the so called Law of Comparative Judgment. 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. Thus, Thurstone's approach may be viewed as a latent variable model where each latent variable corresponds to each of the items (Takane, 1987; Maydeu-Olivares, 2002). Although he focused initially on paired comparisons, Thurstone (1931) recognized later that many other types of choice data, including rankings, could be modeled in a similar way.

Response model for ranking tasks

Consider a random sample of respondents sampled from the population of interest. According to Thurstone (1927, 1931), when a respondent is confronted with a ranking task, each of the n items to be ranked elicits a utility. We shall denote by t_i the utility (a latent variable) associated with item i. Therefore, in Thurstone's model there are exactly n such latent variables when modeling n items. A respondent prefers item i over item k if her or his latent utility for item i is larger than for item k, and consequently ranks item i before item k. Otherwise, he or 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 (2) can be alternatively described by computing differences between the latent utilities. Let

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

be a 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

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

It is convenient to write the response process 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 $\tilde{n} = \dfrac{n(n-1)}{2}$. 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 $\tilde{n}$ paired comparisons. For example, when $n = 2$, $\mathbf{A} = \begin{pmatrix} 1 & -1 \end{pmatrix}$, whereas when $n = 3$, $n = 4$, and $n = 5$

$$\begin{array}{ccc} n = 3 & n = 4 & n = 5 \end{array}$$

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{6}$$

respectively. For instance, in the design matrix for $n = 4$ items, each column corresponds to one of the four items {A, B, C, D}. The corresponding rows give the 6 possible paired comparisons {{A,B},{A,C},{A,D},{B,C}, {B,D}, {C,D}}. Row 4 indicates that B is compared to C; and row 6 indicates that C is compared to D.

Thurstone's model assumes that the utilities $\mathbf{t}$ are normally distributed in the

population of respondents. Thus, we can write

$$\mathbf{t} \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \tag{7}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ denote the mean vector and covariance matrix of the n latent variables $\mathbf{t}$.

When interest lies in scaling the items, two popular models within this class are the so called Case III model, where $\boldsymbol{\Sigma}_t = \boldsymbol{\Psi}^2$, a diagonal matrix, and its special case, the so called Case V model, where $\boldsymbol{\Sigma}_t = \psi^2\mathbf{I}$. However, when interest lies in assessing respondents, items serve as indicators of some latent factors (personality traits, motivation factors, attitudes etc.). Therefore we need to take an extra step and express the latent variables $\mathbf{t}$ as indicators of a set of m common factors (latent traits):

$$\mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{8}$$

In this equation, $\boldsymbol{\mu}_t$ contains the n means of the latent variables $\mathbf{t}$ (i.e., the utilities' means), $\boldsymbol{\Lambda}$ is an $n \times m$ matrix of factor loadings, $\boldsymbol{\eta}$ is an m-dimensional vector of common factors (latent traits, in IRT terminology), and $\boldsymbol{\varepsilon}$ is an n-dimensional vector of unique factors. This factor model assumes that the common factors have mean zero, unit variance and are possibly correlated (their correlation matrix is $\boldsymbol{\Phi}$). The model also assumes that the unique factors have mean zero and are uncorrelated, so that their covariance matrix, $\boldsymbol{\Psi}^2$, is diagonal. In concordance with the distributional assumptions of Thurstonian choice models, the common and unique factors are assumed to be normally distributed.

Response model for paired comparison tasks

In a paired comparison task, respondents need not be consistent in their pairwise choices, possibly yielding intransitive patterns. Inconsistent pairwise responses can be accounted for by adding an error term e_l to the difference judgment (3),

$$y_l^* = t_i - t_k + e_l. \tag{9}$$

This random error e_l is assumed to be normally distributed with zero mean and variance ω_l^2,

uncorrelated across pairs, and uncorrelated with the latent utilities. The error term accounts for intransitive responses by reversing the sign of the difference between the utilities t_i and t_k. For example, suppose that for a given respondent, $t_i = 3$ and $t_k = 2$. Then, whenever $e_l \leq 1$, $y_l^* \geq 0$ and the respondent will choose item i over item k. But if $e_l > 1$, $y_l^* < 0$ and he/she will choose item k over item i, resulting in an intransitivity because $t_i > t_k$.

As in the case of ranking data, the relationship between the observed comparative response y_l and the latent difference judgment y_l^* is given by Equation (4). Similarly, the response process can be written in matrix form as

$$\mathbf{y}^* = \mathbf{A}\,\mathbf{t} + \mathbf{e}\,, \tag{10}$$

where $\mathbf{e}$ is a $\tilde{n} \times 1$ vector of random errors with covariance matrix $\mathbf{\Omega}^2$ which is a diagonal matrix with elements $\omega_1^2, \cdots, \omega^2$.

When the common factor model (8) is embedded in Equation (10) we obtain

$$\mathbf{y}^* = \mathbf{A}\left(\boldsymbol{\mu}_t + \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}\right) + \mathbf{e}\,. \tag{11}$$

Also, the mean vector and covariance matrix of the latent differences $\mathbf{y}^*$ are

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t\,, \text{ and } \qquad \mathbf{\Sigma}_{y^*} = \mathbf{A}\left(\mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}^2\right)\mathbf{A}' + \mathbf{\Omega}^2\,. \tag{12}$$

The model for ranking data can be seen as a special case of the model for paired comparisons. The smaller the diagonal elements of the error covariance matrix $\mathbf{\Omega}^2$, the more consistent the respondents are in evaluating the items. In the extreme case, when all the diagonal elements of $\mathbf{\Omega}^2$ are zero, no intransitivities would be observed in the data and the paired comparison data are effectively rankings. A more restricted model that is often found to be useful in applications involves setting the error variances to be equal for all pairs (i.e., $\mathbf{\Omega}^2 = \omega^2\mathbf{I}$) This restriction implies that the number of intransitivities is approximately equal for all pairs provided the elements of $\boldsymbol{\mu}_t$ are not too dissimilar (Maydeu-Olivares & Böckenholt, 2005).

Thresholds and tetrachoric correlations implied by the model

Because all random variables ($\boldsymbol{\eta}$, $\boldsymbol{\varepsilon}$, and $\mathbf{e}$) are normally distributed, the latent difference responses $\mathbf{y}^*$ are also normally distributed. Since the outcome binary variables $\mathbf{y}$ are obtained by dichotomizing the $\mathbf{y}^*$ variables, the correlations among the $\mathbf{y}^*$ variables are tetrachoric correlations.

To obtain the tetrachoric correlations implied by Thurstone's model we standardize the latent difference responses $\mathbf{y}^*$using

$$\mathbf{z}^* = \mathbf{D}\left(\mathbf{y}^* - \boldsymbol{\mu}_{y^*}\right), \qquad \mathbf{D} = \left(\mathrm{Diag}\left(\boldsymbol{\Sigma}_{y^*}\right)\right)^{-\frac{1}{2}}, \tag{13}$$

where $\mathbf{z}^*$ are the standardized latent difference responses and $\mathbf{D}$ is a diagonal matrix with the reciprocals of the model implied standard deviations of $\mathbf{y}^*$ in the diagonal. The standardized latent difference responses are multivariate normal with a $\mathbf{0}$ mean vector and tetrachoric correlation matrix $\mathbf{P}_{z^*}$, where

$$\mathbf{P}_{z^*} = \mathbf{D}\left(\boldsymbol{\Sigma}_{y^*}\right)\mathbf{D}. \tag{14}$$

Using (12), in the special case where a common factor model is assumed to underlie the utilities, (14) becomes

$$\mathbf{P}_{z^*} = \mathbf{D}\left(\boldsymbol{\Sigma}_{y^*}\right)\mathbf{D} = \mathbf{D}\left(\mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right)\mathbf{A}' + \boldsymbol{\Omega}^2\right)\mathbf{D}. \tag{15}$$

The standardized latent difference responses $\mathbf{z}^*$ are related to the observed comparative responses $\mathbf{y}$ via the threshold relationship

$$y_l = \begin{cases} 1 & \text{if} \quad z_l^* \geq \tau_l \\ 0 & \text{if} \quad z_l^* < \tau_l \end{cases} \tag{16}$$

where the $\tilde{n} \times 1$ vector of thresholds $\boldsymbol{\tau}$ has the following structure (Maydeu-Olivares & Böckenholt, 2005)

$$\boldsymbol{\tau} = -\mathbf{D}\boldsymbol{\mu}_{y^*} = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}_t. \tag{17}$$

Identification of Thurstonian factor models for comparative data

Identification restrictions for these models were given by Maydeu-Olivares and Böckenholt (2005) and they are the same for ranking and paired comparisons models. Consider an unrestricted (aka exploratory) factor model. It is well known (e.g. McDonald,1999: p. 181) that this model applied to continuous data can be identified by setting the factors to be uncorrelated and by setting the upper triangular part of the factor loading matrix equal 0. This amounts to setting $\lambda_{ij} = 0$ for all such i and j that $i = 1, \ldots, m - 1$; $j = i + 1, \ldots, m$. For example, with these constraints the factor loading matrix for a three factor model has the following form

$$\mathbf{\Lambda}_t = \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \vdots & \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n3} \end{pmatrix}. \tag{18}$$

The resulting solution can then be rotated (orthogonally or obliquely) to obtain a more interpretable solution.

For Thurstonian factor models additional constraints are needed to obtain the initial solution because of the comparative nature of the data. Thus, in addition to the constraints on the loading matrix given by the pattern (18) Maydeu-Olivares and Böckenholt (2005) suggested (a) fixing all factor loadings involving the last item to 0, $\lambda_{ni} = 0$, $i = 1, \ldots, m$; and (b) fixing the unique variance of the last item to one, $\psi_n^2 = 1$. These identification constraints define the scales of the factor loadings, and the unique factor variances, respectively. As an illustration, the identification restrictions needed to estimate a Thurstonian two factor model for paired comparisons and ranking data are

$$\mathbf{\Lambda}_t = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \vdots & \vdots \\ \lambda_{n-1,1} & \lambda_{n-1,2} \\ 0 & 0 \end{pmatrix}, \text{ and } \mathbf{\Psi}_t^2 = \begin{pmatrix} \psi_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \psi_{n-1}^2 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{19}$$

The necessary identification constraints imply that at least $n = 5$, 6, 8 and 9 items are required to estimate Thurstonian factor models with $m = 1$, 2, 3 and 4 common factors in both paired comparisons and ranking data. Factor models with smaller number of items can also be estimated, but additional constraints are needed to estimate them.

Regarding the means of the utilities, $\boldsymbol{\mu}_t$, these parameters can be estimated by fixing one of the means to some constant, for instance $\mu_n = 0$.

Thurstonian Models for Ranking and Paired Comparison Data as IRT models

In the previous section, we showed that Thurstonian factor models for ranking and paired comparisons data are indeed a second order factor model for binary data with some special features: a) the number of first order factors $\mathbf{t}$ is fixed by design; it is n, the number of items; b) the first order factor loading matrix, $\mathbf{A}$, is a matrix of constants, see (6); c) the uniquenesses of the first order factors can be estimated (except for one) because the first order factor loading matrix is a matrix of constants; d) one row of the second order factor matrix needs to be fixed to identify the model – see (19); d) the first order factor means may be estimated (these are the items' means, or in Thurstonian terms, the mean utilities); and e) if the binary outcomes arise from a ranking experiment, the uniquenesses of the latent response variables must be fixed to zero.

Because factor models for binary data are equivalent to the normal ogive IRT model (see Takane & de Leeuw, 1987), in this section we exploit this relationship and present Thurstonian models for comparative data as IRT models. First, we introduce a Thurstonian factor model with unconstrained thresholds that it is likely to yield a better fit in applications. Then, we show how the Thurstonian factor model (which is a second-order

model) can be equivalently expressed as a first-order model with structured correlated errors. We refer to this model as the Thurstonian IRT model.

Thurstonian factor models with unrestricted thresholds (unrestricted intercepts)

Recall that Thurstonian factor models are defined by equations (8) and (10), which we repeat here for convenience

$$\mathbf{y}^* = \mathbf{A}\,\mathbf{t} + \mathbf{e}\,, \qquad\qquad \mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,, \tag{20}$$

where for ranking data $\mathbf{e} = \mathbf{0}$, and recall that the n parameters $\boldsymbol{\mu}_t$ are the means of the utilities, i.e., the means of the latent variables underlying each item. In IRT applications, the utilities $\mathbf{t}$ (and in particular, the parameters $\boldsymbol{\mu}_t$) will be seldom of interest. Rather, in IRT applications, the main focus in on estimating the latent traits $\boldsymbol{\eta}$. When the mean utilities are not of interest, we can use instead of (20)

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}\,\mathbf{t} + \mathbf{e}\,, \qquad\qquad \mathbf{t} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,. \tag{21}$$

Model (21) is a Thurstonian factor model with unrestricted intercepts. The original model – given by (20)– is simply a constrained version of (21) where the $\tilde{n}$ intercepts $-\boldsymbol{\gamma}$ are constrained to be a function of the n parameters $\boldsymbol{\mu}_t$,

$$\boldsymbol{\gamma} = -\mathbf{A}\boldsymbol{\mu}_t = -\boldsymbol{\mu}_{y^*}\,. \tag{22}$$

That is, the intercepts are also the means of the latent difference judgments $\mathbf{y}^*$ with a sign change. We refer to model (21) as a Thurstonian factor model with unrestricted thresholds because for this model the threshold structure (17) becomes

$$\boldsymbol{\tau} = \mathbf{D}\boldsymbol{\gamma}\,. \tag{23}$$

Thus, the threshold structure $\boldsymbol{\tau}$ becomes unconstrained since $\boldsymbol{\gamma}$ is simply a re-scaling of $\boldsymbol{\tau}$ by the matrix $\mathbf{D}$.

In applications where the parameters $\boldsymbol{\mu}_t$ are not of interest, we recommend fitting Thurstonian models with unrestricted thresholds (21) as it leads to a considerably less

constrained model.

Thurstonian IRT model for comparative data

If indeed the latent utilities $\mathbf{t}$ are not of interest, as in most typical IRT applications, we can go one step further and reparameterize the Thurstonian factor model with unrestricted thresholds as a first order factor model so that the latent utilities $\mathbf{t}$ effectively disappear from the model:

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\right) + \mathbf{e} = -\boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\boldsymbol{\gamma} + \breve{\boldsymbol{\Lambda}}\boldsymbol{\eta} + \breve{\boldsymbol{\varepsilon}}\,. \tag{24}$$

with $\breve{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}$, and $\operatorname{cov}\left(\breve{\boldsymbol{\varepsilon}}\right) = \breve{\boldsymbol{\Psi}}^2$, where

$$\breve{\boldsymbol{\Lambda}} = \mathbf{A}\boldsymbol{\Lambda}\,, \qquad\qquad \breve{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \boldsymbol{\Omega}^2\,, \tag{25}$$

are a $\tilde{n} \times m$ matrix and a $\tilde{n} \times \tilde{n}$ matrix, respectively.

However, the matrices $\breve{\boldsymbol{\Lambda}}$ and $\breve{\boldsymbol{\Psi}}^2$ are patterned. For instance in the case of $n = 3$, and assuming a single latent trait

$$\breve{\boldsymbol{\Lambda}} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_3 \\ \lambda_2 - \lambda_3 \end{pmatrix}, \tag{26}$$

and $\breve{\boldsymbol{\Psi}}^2$ is patterned as follows: For $n = 3$

$$\breve{\boldsymbol{\Psi}}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2 + \psi_3^2 + \omega_3^2 \end{pmatrix}, \tag{27}$$

whereas for $n = 4$

$$\breve{\boldsymbol{\Psi}}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & & & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & & & & \\ \psi_1^2 & \psi_1^2 & \psi_1^2 + \psi_4^2 + \omega_3^2 & & & \\ -\psi_2^2 & \psi_3^2 & 0 & \psi_2^2 + \psi_3^2 + \omega_4^2 & & \\ -\psi_2^2 & 0 & \psi_4^2 & \psi_2^2 & \psi_2^2 + \psi_4^2 + \omega_5^2 & \\ 0 & -\psi_3^2 & \psi_4^2 & -\psi_3^2 & \psi_4^2 & \psi_3^2 + \psi_4^2 + \omega_6^2 \end{pmatrix}, \tag{28}$$

where recall that all $\omega_l^2 = 0$ in the case of ranking data. Notice that $\breve{\boldsymbol{\Psi}}^2$ is not a diagonal

matrix and that its pattern does not depend on the number of latent traits, but on the number of items. Also, $\breve{\boldsymbol{\Psi}}^2$ is not of full rank. Its rank is the same as the rank of $\mathbf{A}$, $n - 1$.

We refer to model (24) with the constraints (25) as the Thurstonian IRT model for comparative data. It is simply a reparameterization of the Thurstonian factor model with unrestricted thresholds. Both models are equivalent. They have the same number of parameters and lead to the same threshold structure –given by (23)– and model implied tetrachoric correlation matrix $\mathbf{P}_{z^*}$:

$$\mathbf{P}_{z^*} = \mathbf{D}\left(\mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'\mathbf{A}' + \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \boldsymbol{\Omega}^2\right)\mathbf{D} = \mathbf{D}\left(\breve{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\breve{\boldsymbol{\Lambda}}' + \breve{\boldsymbol{\Psi}}^2\right)\mathbf{D}\,. \tag{29}$$

The Thurstonian factor model is a second order factor model, where the first order factors are the latent utilities, and the second order factors are the latent traits. As a result, in this model, there are $n + m$ latent variables. In contrast, the Thurstonian IRT model is a first order factor model involving only m latent variables, the latent traits.

Item parameter estimation of Thurstonian models for paired comparisons and rankings

IRT models are most often estimated using full information maximum likelihood (FIML – often referred to in the IRT literature as marginal maximum likelihood, see Bock & Aitkin, 1981). To obtain parameter estimates using FIML, the probabilities of observing each response pattern are obtained by integrating the product of the item characteristic curves (ICCs) over the density of the latent traits, assuming local independence. For the models under consideration, this assumption does not need to hold. Consider the joint covariance matrix of $\mathbf{y}^*$, $\mathbf{t}$, and $\boldsymbol{\eta}$. This is

$$\operatorname{cov}\left(\mathbf{y}^*, \mathbf{t}, \boldsymbol{\eta}\right) = \begin{pmatrix} \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right)\mathbf{A}' + \boldsymbol{\Omega}^2 & \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right) & \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\Phi} \\ & \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2 & \boldsymbol{\Lambda}\boldsymbol{\Phi} \\ & & \boldsymbol{\Phi} \end{pmatrix}. \tag{30}$$

From (30), we obtain

$$\operatorname{cov}\left(\mathbf{y}^{*}\,\middle|\,\mathbf{t},\boldsymbol{\eta}\right)=\boldsymbol{\Omega}^{2}\,, \tag{31}$$

$$\operatorname{cov}\left(\mathbf{y}^{*}\,\middle|\,\boldsymbol{\eta}\right)=\mathbf{A}\boldsymbol{\Psi}^{2}\mathbf{A}'+\boldsymbol{\Omega}^{2}\equiv\breve{\boldsymbol{\Psi}}^{2}\,. \tag{32}$$

Equation (32) reveals that the latent difference responses $\mathbf{y}^{*}$ are not independent when conditioning only on the latent traits, regardless of whether paired comparisons or ranking data is involved, because $\breve{\boldsymbol{\Psi}}^{2}$ is not a diagonal matrix. On the other hand, equation (31) reveals that the latent difference responses $\mathbf{y}^{*}$ are independent when conditioning on the utilities and latent traits for paired comparisons data (by the diagonal assumption on $\boldsymbol{\Omega}^{2}$). For ranking data, where $\boldsymbol{\Omega}^{2}=\mathbf{0}$, conditioning on both the utilities and the latent traits leads to a degenerate distribution (see Maydeu-Olivares, 2001, p. 215).

This implies that in Thurstonian factor models, where both the n latent utilities $\mathbf{t}$ and the m latent traits $\boldsymbol{\eta}$ are involved, the ICCs are conditionally independent, but to estimate this model by FIML $n+m$ dimensional integration is needed. It is well known that FIML is only computationally feasible unless a very few latent dimensions are involved. In practice, FIML is seldom performed with more than three latent dimensions. On the other hand, in Thurstonian IRT models, where only the m latent traits $\boldsymbol{\eta}$ are involved, the ICCs are conditionally dependent. If standard FIML estimation is used (i.e., assuming local independence), only m dimensional integration is needed, but it would result in biased estimates because of the violation of the local independence assumption. Thus, FIML estimation is ill suited to estimate either model.

Fortunately, the item parameters of Thurstonian models can be straightforwardly estimated using limited information methods as follows. First, the sample thresholds $\hat{\boldsymbol{\tau}}$ and the sample tetrachoric correlations $\hat{\boldsymbol{\rho}}$ are estimated. Then, the item parameters of the model are estimated from the first stage estimates by unweighted least squares (ULS: Muthén, 1993) or diagonally weighted least squares (DWLS: Muthén, du Toit & Spisic, 1997). Limited

information methods and FIML yield very similar IRT parameter estimates and standard errors (Forero & Maydeu-Olivares, 2009). Also, differences between using ULS or DWLS in the second stage of the estimation procedure are negligible (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009). Furthermore, a test of the restrictions imposed on the thresholds and tetrachoric correlations is available, with degrees of freedom equal to the number of thresholds plus the number of tetrachoric correlations, $\tilde{n}(\tilde{n}+1)/2$, minus the number of estimated item parameters (say q).

However, care is needed when testing the model with ranking data. This is because Maydeu-Olivares (1999) showed that when ranking data is used, there are

$$r = n(n-1)(n-2)/6 \tag{33}$$

redundancies among the thresholds and tetrachoric correlations estimated from the binary outcome variables. Hence, the correct number of degrees of freedom when modeling ranking data is $df = \tilde{n}(\tilde{n}+1)/2 - r - q$. This means that the p-value for the chi-square test statistic needs to be recomputed using the correct number of degrees of freedom. Also, goodness of fit indices involving degrees of freedom in their formula, such as the $RMSEA = \sqrt{\frac{T-df}{df \times N}}$, where T denotes the chi-square statistic and N denotes sample size, also need to be recomputed using the correct degrees of freedom for ranking data.

The Thurstonian IRT model

In this section, we provide the item characteristic and information function for the model and discuss item parameter estimation, latent trait estimation, and reliability estimation. We conclude this section providing some remarks about the impact of the choice of identification constraints on item parameter estimates.

Item characteristic function (ICC)

The ICC for binary outcome variable y_l involving items i and k is

$$\Pr\left(y_l = 1 \middle| \boldsymbol{\eta}\right) = \Phi\left(\frac{-\gamma_l + \breve{\boldsymbol{\lambda}}_l{}'\boldsymbol{\eta}}{\sqrt{\breve{\psi}_l^2}}\right), \tag{34}$$

where $\Phi\left(x\right)$ denotes a standard normal distribution function evaluated at x, , γ_l is the threshold for binary outcome y_l, $\breve{\boldsymbol{\lambda}}_l{}'$ is the $1 \times m$ vector of factor loadings, and $\breve{\psi}_l^2$ is the uniqueness for binary outcome y_l.

Equation (34) is simply the ICC of a normal ogive model for binary data except that (a) $\breve{\boldsymbol{\lambda}}_l{}'$ is structured, (b) $\breve{\psi}_l^2$ is structured, and (c) the ICCs are not independent (local independence conditional on the latent traits does not hold). Rather, there are patterned covariances among the unique factors, see (27) and (28) for the case of three and four items, respectively.

Indeed, when only a single trait is involved the ICC for Thurstonian IRT models can be written using (26) and (27) as

$$\Pr\left(y_l = 1 \middle| \eta\right) = \Phi\left(\frac{-\gamma_l + \breve{\lambda}_l \eta}{\sqrt{\breve{\psi}_l^2}}\right) = \Phi\left(\frac{-\gamma_l + \left(\lambda_i - \lambda_k\right)\eta}{\sqrt{\psi_i^2 + \psi_k^2 + \omega_l^2}}\right). \tag{35}$$

With n items being compared, $\tilde{n}$ binary outcome variables are used, and the number of parameters being estimated is $\tilde{n}$ thresholds γ_l, $n-1$ factor loadings λ_i, $n-1$ uniquenesses ψ_i^2 , and $\tilde{n}$ paired-specific error variances ω_l^2. Models for ranking data involve $\tilde{n}$ fewer parameters as $\omega_l^2 = 0$ for all variables. This corresponds to a model with unrestricted thresholds. A model with restrictions on the threshold structure amounts to setting $\gamma_l = -\mu_i + \mu_k$ for all binary outcome variables. Thus, $n-1$ item means μ_i are estimated instead of the $\tilde{n}$ thresholds γ_l,

Equation (35) expresses the model using a threshold/factor loading parameterization. Letting

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_k^2 + \omega_l^2}} \qquad \text{and} \qquad \beta_l = \frac{\lambda_i - \lambda_k}{\sqrt{\psi_i^2 + \psi_k^2 + \omega_l^2}} \tag{36}$$

the ICC for unidimensional Thurstonian IRT models can be written in an intercept α_l and slope β_l form as

$$\Pr\left(y_l = 1 \middle| \eta\right) = \Phi\left(\alpha_l + \beta_l \eta\right). \tag{37}$$

Note that α_l and β_l are not standardized parameters since $\breve{\psi}_l^2 = \psi_i^2 + \psi_k^2 + \omega_l^2$ is not the variance of y_l^*. Also, note that the $\tilde{n}$ intercepts and $\tilde{n}$ slopes are not free parameters to be estimated. Rather, they are functions of the fundamental parameters of the model (thresholds, factor loadings, uniquenesses and paired-specific error variances).

Latent trait estimation, information functions and reliability estimation

After the item parameters have been estimated, latent trait scores can be estimated by treating the estimated parameters as if they were known. This is reasonable if item parameters have been accurately estimated. One approach to estimate the latent trait scores is by maximum likelihood (ML). Two other alternative approaches are a) computing the mean of the posterior distribution of the latent traits, and b) computing the mode of that distribution. The former is known as expected a posteriori (EAP) estimation, and the latter maximum a posteriori (MAP) estimation (see Bock & Aitkin, 1981). Here, we focus on the MAP estimator, as it is the method implemented in the software used throughout this paper, Mplus. In passing, we also provide results for the ML estimator.

Now, recall that in Thurstonian models, the latent traits $\boldsymbol{\eta}$ are assumed to be normally distributed with mean zero, i.e., $\boldsymbol{\mu}_{\boldsymbol{\eta}} = \mathbf{0}$, and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Phi}$, a correlation matrix, and let $P_l\left(\boldsymbol{\eta}\right) = \Pr\left(y_l = 1 \middle| \boldsymbol{\eta}\right)$. For normally distributed traits and assuming local independence, MAP scores can be obtained by minimizing

$$F\left(\boldsymbol{\eta}\right) = \frac{1}{2}\left(\boldsymbol{\eta} - \boldsymbol{\mu}_{\boldsymbol{\eta}}\right)' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}\left(\boldsymbol{\eta} - \boldsymbol{\mu}_{\boldsymbol{\eta}}\right) - \sum_{l=1} P_l\left(\boldsymbol{\eta}\right)^{y_l}\left(1 - P_l\left(\boldsymbol{\eta}\right)\right)^{1-y_l} \tag{38}$$

whereas ML scores are obtained by simply minimizing the second term in (38). In what follows, we just consider an IRT model with a single trait, in which case, (38) simplifies to

$$F\left(\eta\right)=\frac{1}{2}\eta^{2}-\sum_{l=1}P_{l}\left(\eta\right)^{y_{l}}\left(1-P_{l}\left(\eta\right)\right)^{1-y_{l}} \tag{39}$$

The standard error of the ML latent trait estimate is given by $SE\left(\hat{\eta}\right)=\sqrt{1/\mathcal{I}^{-1}\left(\hat{\eta}\right)}$ a function that depends on the latent trait. $\mathcal{I}\left(\eta\right)$ denotes the test information function, which, under local independence, can be written as the sum of the item information functions, i.e. $\mathcal{I}\left(\eta\right)=\sum_{l}\mathcal{I}_{l}\left(\eta\right)$.

In turn, the ML item information for the binary outcome l is obtained as

$$\mathcal{I}_{l}\left(\eta\right)=\frac{\left[P_{l}'\left(\eta\right)\right]^{2}}{P_{l}\left(\eta\right)\left[1-P_{l}\left(\eta\right)\right]}=\frac{\left[\beta_{l}\phi\left(\alpha_{l}+\beta_{l}\eta\right)\right]^{2}}{\Phi\left(\alpha_{l}+\beta_{l}\eta\right)\left[1-\Phi\left(\alpha_{l}+\beta_{l}\eta\right)\right]} \tag{40}$$

where $P_{l}'\left(\eta\right)$ denotes the derivative of (35) with respect to the latent trait η, and $\phi\left(z\right)$ denotes a standard normal density function evaluated at z.

Equation (40) shows that the information provided by an item depends on the magnitude of the slope β_{l} but equation (36) reveals that, for one-dimensional models, the slope β_{l} linearly depends on the difference between the factor loadings λ_{i} and λ_{k} of the two items involved in the comparison. Also, the slope β_{l} will be higher the smaller the ψ_{i}^{2} and ω_{l}^{2} parameters. But when factor loadings λ_{i} and λ_{k} are similar, the slope β_{l} will be close to zero, and the binary outcome will not discriminate well among respondents. In applications, unless items are chosen so that the loadings λ_{i} vary widely in their magnitudes, the item slopes in the one-dimensional Thurstonian IRT model are likely to be low in applications and a large number of items will be needed to accurately estimate the latent trait. Equation (36) also reveals that whenever $\lambda_{i}<\lambda_{k}$, the slope β_{l} will be negative for one-dimensional models. Thus, in applications negative estimates for β_{l} will be commonly found. However, it is the magnitude of the slope parameters β_{l} that matters, not their sign.

Now, the standard error of the MAP latent trait estimate is given by

$$SE\left(\hat{\eta}\right) = \sqrt{1 / \mathcal{I}_P^{-1}\left(\hat{\eta}\right)} \tag{41}$$

where $\mathcal{I}_P\left(\eta\right)$ denotes the test information function of the posterior distribution of the latent trait. For a single latent trait, which is assumed to be normally distributed with mean zero and variance 1, the MAP test information function is

$$\mathcal{I}_p\left(\eta\right) = \mathcal{I}\left(\eta\right) + \frac{\partial^2 \phi\left(\eta\right)}{\partial \eta^2} = \mathcal{I}\left(\eta\right) + 1\,. \tag{42}$$

In applications, it may be convenient to offer a single index of the precision of measurement of the latent trait instead of the standard error function (41), which is a function of the latent trait. Provided the squared standard error function is relatively uniform, a single index of the precision of measurement can be obtained using the reliability coefficient (e.g. Bock, 1997)

$$\rho = \frac{\sigma^2 - \overline{\sigma}^2_{error}}{\sigma^2}\,. \tag{43}$$

There are two ways to estimate this coefficient.

One way, referred to as *theoretical reliability* (du Toit, 2003) involves estimating the average error of measurement as

$$\overline{\sigma}^2_{error} = \int_{-\infty}^{\infty} \mathcal{I}_P^{-1}\left(\eta\right)\phi\left(\eta\right)d\eta\,, \tag{44}$$

and using $\sigma^2 = 1$ in (43) as this the assumed value for the variance of the latent trait. In the case of multiple traits, this procedure becomes unattractive since it involves integrating a multivariate normal distribution.

An alternative way to estimate (43), referred to as *empirical reliability*, involves estimating σ^2 using the sample variance of the estimated MAP scores, and estimating $\overline{\sigma}^2_{error}$ using the mean of the squared standard errors of the estimated MAP scores. That is, given a sample of N respondents, and letting $\hat{\eta}_j$ be the estimated MAP score for respondent j, we

compute

$$\hat{\sigma}^2 = \frac{1}{N}\sum_j \left(\hat{\eta}_j - \bar{\hat{\eta}}\right)^2, \qquad \hat{\bar{\sigma}}^2_{error} = \frac{1}{N}\sum_j \left(SE\left(\hat{\eta}_j\right)\right)^2 = \frac{1}{N}\sum_j \frac{1}{\mathcal{I}_P\left(\hat{\eta}_j\right)}. \quad (45)$$

In our experience, for long tests, the theoretical and empirical reliabilities are quite close to each other. In short tests, MAP estimates may shrink towards the mean, and $\hat{\sigma}^2$ computed using (45) may be low, in which case the empirical estimate will underestimate the reliability.

In either case, given the estimated reliability, we can estimate the correlation between the true latent trait and the estimated scores using

$$\mathrm{corr}\left(\eta, \hat{\eta}\right) = \sqrt{\rho}\ . \quad (46)$$

In closing this subsection, we emphasize that the above standard results for unidimensional IRT models do not hold if local independence does not hold. In particular, when local independence does not hold the test information cannot be decomposed into the sum of item information functions. Thus, we shall investigate the extent to which the above expressions (using the simplifying assumption that the ICCs of Thurstonian IRT models are locally independent) provide a sufficiently accurate approximation in applications. Note that this simplifying assumption is only employed for latent trait estimation, not for item parameter estimation.

Some remarks about parameterizations and the choice of identification constraints

Here we have followed Maydeu-Olivares and Böckenholt's (2005) suggestions regarding the choice of identification constraints, perhaps the most striking of which is to fix one of the factor loadings to zero. In this subsection we examine the implications of these identification choices. For ease of exposition, we focus on a set of items that substantively are assumed to be positively related to a single latent trait.

Statistically, the choice of identification constraints has no impact whatsoever. In the

previous subsection we have shown that it is the intercepts and slopes (i.e., the ICC) which govern item information, and consequently latent trait recovery. Intercepts and slopes are invariant to the choice of identification constraints. This is shown in Appendix A.

Substantively, it is unappealing to fix a factor loading to zero because it suggests that one particular item is unrelated to the latent trait. From this point of view, it may be better to fix one of the loadings to 1 instead, or to estimate all loadings using a sum constraint (e.g., $\sum_i \lambda_i = 1$) which would enable computing standard errors for all loadings. We prefer to fix a factor loading because it is easier to implement, to remind researchers that there is a constraint among the loadings, and because using a sum constraint will lead to some factor loadings to be negative. If one factor loading is fixed to some constant for identification some factor loading estimates may be negative as well. If item n is fixed for identification and a negative factor loading for item i is obtained, this indicates that the absolute value of λ_i is smaller than λ_n. It should not be interpreted as a negative relationship between item i and the trait. With comparative data, the usual interpretation of the signs of factor loadings does not hold. This is because when comparative data is modeled, the scale origin is arbitrary (Böckenholt, 2004), and there are many sets of thresholds and factor loadings that are consistent with any given model and a researcher is free to choose the most substantively meaningful model among the set of equivalent models (Maydeu-Olivares & Hernández, 2007). In fact, one can change the signs of one or more factor loadings to ease the interpretation of the model according to the substantive theory simply by changing the identification constraints. The formula presented in Appendix A can be used to explore the set of thresholds and factor loadings that are equivalent to those estimated in a given application. The important point is that the chosen constraints will not alter the binary outcomes' intercepts and slopes.

Simulation studies

It is of interest to know how well the fundamental parameters of the Thurstonian IRT model ($\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, ψ^2, and in the case of paired comparisons models $\boldsymbol{\omega}^2$) can be estimated. These parameters are difficult to interpret substantively, because of the existence of equivalent models. Thus, it is also of interest to know how well the intercepts $\boldsymbol{\alpha}$ and slopes $\boldsymbol{\beta}$ are estimated as these parameters are invariant to the choice of identification constraints and the ICCs and information function are a direct function of them. The $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters are obtained as a function of the parameters $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, ψ^2, and $\boldsymbol{\omega}^2$. Finally, it is also of interest to investigate latent trait recovery. To address these issues, we performed a number of simulation studies.

Item parameter recovery and goodness of fit tests

We considered 12 conditions by crossing 3 sample sizes (200, 500, and 1000 respondents), two model sizes (6 and 12 items), and 2 model conditions (paired comparison models with equal and unequal paired specific variances $\boldsymbol{\omega}^2$). 1000 replications were used in each condition. Estimation of the Thurstonian IRT model was performed via tetrachoric correlations using Mplus. ULS estimation was used to estimate the fundamental model parameters from the tetrachoric correlations. The intercepts and slopes were computed in Mplus from the model parameters and their standard errors obtained using the delta method.

For 6 items, the true parameters used were $\boldsymbol{\lambda}' = (1.5,\ 1,\ 0,\ 0,\ -1,\ -1.5)$, $\boldsymbol{\mu}_t' = \left(-0.2,\ 0.2,\ -.7,\ .7,\ 0.2,\ -0.2\right)$, $\psi^{2\prime} = \left(1,\cdots,1\right)$, $\boldsymbol{\omega}^{2\prime} = \left(0.3,\ldots,0.3\right)$. For 12 items, this setting was simply duplicated. Table 1 provides the minimum and maximum relative bias, expressed as a percentage, of the parameter estimates and standard errors. If we use 10% as cut off for good performance, the results shown in Table 1 reveal that a sample size of 1000 observations is needed for good recovery of the fundamental parameters of the model (i.e., $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, ψ^2, and $\boldsymbol{\omega}^2$) when 6 items are used. Item parameter recovery improves dramatically with

increasing model size. As few as 200 observations provide accurate item parameters when the paired specific variances are equal with 12 items. 500 observations are needed to accurately estimate the thresholds, factor loadings and uniquenesses when the paired specific variances are unequal. Much larger sample sizes are needed to estimate the unequal paired specific variances.

Most interestingly, the intercepts and slopes (i.e., the ICCs) are very accurately estimated in all conditions even when the fundamental parameters themselves are extraordinarily poorly estimated. This is shown in Table 2, which provides the minimum and maximum relative bias of the intercept and slope estimates, as well as of its standard errors. This is a very important and surprising finding, as latent trait estimation, and the goodness of fit of the model depend on how well the ICCs are estimated not on how well each fundamental parameter is estimated.

Insert Tables 1 to 3 about here

Turning to the results for goodness of fit tests, Table 3 provides the empirical rejection rates of the mean corrected goodness of fit test of the model to the tetrachoric correlations. As this table shows, the test maintains its nominal rates for all the small models considered, whereas it is slightly too conservative for 12 items (it rejects slightly less than it should), particularly when sample size is 200.

In the above simulations we investigated item parameter recovery for the Thurstonian IRT model (i.e., a first order model with correlated residuals and restrictions on the parameters). In terms of item parameters, this model and the Thurstonian factor model with unconstrained thresholds (i.e., a second order model) are equivalent. Nevertheless we also run some conditions also using the Thurstonian factor model to investigate whether the choice of parameterization affected in any way the results. It did not, results were absolutely

identical in all replications and conditions. However, the IRT model runs considerably faster than the Thurstonian factor model. However, the Thurstonian IRT model and the Thurstonian factor model are not equivalent when used to score the latent traits as in the former we use the simplifying assumption that ICCs are locally independent.

Latent trait recovery

To investigate how well MAP scores can recover the true latent trait scores we performed additional simulations. Fourteen conditions were considered. The conditions were obtained by crossing two model sizes (6 and 12 items), four values of the paired specific error variances (0, 0.1, 0.3, and 0.5), and two models (the Thurstonian factor model and the Thurstonian IRT model). The ICC for the Thurstonian factor model and details on how to estimate MAP scores under this model are given in Appendix B. The same values for the factor loadings, thresholds and uniquenesses used in the previous simulations were used here. Here, however, we varied the value of the common paired specific error variance to investigate if it affected in any way latent trait recovery. All simulations were performed using Mplus. In all cases, item parameters were treated as known and true latent trait scores were generated using the Thurstonian factor model. Hence, the use of the Thurstonian IRT model for scoring assuming local independence involves the use of a misspecified model. MAP estimates can not be computed for the Thurstonian factor model when $\boldsymbol{\omega}^2 = 0$ (i.e. for ranking data). Hence, only 14 conditions were investigated (rather than 16). For each of the conditions, 100 datasets of 1000 respondents were used.

Table 4 provides the average correlation between true and recovered scores for each of the conditions. One clear result from this table is that the value of the paired specific error variance has negligible impact on latent trait recovery. In particular, latent trait recovery is very similar for ranking data ($\boldsymbol{\omega}^2 = 0$) and paired comparisons data ($\boldsymbol{\omega}^2 > 0$). Another clear result apparent in this table is the negligible impact of ignoring local

dependencies in latent trait estimation for these models. As can be seen in this table, the correlation between MAP scores obtained with or without a local independence assumption are in all cases around 0.998. Using the simplifying assumption of local independence only negligibly affects MAP scores. The only factor that has a clear impact on latent trait recovery is test length: the correlation between true and estimated scores is around 0.935 with 12 items, but only around 0.872 with 6 items. This is because MAP scores are biased towards the mean, particularly in small models, which leads to a small variance of the estimated MAP scores.

Insert Table 4 about here

Numerical examples

We provide two empirical applications to illustrate the features of the model introduced here. The first one involves assessing vocational interests using a paired comparison task, whereas the second one involves assessing work motivation using a ranking task. We provide the modeling results, selected ICCs and information functions, and estimations of true score recovery for these applications.

Example 1. Modeling vocational interests using a paired comparisons task

Elosua (2007) collected data from 1,069 adolescents in the Spanish Basque Country using the 16PF Adolescent Personality Questionnaire (APQ; Schuerger, 2001). The Work Activity Preferences section of this questionnaire includes a paired comparisons task involving the 6 types of Holland's RIASEC model (see Holland, 1997): Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. For each of the 15 pairs, respondents were asked to choose their future preferred work activity. Typically, one would be interested in the actual utilities of vocational interests in this paired comparison task (first-order latent variables), but other higher-order vocational factors might also be of

interest. Factorial representations of the RIASEC model have been extensively researched and discussed in the literature. Rounds and Tracey (1993) examined 77 published RIASEC correlation matrices and concluded that, taken together, these studies suggested the presence of a general factor with equal loadings on all specific interests, which they interpreted as bias. However, this uniform biasing factors would not be observed here due to the comparative nature of the task (Cheung & Chan, 2002). The remaining variance, Rounds and Tracey (1993) suggested, is best explained by the original theory-based circumplex. In Hogan's interpretation, for instance, one of the two orthogonal axes on the circumplex was Conformity, with Conventional at the positive pole and Artistic at the negative pole, Enterprising and Realistic loading positively, and Social and Investigative negatively. For the purposes of illustration we will fit a unidimensional Thurstonian IRT model here, with the latent trait representing Conformity.

Thus, we fitted a one-dimensional model with unrestricted thresholds. The model yields a chi-square of 102.427 on 80 df, $p = 0.046$, RMSEA = 0.016. The model fits rather well. Next, we consider obtaining a more parsimonious model. One way to do this is to set all the variances of the paired comparison specific errors ω_l^2 equal. In so doing, we obtain a chi-square of 155.940 on 94 df, RMSEA = 0.025. Clearly, this model fits more poorly, suggesting that the number of intransitivities may not be approximately equal across pairs. Another way to obtain a more parsimonious model is to constrain the thresholds γ_l by estimating the mean utilities μ_i. In this case, we obtain a chi-square of 150.873 on 90 df, RMSEA = 0.025. Therefore, this model also fits more poorly that our initial model. The best fitting unidimensional model for these data is the unrestricted one dimensional model. We provide in Table 5 the parameter estimates and standard errors for this model.

Insert Tables 5 and 6 and Figure 1 about here

It can be seen that an arbitrary choice of identification constraints in this case yielded a set of parameters that match well with the substantive theory. In line with the definition of Conformity, the scale Conventional has the highest positive loading and Artistic has the lowest negative loading on the common factor. However, these estimates are not unique. The results presented in Appendix A imply that alternative sets of parameters can be obtained that yield the same fit to the data. For instance, using equation (55) we find out that if instead of fixing the last factor loading to 0 we were to fix it to 1, we would obtain the following factor loadings estimates: 0.974, 0.716, 0.102, 1.511, 0.363, 1 (i.e., this particular change of identification constraint simply amounts to adding 1 to the estimates shown in Table 5). The standard errors are unaffected by the choice of identification constraint. Goodness of fit tests, intercepts and slopes, information functions and latent trait estimates are also unaffected by the choice of identification constraints.

Estimated intercepts $\boldsymbol{\alpha}$ and slopes $\boldsymbol{\beta}$ computed using (36) are shown in Table 6. Notice that about half of the slopes in the table are negative, whereas the other half are positive. Also, we notice that the magnitudes of the estimated slopes are in general very low. The only large slope in this example (-1.223) is for pair {3,4}. Not surprisingly, this slope relates to the pair {Artistic, Conventional}, two interests serving as the main negative and positive indicators for the latent trait, Conformity. The rest of the paired comparisons do not provide much information about the latent trait.

Given the little information about the latent trait contained in the binary outcome variables in this example it is not surprising that the MAP test information function is rather low and the latent trait standard errors are high (see Figure 1). The standard error function is relatively uniform, which justifies computing a single reliability index to summarize the precision of measurement across the latent trait continuum. Using (44), the estimated average error of estimation of MAP scores is 0.38, which yields a theoretical

estimate of reliability of 1 – 0.38 = 0.62. The empirical estimated average error of estimation, computed using (45), is 0.36, quite close to the theoretical estimate. However, the MAP estimates in this application are quite shrunken towards the mean, the sample variance of the estimated MAP scores, computed using (45), is only 0.64, which leads to a very low empirical estimate of reliability, 0.43. Thus, in this application the empirical estimate of reliability underestimates quite markedly the theoretical reliability. In either case, we conclude that although the model appears to fit well, the precision of measurement obtained is unacceptable. However, this particular paired comparisons task was used as an illustration as it was not designed to measure a single underlying trait. Instead, population parameters of the utilities (vocational interests) would be of interest here.

Example 2: Modeling work motivation using a ranking task

This empirical example is based on ranking data collected as a part of research in the area of work motivation (Yang, Inceoglu & Silvester, 2010). Nine broad features of the work environment that are positively related to employee well-being, for example “personal development”, were developed from ideas found in the literature on person-environment fit and the vitamin model of Warr (2007).

1. Supportive Environment
2. Challenging Work
3. Career Progression
4. Ethics
5. Personal Impact
6. Personal Development
7. Social Interaction
8. Competition
9. Work Security

A hypothesized common factor underlying these generally desirable work features is the general work motivation, i.e. having strong drive for working and achieving. One-thousand-and-eighty volunteers were asked to rank these job features "according to how important it is for you to have these in your ideal job". Extended descriptions of the job features were presented to the participants, for example: “The opportunity to develop your knowledge and skills and to get feedback on what you do well and less well”.

After transforming the observed ranks into binary outcomes, we fitted a unidimensional Thurstonian IRT model. Using DWLS estimation Mplus yielded a mean corrected chi-square of 3121.126 on 614 df, RMSEA = 0.062. However, since the binary outcomes arise from rankings the degrees of freedom (and the RMSEA) need to be adjusted using (33). The correct number of degrees of freedom is 594 but the RMSEA is still 0.062. The model fits acceptably. Table 7 displays the estimated factor loadings and uniquenesses. As we can see in this table, the job characteristic that is more strongly related to general work motivation is having a challenging work environment, followed by career progression and supportive environment. Interestingly, the characteristic that is least strongly related to work motivation is having work security.

Figure 2 shows the MAP information function (and the SE function) for this example. Interestingly, individuals scoring low on work motivation are measured with higher precision than individuals high on work motivation. Also, we obtain smaller SEs in this application than in the vocational interest application (there are more binary outcomes in this application). The standard error function is not too uniform, but we compute the reliability estimate for this example. Using (44), the estimated average error of MAP scores is 0.26. Thus, the theoretical estimate of reliability is 0.74. The empirical estimated average error of estimation, computed using (45), is 0.27, quite close to the theoretical estimate, and the sample variance of the estimated MAP scores, computed using (45), is 1.09, which leads to an empirical estimate of reliability of 0.76. Thus, in this application both estimates of reliability suggest an adequate level of measurement across the latent trait continuum. Also, the empirical estimate is very close to the theoretical estimate.

Discussion

Item response modeling is generally applied to single-stimulus or Likert-type items. However, it can also be applied to items presented in a comparative manner, for instance

using paired comparisons or ranking. Thurstonian models for comparative data become IRT models when the latent utilities (discriminal processes) in these models depend on a set of latent traits (Maydeu-Olivares, 2001; Maydeu-Olivares & Böckenholt, 2005). In this article we have deepened our understanding of Thurstonian IRT models, with a particular emphasis on unidimensional models (models with a single latent trait underlying the items).

Unidimensional Thurstonian IRT models are simply normal ogive models with structured factor loadings $\breve{\lambda}_l = \lambda_i - \lambda_k$, structured uniquenesses $\breve{\psi}_l^2 = \psi_i^2 + \psi_k^2 + \omega_l^2$, and structured local dependencies (i.e., local independence does not hold). These features of Thurstonian IRT models have important implications for item parameter estimation, latent trait estimation, and test construction. We discuss each of these topics in turn.

Full information maximum likelihood (FIML aka marginal maximum likelihood) is ill-suited for item parameter estimation in these models. For full information estimation, Markov Chain Monte Carlo (MCMC: Tsai & Böckenholt, 2001) may be better suited than FIML, but MCMC estimation is computationally very intensive. On the other hand, limited information estimation via thresholds and tetrachoric correlations is computationally very efficient and can be implemented using existing software. Here we used Mplus to this aim. Thurstonian models for comparative data can be specified in two equivalent ways: as a second-order factor analysis model for binary data, or as a first-order model with structured correlated errors. To distinguish them, we refer to the first approach as Thurstonian factor model, and to the latter as Thurstonian IRT model. It is simpler to write scripts for the Thurstonian factor model than for the Thurstonian IRT model as in the latter case one needs to impose constraints on the model parameters of the type (26) and (27). Also, when fitting the Thurstonian IRT model, Mplus warns that the $\tilde{n}$ by $\tilde{n}$ covariance matrix of residuals, $\breve{\mathbf{\Psi}}^2$, is not of full rank. We have pointed out that this matrix is of rank $n - 1$. Mplus input files for the examples in this article are available from the authors upon request.

Mplus also yields MAP trait scores as a side product of the parameter estimation process. However, it does so using the simplifying assumption of local independence for latent trait estimation. This has no effect when the Thurstonian factor model is used, as in this case local independence holds. Hence, one can obtain 'correct' latent trait estimates using the Thurstonian factor model, but only for paired comparisons models. No latent trait estimates can be obtained for ranking data. On the other hand, when the Thurstonian IRT model is used one obtains latent trait estimates for both paired comparisons and ranking data, but in this case local independence does not hold. However, as our simulation studies show the use of this simplifying assumption has negligible effect on the quality of the latent trait estimates.

Our simulation studies also show that model size (i.e. the number of items being compared) has a major impact on the accuracy of the item parameter estimates. Thresholds, factor loadings and uniquenesses are well estimated in large models (i.e. 12 items) even in small samples (200 observations) but very poorly estimated in small models (6 items). Very large samples (larger than 1000 observations) are needed to accurately estimate paired specific error variances (in paired comparisons models). Perhaps the most interesting finding is that the item characteristic curves (i.e., intercept and slopes) are very accurately estimated in these models even when individual parameters are not. We found that in all cases considered a sample of size 200 sufficed to estimate very accurately the ICCs. This is important, as latent trait recovery, information functions, even the goodness of fit tests depend on how well the ICCs are estimated and not on how well individual parameters are estimated.

No simulation studies have been presented comparing the standard errors obtained using the Thurstonian IRT model vs. the Thurstonian factor model because in the latter the standard errors also depend not only on the value of the latent trait but also on the values of

the utility errors. This is discussed in Appendix B.

Concluding remarks

Test design when comparative tasks are used is a different endeavor than in the case of single-stimulus or rating tasks. In rating tasks, items are selected so that their factor loadings are as high as possible because test information is a function of the loadings' magnitudes. In contrast, in comparative tasks, test information is a function of differences of factor loadings when one latent trait is measured. Hence, maximum information is obtained when these differences are largest, that is, when factor loadings are of widely different magnitudes. If all items to be compared are highly related to the latent trait, as in rating applications, test information will be low and latent traits will be estimated so poorly as to make the application useless. The problem with low discrimination when items have factor loadings that are too similar to each other is easy to illustrate if one considers comparing two equally discriminating statements from the same trait. Utilities for the two statements are likely to be very similar for the respondent and preference for one of them, therefore, will be random. Conversely, if items with varying discriminations are compared (particularly when one item is positively keyed and the other is negatively keyed), making a choice is easy because the utilities for the items are likely to be very different. Thus, it is important in comparative data applications with one underlying latent trait to select items with widely different expected factor loadings. Also, it is not important if the signs of factor loadings estimates are of the 'wrong' sign according to theory, as the sign of the loading depends on the values used to identify the model. On the other hand, intercepts and slopes are invariant to the choice of identification constraints, and so are information functions, reliability estimates, and latent trait scores.

Sufficient consideration has also to be given to the pairwise intercepts. In comparative tasks, intercepts are a function of differences of the utilities' means of the items.

The intercepts will influence the test information function, and to obtain sufficient information along the whole latent trait continuum, it is recommended to combine items so that the differences in their utilities' means are widely varying.

The above considerations are important for designing ranking and paired comparison tasks involving a single trait. Most often, however, ranking and paired comparison tasks are used to assess multiple traits. In multidimensional applications the number of items is much larger and it becomes unfeasible to present all items in a single block as in the examples shown in this paper. Rather, an incomplete paired comparisons design or a ranking task where items are presented in multiple blocks of rankings, typically triplets or quads, is called for. Multidimensional tests involving multiple blocks of rankings are generally referred to in the literature as forced-choice tests, and they may involve as many as 30 latent traits. The extension of the present setup to applications presented in these forms is straightforward: the two items in a paired comparison belong to different traits, the item characteristic function becomes a two-dimensional normal ogive model, and the item information involves computing directional derivatives (Brown & Maydeu-Olivares, 2009). These models have similarities and differences to the one-dimensional models described here. For instance, in the multidimensional case the consideration of widely varying factor loadings does not apply to the same extent, whereas other considerations such as the number of traits assessed become more important for efficient trait estimation. A detailed account of multidimensional Thurstonian IRT models for forced-choice tests is given in Brown and Maydeu-Olivares (in press).

References

Bock, R.D. & Jones, L.V. (1968). *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day.

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

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory.* New York: Springer Verlag, 33-49.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*, 453-465.

Brown, A. & Maydeu-Olivares, A. (2009). Improving forced-choice tests with IRT. *Paper presented at the 16th International Meeting of the Psychometric Society, 20-24 July 2009, Cambridge.*

Brown, A. & Maydeu-Olivares, A. (in press). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement.*

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.

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

Elosua, P. (2007). Assessing vocational interests in the Basque Country using paired comparison design. *Journal of Vocational Behavior, 71,* 135-145.

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.

Forero, C.G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating

data: Limited vs. full information methods. *Psychological Methods*, *14*, 275-299.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Eglewood Cliffs, NJ: Prentice Hall.

Yang, M., Inceoglu, I. & Silvester, J. (2010). Exploring ways of measuring Person-Job fit to predict engagement. *Paper presented at the BPS Division of Occupational Psychology conference*, January 13-15, Brighton, UK.

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

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*, 209-228.

Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data. *Mathematical Social Sciences*, *43*, 467-483.

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

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

Maydeu-Olivares, A. & Hernández, A. (2007). Identification and small sample estimation of Thurstone's unrestricted model for paired comparisons data. *Multivariate Behavioral Research*, *42*, 323-347.

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

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78, 218-225.

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

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). *Robust inference using weighted least*

squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript. College of Education, UCLA. Los Angeles, CA.

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

Rounds, J., & Tracey, T.J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology, 78(6)*, 875-890.

Schuerger, J. M. (2001). *16PF-APQ Manual.* Champaign, IL: Institute for Personality and Ability Testing.

Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data. *Communication and Cognition*, *20*, 45-62.

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.

Tsai, R.C. & Böckenholt, U. (2001). Maximum likelihood estimation of factor and ideal point models for paired comparison data. *Journal of Mathematical Psychology*, *45*, 795-811.

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.

Warr, P. (2007). *Work, happiness, and unhappiness.* Mahwah, NJ: Lawrence Erlbaum.

Appendix A: Relationship between item parameters in equivalent one-dimensional Thurstonian IRT models

Consider a Thurstonian model with parameter matrices $\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$, and $\boldsymbol{\Omega}^2$. Any model with parameter matrices $\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t$, and $\tilde{\boldsymbol{\Omega}}^2$ satisfying

$$\mathbf{DA}\tilde{\boldsymbol{\mu}}_t = \mathbf{DA}\boldsymbol{\mu}_t\,, \tag{47}$$

$$\tilde{\boldsymbol{\Sigma}}_t = c\boldsymbol{\Sigma}_t + \mathbf{d1}' + \mathbf{1d}'\,, \tag{48}$$

and

$$\tilde{\boldsymbol{\Omega}}^2 = c\boldsymbol{\Omega}^2\,, \tag{49}$$

is equivalent to the estimated model (Tsai, 2003; Corollary 1). That is, it yields the same fit to the data. In Equations (48) and (49) c is a positive constant and $\mathbf{d}$ is an $n \times 1$ vector of constants. These constants are arbitrary as long as $\tilde{\boldsymbol{\Sigma}}_t$ and $\tilde{\boldsymbol{\Omega}}^2$ are positive definite.

Assume $\boldsymbol{\Psi}^2$ and $\boldsymbol{\Omega}^2$ are diagonal matrices. Given a set of population item parameters of a unidimensional Thurstonian IRT model $\boldsymbol{\mu}_t, \boldsymbol{\lambda}, \boldsymbol{\Psi}^2$, and $\boldsymbol{\Omega}^2$, we can use equations (47) to (49) to obtain another set of population parameters, say $\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\lambda}}$, and $\tilde{\boldsymbol{\Psi}}^2$, that will yield the same fit to the data. With $\boldsymbol{\mu}_t, \boldsymbol{\lambda}, \boldsymbol{\Psi}^2$, and $\boldsymbol{\Omega}^2$ the true and unknown population parameters, the results below can be used to determine the population parameters that will be estimated when the f^{th} element of $\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\lambda}}$, and $\tilde{\boldsymbol{\Psi}}^2$ is fixed for identification (we fix the n^{th} element throughout this paper). Or with $\boldsymbol{\mu}_t, \boldsymbol{\lambda}, \boldsymbol{\Psi}^2$, and $\boldsymbol{\Omega}^2$ the parameter estimates obtained with a given set of identification constraints, the results below can be used to determine the parameter estimates that will be obtained when a different set of identification constraints involving the f^{th} element of $\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\lambda}}$, and $\tilde{\boldsymbol{\Psi}}^2$ is used.

To establish relationships between $\tilde{\boldsymbol{\Psi}}^2$ and $\boldsymbol{\Psi}^2$, and between $\tilde{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}$, we use equation (48). By fixing λ_f to $\tilde{\lambda}_f$, and ψ_f^2 to $\tilde{\psi}_f^2$, we obtain an equivalent model if and only if $\tilde{\boldsymbol{\Sigma}}_t = c\boldsymbol{\Sigma}_t + \mathbf{d1}' + \mathbf{1d}'$, which for one-dimensional Thurstonian models implies that

$$\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}' + \tilde{\boldsymbol{\Psi}}^2 = c\left(\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}^2\right) + \mathbf{d}\mathbf{1}' + \mathbf{1}\mathbf{d}' = \left[c\boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{d}\mathbf{1}' + \mathbf{1}\mathbf{d}'\right] + c\boldsymbol{\Psi}^2 . \tag{50}$$

The rightmost part of (50) is the only way to present the utilities covariance structure as a sum of two matrices one of which is diagonal (the uniqueness component). Therefore we can write

$$\tilde{\boldsymbol{\Psi}}^2 = c\boldsymbol{\Psi}^2 , \tag{51}$$

$$\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}' = c\boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{d}\mathbf{1}' + \mathbf{1}\mathbf{d}' . \tag{52}$$

The diagonal matrix of uniquenesses for the model where ψ_f^2 is fixed to $\tilde{\psi}_f^2$ contains $n-1$ elements $\tilde{\psi}_i^2 = c\psi_i^2$. It means that the ratio between any diagonal element in this matrix and the corresponding diagonal element in the matrix containing "true" uniquenesses is equal to c. It then follows that that the equality $c = \dfrac{\tilde{\psi}_i^2}{\psi_i^2} = \dfrac{\tilde{\psi}_f^2}{\psi_f^2}$ holds for any i, and therefore any uniqueness parameter in the equivalent model can be expressed through its "true" value multiplied by the ratio of the fixed parameter to its "true" value:

$$\tilde{\psi}_i^2 = \psi_i^2 \frac{\tilde{\psi}_f^2}{\psi_f^2} . \tag{53}$$

Now, it follows from (52) that for any i the following equations also hold:

$$\begin{array}{l} \tilde{\lambda}_i^2 = c\lambda_i^2 + 2d_i \\ \tilde{\lambda}_f^2 = c\lambda_f^2 + 2d_f \\ \tilde{\lambda}_i\tilde{\lambda}_f = c\lambda_i\lambda_f + d_i + d_f \end{array} . \tag{54}$$

Adding the first and the second equations, and subtracting the third multiplied by 2, we derive the following equality: $\tilde{\lambda}_i^2 + \tilde{\lambda}_f^2 - 2\tilde{\lambda}_i\tilde{\lambda}_f = c(\lambda_i^2 + \lambda_f^2 - 2\lambda_i\lambda_f)$, or $\left(\tilde{\lambda}_i - \tilde{\lambda}_f\right)^2 = c\left(\lambda_i - \lambda_f\right)^2$. It then follows that

$$\tilde{\lambda}_i = \tilde{\lambda}_f + \sqrt{\frac{\tilde{\psi}_f^2}{\psi_f^2}}\left(\lambda_i - \lambda_f\right). \tag{55}$$

It can be similarly shown that the relationship between the utilities' means is

$$\tilde{\mu}_i = \tilde{\mu}_f + \sqrt{\frac{\tilde{\psi}_f^2}{\psi_f^2}}\left(\mu_i - \mu_f\right). \tag{56}$$

And finally, it follows straight from (49) that

$$\tilde{\omega}_i^2 = \omega_i^2 \frac{\tilde{\psi}_f^2}{\psi_f^2}. \tag{57}$$

In models with unrestricted thresholds, (47) is replaced by

$$\mathbf{D}\tilde{\boldsymbol{\gamma}} = \mathbf{D}\boldsymbol{\gamma}, \tag{58}$$

and the relationship between the true thresholds and the estimated thresholds is

$$\tilde{\gamma}_i = \gamma_i \sqrt{\frac{\tilde{\psi}_f^2}{\psi_f^2}}. \tag{59}$$

For example, consider a model for paired comparison data involving $n = 5$ items with true parameters

$$\begin{aligned}
\boldsymbol{\lambda}' &= \left(1.5,\ 0.6,\ 1,\ 0.8,\ 1.5\right),\\
\boldsymbol{\psi}^{2\prime} &= \left(0.5,\ 1.2,\ 0.8,\ 1,\ 0.7\right),\\
\boldsymbol{\mu}_t{}' &= \left(1.3,\ 0.4,\ \text{-}0.2,\ 0.4,\ 0.5\right),\\
\boldsymbol{\omega}^{2\prime} &= \left(0.2,\ 0.1,\ 0.3,\ 0.2,\ 0.1,\ 0.9,\ 0.3,\ 0.5,\ 0.3,\ 0.5\right).
\end{aligned} \tag{60}$$

To estimate the model with a threshold structure we arbitrarily fix $\tilde{\lambda}_5 = 0$, $\tilde{\psi}_5^2 = 1$, and $\tilde{\mu}_5 = 0$. Using (53), (55), (56) and (57) the population factor loadings that would be estimated are

$$\begin{aligned}
\tilde{\boldsymbol{\lambda}}' &= \left(0.,\ \text{-}1.076,\ \text{-}0.598,\ \text{-}0.837,\ 0^*\right),\\
\tilde{\boldsymbol{\psi}}^{2\prime} &= \left(0.714,\ 1.714,\ 1.143,\ 1.43,\ 1^*\right),\\
\tilde{\boldsymbol{\mu}}_t{}' &= \left(\text{-}0.956,\ 0.084,\ 0.837,\ 0.120,\ 0^*\right),\\
\tilde{\boldsymbol{\omega}}^{2\prime} &= \left(0.14,\ 0.07,\ 0.21,\ 0.14,\ 0.07,\ 0.63,\ 0.21,\ 0.35,\ 0.21,\ 0.35\right),
\end{aligned} \tag{61}$$

where we have marked using an asterisk the parameters fixed for identification. If a model with unrestricted thresholds is estimated, then the true thresholds are $\boldsymbol{\gamma}' = (0.9,\ 1.5,\ 0.9,\ 0.8,\ 0.6,\ 0,\ \text{-}0.1,\ \text{-}0.6,\ \text{-}0.7,\ \text{-}0.1)$ and the population thresholds that would

be estimated are $\tilde{\boldsymbol{\gamma}}' = \big(1.040,\ 1.793,\ 1.08,\ 0.956,\ 0.753,\ 0.036,\ \text{-}0.084,\ \text{-}0.717,\ \text{-}0.837,\ \text{-}0.120\big)$.

This example shows why in applications one can get estimated factor loadings with the wrong sign according to substantive theory. If a solution with all loadings being positive is desired, all that is needed is to re-estimate the model fixing at zero the loading with smallest negative estimate instead of the last one. Indeed, equations (53), (55), (56) and (57) show that if $\tilde{\lambda}_2 = 0$ is used to identify the model instead of $\tilde{\lambda}_5 = 0$, we would estimate

$$\tilde{\boldsymbol{\lambda}}' = \big(1.076,\ 0^{*},\ 0.478,\ 0.239,\ 1.076\big), \tag{62}$$

and the remaining parameters shown in (61).

In closing, for equivalent models, slopes and intercepts are invariant to the choice of identification constraints $\tilde{\mu}_f, \tilde{\lambda}_f,$ and $\tilde{\psi}_f^2$. This is because

$$\tilde{\beta}_l = \frac{\tilde{\lambda}_i - \tilde{\lambda}_k}{\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\big(\lambda_i - \lambda_k\big)\sqrt{\dfrac{\tilde{\psi}_f^2}{\psi_f^2}}}{\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\lambda_i - \lambda_k}{\sqrt{\dfrac{\psi_f^2}{\tilde{\psi}_f^2}}\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\lambda_i - \lambda_k}{\sqrt{\psi_i^2 + \psi_k^2}} = \beta_l\,, \tag{63}$$

$$\tilde{\alpha}_l = \frac{\tilde{\mu}_i - \tilde{\mu}_k}{\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\big(\mu_i - \mu_k\big)\sqrt{\dfrac{\tilde{\psi}_f^2}{\psi_f^2}}}{\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\mu_i - \mu_k}{\sqrt{\dfrac{\tilde{\psi}_f^2}{\psi_f^2}}\sqrt{\tilde{\psi}_i^2 + \tilde{\psi}_k^2}} = \frac{\mu_i - \mu_k}{\sqrt{\psi_i^2 + \psi_k^2}} = \alpha_l\,. \tag{64}$$

Appendix B: Information function for the Thurstonian factor model for paired comparisons data

Letting $\boldsymbol{\eta}^* = \left(\mathbf{t}, \boldsymbol{\eta}\right)'$, from (24), the Thurstonian factor model with unrestricted thresholds can be written as

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \begin{pmatrix}\mathbf{A} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{t} \\ \boldsymbol{\eta}\end{pmatrix} + \mathbf{e} = -\boldsymbol{\gamma} + \boldsymbol{\Lambda}^*\boldsymbol{\eta}^* + \mathbf{e}\,, \tag{65}$$

with $\boldsymbol{\mu}_{\boldsymbol{\eta}^*} = \mathbf{0}$ and

$$\operatorname{cov}\left(\boldsymbol{\eta}^*\right) \equiv \boldsymbol{\Phi}^* = \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2 & \boldsymbol{\Lambda}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Lambda}' & \boldsymbol{\Phi}\end{pmatrix} \tag{66}$$

and we obtain MAP scores by minimizing (38) where

$$\boldsymbol{\Phi}^{*-1} = \begin{pmatrix}\boldsymbol{\Psi}^{-2} & -\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2} & \boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}\end{pmatrix} \tag{67}$$

Thus, for paired comparisons models, the ICC is

$$P_l\left(\boldsymbol{\eta}^*\right) = \Phi\left(\frac{-\gamma_l + \boldsymbol{\lambda}_l^{*\prime}\boldsymbol{\eta}^*}{\sqrt{\omega_l^2}}\right) = \Phi\left(\frac{-\gamma_l + t_i - t_k}{\sqrt{\omega_l^2}}\right), \tag{68}$$

since recall that for ranking models $\omega_l^2 = 0$ and the following discussion is not applicable. We note that the ICC does not depend on the latent traits, $\boldsymbol{\eta}$. It only depends on the utilities $\mathbf{t}$.

From here on we concentrate, for ease of exposition, on models with a single latent trait. In this case,

$$\operatorname{cov}\left(\boldsymbol{\eta}^*\right) \equiv \boldsymbol{\Phi}^* = \begin{pmatrix}\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}^2 & \boldsymbol{\lambda} \\ \boldsymbol{\lambda}' & 1\end{pmatrix}, \tag{69}$$

$$\boldsymbol{\Phi}^{*-1} = \begin{pmatrix} \frac{1}{\psi_1^2} & 0 & 0 & -\frac{\lambda_1}{\psi_1^2} \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & \frac{1}{\psi_n^2} & -\frac{\lambda_n}{\psi_n^2} \\ -\frac{\lambda_1}{\psi_1^2} & \cdots & -\frac{\lambda_n}{\psi_n^2} & 1+\boldsymbol{\lambda}'\boldsymbol{\Psi}^2\boldsymbol{\lambda} \end{pmatrix}. \tag{70}$$

Now, akin to (42) the information function about the latent trait η is

$$\begin{aligned} \mathcal{I}_P^{\eta}\left(\boldsymbol{\eta}^*\right) &= \mathcal{I}^{\eta}\left(\boldsymbol{\eta}^*\right) - \frac{\partial^2 \ln\left(\phi\left(\boldsymbol{\eta}^*\right)\right)}{\partial \eta^2} = \mathcal{I}^{\eta}\left(\boldsymbol{\eta}^*\right) + \left[\boldsymbol{\Phi}^{*-1}\right]_{\eta}, \\ &= \sum_l \mathcal{I}_l^{\eta}\left(\boldsymbol{\eta}^*\right) + 1 + \boldsymbol{\lambda}'\boldsymbol{\Psi}^2\boldsymbol{\lambda} \end{aligned} \tag{71}$$

where $\left[\boldsymbol{\Phi}^{*-1}\right]_{\eta}$ denotes the diagonal element of $\boldsymbol{\Phi}^{*-1}$ corresponding to the latent trait. Also, when conditioning on the utilities and the latent trait local independence holds, so the information function is additive. The ML item information about η is

$$\mathcal{I}_l^{\eta}\left(\boldsymbol{\eta}^*\right) = \frac{\left[\nabla_{\eta} P_l\left(\boldsymbol{\eta}^*\right)\right]^2}{P_l\left(\boldsymbol{\eta}^*\right)\left[1 - P_l\left(\boldsymbol{\eta}^*\right)\right]}, \tag{72}$$

where

$$\begin{aligned} \nabla_{\eta} P_l\left(\boldsymbol{\eta}^*\right) &= \sum_{i=1}^{n+1}\left(\frac{\partial P_l\left(\boldsymbol{\eta}^*\right)}{\partial \eta_i^*}\operatorname{corr}\left(\eta_i^*, \eta\right)\right) = \\ &= \frac{\partial P_l\left(\boldsymbol{\eta}^*\right)}{\partial t_i}\operatorname{corr}\left(t_i, \eta\right) + \frac{\partial P_l\left(\boldsymbol{\eta}^*\right)}{\partial t_k}\operatorname{corr}\left(t_k, \eta\right) \end{aligned}, \tag{73}$$

is the derivative in the direction of the latent trait (see Reckase, 2009) and from (69)

$$\operatorname{corr}\left(t_i, \eta\right) = \frac{\lambda_i}{\sqrt{\lambda_i^2 + \psi_i^2}}. \tag{74}$$

Finally, with $z_l = \dfrac{-\gamma_l + t_i - t_k}{\sqrt{\omega_l^2}}$, $\dfrac{\partial P_l\left(\boldsymbol{\eta}^*\right)}{\partial t_i} = \dfrac{\phi\left(z_l\right)}{\sqrt{\omega_l^2}}$ and $\dfrac{\partial P_l\left(\boldsymbol{\eta}^*\right)}{\partial t_k} = -\dfrac{\phi\left(z_l\right)}{\sqrt{\omega_l^2}}$.

Thus, the item information function for the Thurstonian factor model is

$$\mathcal{I}_l^{\eta}\left(\boldsymbol{\eta}^*\right)=\frac{\left(\dfrac{\lambda_i}{\sqrt{\lambda_i^2+\psi_i^2}}-\dfrac{\lambda_k}{\sqrt{\lambda_k^2+\psi_k^2}}\right)^2}{\omega_l^2}\frac{\left[\phi\left(z_l\right)\right]^2}{\Phi\left(z_l\right)\left[1-\Phi\left(z_l\right)\right]}, \tag{75}$$

This is to be compared with the item information function for the Thurstonian IRT model (40), which using the threshold/factor loading parameterization and $x_l=\dfrac{-\gamma_l+\left(\lambda_i-\lambda_k\right)\eta}{\sqrt{\psi_i^2+\psi_k^2+\omega_l^2}}$ is

$$\mathcal{I}_l\left(\eta\right)=\frac{\left(\lambda_i-\lambda_k\right)^2}{\left(\psi_i^2+\psi_k^2+\omega_l^2\right)}\frac{\left[\phi\left(x_l\right)\right]^2}{\Phi\left(x_l\right)\left[1-\Phi\left(x_l\right)\right]}, \tag{76}$$

where recall that $\breve{\lambda}_l=\lambda_i-\lambda_k$.

We did not perform a simulation study comparing the SEs for the MAP scores of the latent trait obtained using the Thurstonian factor model and the Thurstonian IRT model, because in the former SEs for the latent trait estimates depend on the utilities, that is, on the value of the latent trait, but also on the values of the utility errors ε, see (8). In other words, in a Thurstonian factor model with a single trait, the SE of a MAP latent trait estimate is not unique since it depends also on the values of the utility errors. In contrast, in the Thurstonian IRT model the SE for a MAP latent trait estimate is unique.

However, we can compare the SE function (76) for the Thurstonian IRT model to the *average* SE function for the Thurstonian factor model. This is Equation (75) with the utility errors ε evaluated at their mean, 0. As an illustration, we provide the Figure 3 both functions for the 12 item condition described above with ω^2= 0.3. As can be seen in this Figure, the SE obtained for the Thurstonian IRT model (under the simplifying assumption of local independence) is very close to the average of the 'correct' SEs (those obtained for the Thurstonian factor model) in the latent trait range (-3, 3). Outside this range, the Thurstonian IRT model SE is larger. Also, note the 'bump' in the average SE function for the Thurstonian factor model, which we believe is the result of being a second order model.

Table 1

Minimum and maximum relative bias (in percentage) of estimates and standard errors for fundamental parameters

			$\widehat{\gamma}$		$\widehat{\lambda}$		$\widehat{\psi}^2$		$\widehat{\omega}^2$	
n	N	ω^2	bias est.	bias SE	bias est.	bias SE	bias est.	bias SE	bias est.	bias SE
6	1000	common	1; 2	-2; 5	-1; 4	-3; -1	3; 5	-3; 1	4	-2
6	500	common	3; 6	-15; 3	-4; 12	-17; -14	12; 18	-34; -30	13	-32
6	200	common	11; 16	47; 86	-12; 41	79; 85	71; 107	178; 192	78	183
6	1000	unequal	1; 3	-2; 5	-1; 5	-4; -2	4; 5	-3; 1	4; 14	-4; 4
6	500	unequal	3; 6	-9; 4	-4; 12	-9; -7	12; 9	-22; -17	12; 23	-14; 3
6	200	unequal	11; 16	15; 35	-11; 40	30; 32	53; 76	65; 72	38; 137	29; 72
12	1000	common	-1; 2	-4; 6	0; 1	-5; 0	1; 2	-1; 7	1	2
12	500	common	-1; 2	-4; 6	-1; 2	-4; 1	2; 3	-4; 2	2	-3
12	200	common	0; 4	-5; 5	-2; 4	-4; 4	3; 5	-8; 0	3	4
12	1000	unequal	-1; 2	-4; 6	-1; 1	-5; 1	1; 2	-2; 6	0; 13	-5; 4
12	500	unequal	-1; 3	-4; 7	-1; 3	-5; 1	1; 3	-4; 2	1; 22	-11; 5
12	200	unequal	0; 5	-7; 6	-3; 7	-5; 2	3; 34	-8; 0	3; 54	-13; 5

Notes: 1000 replications per condition. For 6 items, $\boldsymbol{\lambda}' = \left(1.5, 1, 0, 0, -1, -1.5\right)$, $\boldsymbol{\mu}_t' = \left(-0.2, 0.2, -.7, .7, 0.2, -0.2\right)$, $\psi^{2\prime} = \left(1,\cdots,1\right)$, $\boldsymbol{\omega}^{2\prime} = \left(0.3,\ldots,0.3\right)$. For 12 items, this setting was duplicated. When $\boldsymbol{\omega}^2$ elements are constrained to a common value, the minimum and maximum coincide.

Table 2

Minimum and maximum relative bias (in percentage) of estimates and standard errors for derived parameters

			$\widehat{\boldsymbol{\alpha}}$		$\hat{\boldsymbol{\beta}}$	
n	N	$\boldsymbol{\omega}^2$	bias est.	bias SE	bias est.	bias SE
6	1000	common	0; 1	-2; 5	0; 1	-5; 2
6	500	common	0; 1	-2; 2	1; 2	-4; 3
6	200	common	1; 3	-6; 4	2; 5	-4; 1
6	1000	unequal	0; 1	-2; 5	0; 1	-4; 3
6	500	unequal	0; 2	-2; 3	1; 3	-6; 4
6	200	unequal	1; 5	-6; 4	2; 7	-8; 1
12	1000	common	-2; 1	-5; 6	-1; 1	-7; 3
12	500	common	-2; 2	-4; 7	-1; 1	-5; 3
12	200	common	-2; 3	-4; 5	-1; 3	-6; 4
12	1000	unequal	-2; 1	-5; 6	-1; 1	-7; 2
12	500	unequal	-2; 2	-4; 8	-1; 1	-7; 3
12	200	unequal	-2; 4	-5; 5	0; 3	-11; 4

Notes: 1000 replications per condition. For 6 items, $\boldsymbol{\lambda}' = \left(1.5,\ 1,\ 0,\ 0,\ -1,\ -1.5\right)$, $\boldsymbol{\mu}_t{}' = \left(-0.2,\ 0.2,\ -.7,\ .7,\ 0.2,\ -0.2\right)$, $\boldsymbol{\psi}^{2\prime} = \left(1,\cdots,1\right)$, $\boldsymbol{\omega}^{2\prime} = \left(0.3,\ldots,0.3\right)$. For 12 items, this setting was duplicated.

Table 3

Empirical rejection rates of the chi-square test of exact fit across 1000 replications

			rejection rates			
n	N	ω^2	1%	5%	10%	20%
6	1000	common	1.3	4.6	10.3	19.1
6	500	common	0.8	5.2	9.9	16.8
6	200	common	0.7	3.7	8.8	19.8
6	1000	unequal	1.4	5.2	10.3	18.9
6	500	unequal	0.8	5.6	9.7	18.4
6	200	unequal	1.3	4.1	8.8	18.7
12	1000	common	0.1	3.1	7.6	16.9
12	500	common	0.1	1.4	4.4	14.8
12	200	common	.0	1.1	3.5	12.2
12	1000	unequal	0.5	2.8	7.1	15.8
12	500	unequal	0.2	1.3	5.3	15.9
12	200	unequal	0	0.8	3.4	11.8

Table 4

Average correlations between true latent trait scores and MAP scores across 100 sets of 1000 respondents

		correlations between		
items	ω^2	true scores and MAP scores	true scores and MAP scores assuming local independence	MAP scores and MAP scores assuming local independence
6	0	–	.873	–
6	.1	.872	.871	.997
6	.3	.871	.870	.998
6	.5	.871	.869	.998
12	0	–	.936	–
12	.1	.937	.935	.997
12	.3	.936	.932	.997
12	.5	.934	.928	.997

Notes: Item parameters are assumed to be known. For 6 items, $\boldsymbol{\lambda}' = \left(1.5,\ 1,\ 0,\ 0,\ -1,\ -1.5\right)$, $\boldsymbol{\mu}_t{}' = \left(-0.2,\ 0.2,\ -.7,\ .7,\ 0.2,\ -0.2\right)$, $\boldsymbol{\psi}^{2\prime} = \left(1, \cdots, 1\right)$. For 12 items, this setting was duplicated. $\omega^2 = 0$ implies ranking data, in this case MAP scores can not be computed easily without assuming local independence

Table 5

One-dimensional Thurstonian IRT model for paired comparisons data. Vocational interests example. Parameter estimates and standard errors.

$l = i,j$	γ_l	ω_l^2
1,2	0.742 (0.093)	1.003 (0.302)
1,3	0.421 (0.081)	1.146 (0.296)
1,4	0.055 (0.063)	0.464 (0.189)
1,5	0.807 (0.103)	1.213 (0.358)
1,6	-0.035 (0.067)	0.346 (0.193)
2,3	-0.35 (0.068)	0.778 (0.233)
2,4	-0.644 (0.084)	0.831 (0.256)
2,5	0.172 (0.07)	0.572 (0.252)
2,6	-0.858 (0.084)	0.505 (0.215)
3,4	-0.517 (0.079)	0.639 (0.219)
3,5	0.329 (0.067)	0.521 (0.222)
3,6	-0.48 (0.072)	0.639 (0.209)
4,5	0.768 (0.106)	1.799 (0.444)
4,6	0.079 (0.07)	1.815 (0.483)
5,6	-1.45 (0.14)	2.523 (0.560)

i	λ_i	ψ_i^2
1	-0.026 (0.089)	1.692 (0.226)
2	-0.284 (0.083)	0.892 (0.132)
3	-0.898 (0.143)	0.464 (0.154)
4	0.511 (0.120)	0.224 (0.178)
5	-0.636 (0.106)	1.534 (0.253)
6	0 (*fixed*)	1 (*fixed*)

Notes: Standard errors in parentheses. The items are: 1 = Realistic, 2 = Investigative, 3 = Artistic, 4 = Conventional, 5 = Social, 6 = Enterprising.

Table 6

Intercepts and slopes for the Vocational interests example. Parameter estimates and standard errors.

$l = i,k$	α_l	β_l
1,2	-0.392 (0.057)	0.136 (0.047)
1,3	-0.232 (0.048)	0.480 (0.079)
1,4	-0.036 (0.041)	-0.347 (0.092)
1,5	-0.383 (0.056)	0.290 (0.051)
1,6	0.020 (0.038)	-0.015 (0.051)
2,3	0.240 (0.052)	0.421 (0.090)
2,4	0.461 (0.078)	-0.569 (0.112)
2,5	-0.099 (0.041)	0.204 (0.054)
2,6	0.554 (0.063)	-0.184 (0.053)
3,4	0.448 (0.087)	-1.223 (0.152)
3,5	-0.207 (0.046)	-0.165 (0.079)
3,6	0.331 (0.055)	-0.620 (0.107)
4,5	-0.407 (0.068)	0.608 (0.084)
4,6	-0.045 (0.040)	0.293 (0.081)
5,6	0.645 (0.077)	-0.284 (0.047)

Notes: The items are: 1 = Realistic, 2 = Investigative, 3 = Artistic, 4 = Conventional, 5 = Social, 6 = Enterprising.

Table 7

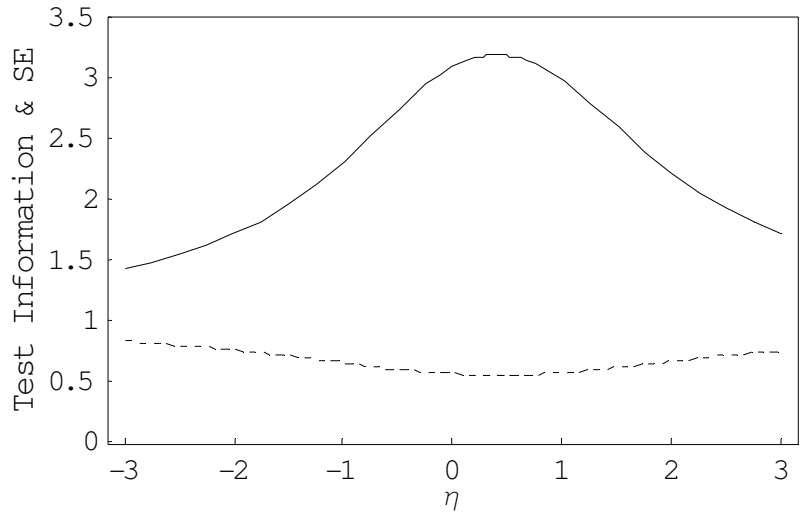
One-dimensional Thurstonian IRT model for ranking data. Work motivation example. Factor loading and uniqueness estimates and their standard errors.

i	λ_i	ψ_i^2
1	1.028 (.158)	1.330 (.222)
2	1.313 (.157)	.851 (.167)
3	1.104 (.154)	1.123 (.193)
4	.931 (.145)	.998 (.164)
5	.882 (.136)	.878 (.144)
6	.908 (.143)	.566 (.092)
7	.539 (.122)	.613 (.108)
8	.330 (.120)	1.346 (.249)
9	0 (*fixed*)	1 (*fixed*)

Notes: Standard errors in parentheses. The thresholds are not shown. The paired specific errors are fixed to zero. The items are: 1 = Supportive Environment, 2 = Challenging Work 3 = Career Progression, 4 = Ethics, 5 = Personal Impact, 6 = Personal Development, 7 = Social Interaction, 8 = Competition, 9 = Work Security.

Figure 1

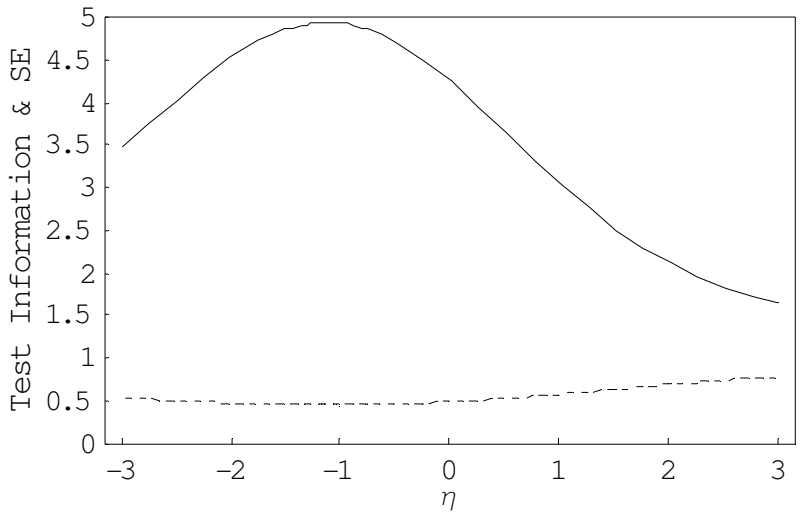
MAP test information function and SE function for the Vocational interests example



Notes: The dotted line is the SE function.

Figure 2

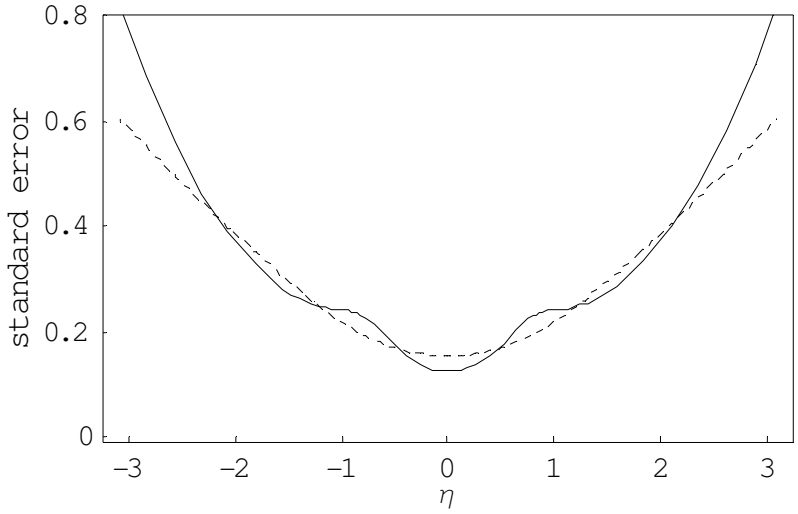
MAP test information function and SE function for the work motivation example



Notes: The dotted line is the SE function.

Figure 3

MAP SE function for the Thurstonian IRT model and average MAP SE function for the Thurstonian factor model: 12 item condition with $\omega^2 = 0.3$



Notes: The dotted line is the SE function for the Thurstonian IRT model; the solid line is the average SE function for the Thurstonian factor model.