Citation:
Brown, A. & Maydeu-Olivares, A. (in press). Ordinal Factor Analysis of Graded-Preference Questionnaire Data. *Structural Equation Modeling: A Multidisciplinary Journal*. DOI: 10.1080/10705511.2017.1392247

Ordinal Factor Analysis of Graded-Preference Questionnaire Data

Anna Brown

University of Kent

Alberto Maydeu-Olivares

University of South Carolina, University of Barcelona

Author Note

Anna Brown, PhD, Senior Lecturer in Psychological Methods and Statistics, School of Psychology, University of Kent.

Alberto Maydeu-Olivares, PhD, Professor, Psychology Department, University of South Carolina.

Correspondence should be addressed to Anna Brown, School of Psychology, University of Kent, Canterbury, Kent CT2 7NP, United Kingdom. E-mail: A.A.Brown@kent.ac.uk

Abstract

We introduce a new comparative response format, suitable for assessing personality and similar constructs. In this "graded-block" format, items measuring different constructs are first organized in blocks of 2 or more; then, pairs are formed from items within blocks. The pairs are presented one at a time, to enable respondents expressing the extent of preference for one item or the other using several graded categories. We model such data using confirmatory factor analysis (CFA) for ordinal outcomes. We derive Fisher information matrices for the graded pairs, and supply R code to enable computation of standard errors of trait scores. An empirical example illustrates the approach in low-stakes personality assessments and shows that similar results are obtained when using graded blocks of size 3 and a standard Likert format. However, graded-block designs may be superior when insufficient differentiation between items is expected (due to acquiescence, halo or social desirability).

Keywords: Thurstonian IRT model, ipsative data, graded preferences, graded response model

Ordinal Factor Analysis of Graded-Preference Questionnaire Data

The most common method to measure personality traits, personal values and similar constructs is using Likert-type items (aka *ratings*). However, when this method is used, respondents may endorse all items regardless of their valence (so-called "acquiescence") or trait allocation (cognitive bias of exaggerated coherence, or "halo" effect). In applications where these effects are common, the validity of inferences is threatened. In such applications, the use of *comparative* judgments (i.e., asking respondents about their preferences for one or another item) is an attractive alternative because comparisons between items facilitate better differentiation and calibration thus reducing halo effects (Kahneman, 2011). Also, when forced to compare items, one cannot agree with all of them indiscriminately thus alleviating acquiescent responding (Cheung & Chan, 2002).

Preferences can be expressed as choices among two items, and as rankings or partial rankings among three or more items. Data collected by this method represent binary choices involved for each pair of items within a set. Simple choice, however, is not the only way of expressing preferences. We may want to obtain quantitative information about the relative merits of items within the set. For example, we may ask respondents to distribute a fixed number of points (say, 100) between the items, resulting in so-called *compositional* data (Brown, 2016b). Or, we may ask respondents to indicate how much they prefer item A to B using a number of ordered categories, such as "much more – a little more – a little less – much less". Each subsequent category represents diminishing preference for item A and increasing preference for item B. Such *graded-preference* format is the focus of the present paper.

Why would we consider collecting graded preferences, if binary preferences have already proven themselves an attractive alternative to ratings, particularly for their resistance to response biases? We believe this extension is desirable for at least two reasons. First, test

takers often criticize forced-choice formats for the perceived "lack of choice" when presented with items that either all apply to them or none apply; as one test taker put it "…*responding correctly was impossible because it forced a choice between equally ranked options*" (Bartram & Brown, 2003). Allowing the test takers to indicate the extent of their preference could increase their engagement and the face validity of the questionnaire. Second, scores derived from forced-choice responses (representing binary choices among pairs of items) generally have lower reliability than scores obtained from Likert ratings of the same items. It is easy to see when considering a simple choice between two items A and B, which can result in only one of two possible outcomes: either A is preferred to B or otherwise. Clearly, such a binary variable contains less information than Likert ratings of the same two items using, say, 5 ordered categories. More information can be obtained per item in forced-choice tasks when items are combined in larger blocks (Brown & Maydeu-Olivares, 2011); however, blocks of 4 items still yield lower reliability than the 5-point Likert scales (Brown & Maydeu-Olivares, 2013). As a result, more items are needed in general in forced-choice questionnaires to reach the same precision of measurement as their Likert-scales counterparts. The additional information obtained from every comparison by asking participants to quantify the preferences may help solve this problem.

How would we score a questionnaire composed of graded-preference items? The simple summative schemes, where preference for one item adds points to that item while decreasing by the same amount the points awarded to the other item will result in *ipsative* scores. Ipsative, or relative-to-self scores, are problematic for interpersonal comparisons and preclude application of standard psychometric analyses (Brown & Maydeu-Olivares, 2013; Clemans, 1966; Closs, 1996; Dunlap & Cornwell, 1994). However, recent advances in modeling forced-choice data (Brown & Maydeu-Olivares, 2011b, 2012; Maydeu-Olivares & Brown, 2010), which have enabled proper scoring of personal attributes without artefacts of

ipsative data, have not yet been extended to graded comparisons. The present paper aims to fill this gap. The objectives of this paper are as follows. The first objective is to introduce a response format for gathering measurements on latent attributes using graded comparisons. We refer to this new format as *graded blocks*. In graded-block designs, individuals are presented with pairs of items and are asked to indicate the extent to which they prefer one item to the other (or the extent to which one item describes their personality or attitudes better than the other item) using a graded scale. The second objective is to propose a model suitable for such data. Such a model needs to take into account: a) the ordinal and comparative nature of the data, b) dependencies when the same item is administered in more than one pair, c) potential intransitivity of responses to pairs involving the same items (an individual may prefer A to B, and B to C, but not prefer A to C). The third objective is to provide the item and test information functions suitable for the proposed model. Armed with such a model, researchers may analyze existing graded-preference data, design optimal graded-block questionnaires, or infer the expected properties of their questionnaires before data are gathered.

The remainder of this article is organized as follows. First, we describe the graded-block design. In a nutshell, items are first organized into blocks of n items (the block size n can be 2, 3, 4, etc.). All possible pairs are drawn from items within each block. Then, the resulting pairs of items are administered using a graded scale. Next, we describe a model suitable for these data. The model is based on Thurstone's (1927) law of comparative judgment, where utilities of items under comparison are linked to graded preference decisions via a threshold process to accommodate ordinal data. We show that the proposed model is an ordinal factor analysis model with specific constraints, and it can be estimated using standard software such as Mplus (Muthén & Muthén, 2016). Since ordinal factor analysis models belong to the general class of IRT models, in technical appendices, we provide the item and

test information functions. Our derivation takes into account the inherent multidimensionality of responses when items measuring different attributes are compared, and the fact that it is impossible to estimate the latent traits separately in such designs. We provide R functions to compute the item and test information, allowing computation of standard errors for estimated scores, and reliability estimates. To illustrate the graded-preference model, we provide an empirical example, in which the Five Factor markers (Goldberg, 1992) are measured using two alternative response formats: standard Likert ratings, and graded blocks. We conclude with a general discussion and a set of recommendations for applied researchers.

The Graded-Block Design

In forced-choice questionnaires, items are uniquely assigned to blocks of size n, and respondents are asked to provide a ranking or a partial ranking of the items within the blocks. In a graded-block design, items to be compared with each other are still drawn from within blocks, but they are presented as pairs to enable graded comparisons. For each pair, respondents are asked to express the extent of their preference for one item or the other using several graded categories. For instance, they may prefer item A "much more" or "slightly more" than item B, be ambivalent about items A and B, or they may prefer item B "slightly more" or "much more" than item A.

	Much more	Slightly more	About the same	Slightly more	Much more	
Item A		X				Item B

If the block size is $n = 2$, the two items from each block are simply presented as one pair. If the block size is $n \geq 3$, all possible pairs of items are drawn from within each block, and the resulting $\tilde{n} = n(n-1)/2$ pairs per block are presented to respondents one at a time. In this case, the same items will appear in more than one pair, but pairs drawn from the same block need not be administered sequentially. Instead, researchers may want to randomize the

presentation of such pairs across the questionnaire to minimize the carry-over effect. Importantly, the model for such designs needs to take into account these patterns of within-block dependencies arising from the repeated item use.

The reason we might want to draw paired comparisons from blocks of 3 or more items is to increase the amount of information obtained per one item. Indeed, when pairs are drawn from blocks of size $n = 2$, the questionnaire has half the number of tasks than a standard Likert-type questionnaire in which items are presented one at a time. When pairs are drawn from blocks of size $n = 3$, there are $\tilde{n} = 3$ pairs arising from each block, and the questionnaire has the same number of tasks as a standard rating task. When items are drawn from blocks of size $n = 4$, there are $\tilde{n} = 6$ pairs arising from each block and the questionnaire contains more tasks than a standard rating task, and therefore may gather more information per item than a questionnaire created from smaller blocks.

To code the graded preferences appropriately, we will always consider the degree of preference for the **first** item in the pair $\{i, k\}$, item i, arbitrarily using descending integers[1], for example:

$$y_{\{i,k\}} = \begin{cases} 5, & \text{if } i \text{ is preferred "much more" than } k \\ 4, & \text{if } i \text{ is preferred "slightly more" than } k \\ 3, & \text{if } i \text{ and } k \text{ are "about the same"} \\ 2, & \text{if } k \text{ is preferred "slightly more" than } i \\ 1, & \text{if } k \text{ is preferred "much more" than } i \end{cases} . \quad (1)$$

Responses coded in this way are the observed outcomes in graded-preference analysis. It is easy to see that the observed outcomes are *ordinal* variables.

[1] The coding 5, 4, ..., 1 is consistent with previous work on factor analysis of binary outcomes (Maydeu-Olivares & Böckenholt, 2005), in which preference for the first stimuli in a pair is coded 1 and for the second is coded 0. Should the ordinal outcomes be coded as ascending integers 1, 2.., 5, all factor loadings will have signs opposite to the ones in the present paper.

Modeling Graded-Preference Questionnaire Data

To model graded preferences when items are presented in pairs, we use the law of comparative judgment (Thurstone, 1927), which attributes preference decisions to the relative *utilities* (or psychological values) of items under comparison. Thus, person *j* prefers item *i* to item *k* if his/her utility for item *i* (t_{ji}) is greater than the utility for *k* (t_{jk}). Therefore, the unobserved *difference* of utilities

$$y^*_{j\{i,k\}} = t_{ji} - t_{jk} \tag{2}$$

is the fundamental quantity in the analysis, which determines the observed preference decision $y_{j\{i,k\}}$ via a threshold process (Böckenholt & Dillon, 1997; Maydeu-Olivares, 2002):

$$y_{j\{i,k\}} = \begin{cases} C, & \text{if } \; y^*_{j\{i,k\}} \geq \tau_{\{i,k\}C-1} \\ C-1, & \text{if } \; \tau_{\{i,k\}C-2} \leq y^*_{j\{i,k\}} < \tau_{\{i,k\}C-1} \\ \quad \ldots & \\ 2, & \text{if } \; \tau_{\{i,k\}1} \leq y^*_{j\{i,k\}} < \tau_{\{i,k\}2} \\ 1, & \text{if } \; y^*_{j\{i,k\}} < \tau_{\{i,k\}1} \end{cases} . \tag{3}$$

According to this threshold process, person *j* selects one of *C* graded options depending on the size of the latent difference $y^*_{j\{i,k\}}$, and a set of $C - 1$ thresholds.

However, when graded paired comparisons are drawn from blocks of three or more items ($n \geq 3$), respondents need not be consistent in their pairwise preferences, possibly yielding intransitive patterns of preference. That is, they may prefer item *i* to item *k*, item *k* to item *l* but not prefer item *i* to *l*. Intransitive pairwise preferences can be accommodated by adding an error term to the difference of utility judgements (Maydeu-Olivares & Böckenholt, 2005; Takane, 1989):

$$y^*_{j\{i,k\}} = t_{ji} - t_{jk} + e_{j\{i,k\}} . \tag{4}$$

The next section describes the distributional assumptions for the unobserved utilities and intransitivity error terms that are necessary to model graded preferences.

Ordinal Factor Model for Graded-Block Preferences

Consider a questionnaire containing b blocks of $n \geq 2$ items where items are to be presented in pairs using a graded scale. Since for each block $\tilde{n} = n(n-1)/2$ item pairs can be obtained, there are $b\tilde{n}$ ordinal responses for each respondent.

In matrix form, the model can be written as follows. Let $\mathbf{y}$ be a $b\tilde{n}$ vector of observed ordinal variables, which are related to the corresponding latent utility differences $\mathbf{y}^*$ via the threshold process (3). The $b\tilde{n}$ vector of latent utility differences $\mathbf{y}^*$ is given by (4)

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

where $\mathbf{t}$ is a bn vector of item utilities, $\mathbf{A}$ is a $b\tilde{n} \times bn$ block-diagonal design matrix of contrasts, and $\mathbf{e}$ is a $b\tilde{n}$ vector of pairwise intransitivity errors needed when block size $n \geq 3$ (these are zero when block size $n = 2$ since there cannot be any intransitivity in a single pair). The errors $\mathbf{e}$ are assumed to have mean zero and uncorrelated with the utilities. They are also assumed uncorrelated with each other so that their covariance matrix $\mathbf{\Omega}^2$ is diagonal. The block-diagonal matrix $\mathbf{A}$ contains contrasts of utilities arising from each block. For $n = 2$, the diagonal entries contrast the first item in a pair with the second $\mathbf{A}_2 = \begin{pmatrix} 1 & -1 \end{pmatrix}$; and for $n = 3$ and $n = 4$, respectively, the contrasts are pairwise:

$$\mathbf{A}_3 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \quad \mathbf{A}_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{6}$$

Because questionnaires are designed to measure some personal attributes (latent traits), we assume that the item utilities depend linearly on a set of d common factors $\boldsymbol{\eta}$ representing the attributes, and the unique factors $\boldsymbol{\varepsilon}$

$$\mathbf{t} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,, \tag{7}$$

where $\boldsymbol{\Lambda}$ is a $bn \times d$ matrix of the factor loadings. The factor analysis model assumes that the common and unique factors have mean zero and they are uncorrelated. The unique factors are assumed uncorrelated so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal. The common factors may be correlated among themselves, with covariance matrix $\boldsymbol{\Phi}$.

Putting together the first-order structure (5) and the second-order structure (7),

$$\mathbf{y}^* = \mathbf{A}(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}\,. \tag{8}$$

Assuming that the common and unique factors, as well as the pairwise intransitivity errors are normally distributed, the latent utility differences $\mathbf{y}^*$ are also normally distributed. Then, their mean is zero and their covariance matrix is

$$\boldsymbol{\Sigma}_{y^*} = \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right)\mathbf{A}' + \boldsymbol{\Omega}^2\,. \tag{9}$$

The model just described is an extension of the Thurstonian factor model for polytomous data (Maydeu-Olivares, 2002) to items presented in more than one block. It is also an extension of the Thurstonian IRT model designed for forced-choice blocks (Brown & Maydeu-Olivares, 2011) to ordinal data with possibly intransitive preferences.

Model Estimation

We recognize (9) as the covariance structure of a second-order factor analysis model where **A**, the matrix of fixed contrasts, represents the first-order factor loadings of the

pairwise outcomes on their respective utilities, and $\boldsymbol{\Lambda}$ represents the second-order factor loadings of the utilities on their respective personal attributes. Since the latent utility differences $\mathbf{y}^*$ are assumed to be normally distributed and the observed variables are ordinal, the model is akin to an ordinal (second-order) factor analysis and it may be estimated from polychoric correlations. Importantly, when items are presented one a time as in standard Likert type, $\mathbf{A}$ is an identity matrix and the model reduces to the standard ordinal factor analysis model.

To enable estimation of the covariance structure (9) from ordinal data, the latent utility differences $\mathbf{y}^*$ are standardized using $\mathbf{z}^* = \mathbf{D}\left(\mathbf{y}^* - \boldsymbol{\mu}_{\mathbf{y}^*}\right) = \mathbf{D}\mathbf{y}^*$, where

$\mathbf{D} = \left(\text{Diag}\left(\boldsymbol{\Sigma}_{y^*}\right)\right)^{-\frac{1}{2}}$ is a diagonal matrix with the reciprocals of the standard deviations of $\mathbf{y}^*$ in the diagonal (Maydeu-Olivares, 2002; Maydeu-Olivares & Böckenholt, 2005). Therefore standardized latent difference responses $\mathbf{z}^*$ are multivariate normal with mean zero and correlation matrix

$$\mathbf{P} = \mathbf{D}\left(\boldsymbol{\Sigma}_{y^*}\right)\mathbf{D}. \tag{10}$$

If we organize the thresholds in (3) in a $bn \times C$ matrix $\boldsymbol{\tau}$, then the thresholds relating the standardized latent utility differences $\mathbf{z}^*$ to the observed ordinal variables $\mathbf{y}$ are

$$\boldsymbol{\alpha} = \mathbf{D}\boldsymbol{\tau}. \tag{11}$$

First, the sample thresholds $\hat{\boldsymbol{\alpha}}$ and polychoric correlations $\hat{\mathbf{P}}$ are estimated. Then the model parameters are estimated from these sample statistics using unweighted or diagonally weighted least squares (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). This can be accomplished using standard software such as Mplus (Muthén & Muthén, 2016). When using this program, researchers only need to specify the first-order structure (5) and the second-

order structure (7), as Mplus automatically implements the constraints (10) and (11). Writing Mplus code can be tedious when block size is greater than 2, as many utility contrasts (matrix **A**) have to be specified. An Excel macro that automates writing the full code, including the necessary identification constraints described below, is available from the first author's webpage.

Estimable Parameters and Identification

Although items measuring different personal attributes are often combined in blocks, most questionnaires are constructed so that each item measures only one attribute (utility factor loadings $\mathbf{\Lambda}$ forming "independent clusters"; McDonald, 1999). We provide identification conditions for this case.

As in any other factor analysis model, we begin by setting the metrics for the common factors by setting their variances to one so that $\mathbf{\Phi}$ is a correlation matrix. However, due to the categorical nature of the data, the metrics of the unique factors need to be set as well. To do so, in blocks of size $n \geq 3$, it suffices to set the uniqueness (i.e., variance of the unique factor) of just one item per block to an arbitrary constant. It is usual to set the uniqueness of the last (or the first) item in each block to one. These are the constraints needed to identify the elements of $\mathbf{\Psi}^2$.

The diagonal elements of $\mathbf{\Omega}^2$ capturing the degree of intransitivity in pairwise comparisons can be freely estimated. However, in this case the model has a large number of parameters and may be nearly non-identified in applications (the standard errors for some parameters may be poorly estimated). To reduce the number of parameters, Maydeu-Olivares and Böckenholt (2005) suggested setting all intransitivity variances equal, i.e., $\mathbf{\Omega}^2 = \omega^2\mathbf{I}$.

A special case arises when the block size is $n = 2$. In this case, $\boldsymbol{\Omega}^2 = \mathbf{0}$ as there can be no intransitivity. Also, the two items' unique variances cannot be identified independently, so we set $\boldsymbol{\Psi}^2 = \mathbf{I}$.

A further special case arises when exactly two attributes ($d = 2$) are measured using multidimensional pairs ($n = 2$). Because each pairwise ordinal outcome loads on both factors, this is essentially an exploratory factor model, and additional identification constraints need to be imposed on some factor loadings (Brown & Maydeu-Olivares, 2012).

Person Score Estimation

After the model parameters have been estimated, factor scores for each person may be estimated using maximum likelihood or, alternatively, Bayesian estimation with the multivariate normal prior with covariance matrix $\boldsymbol{\Phi}$. Either the mean of the posterior distribution can be estimated (expected a-posteriori or EAP), or the mode (maximum a-posteriori or MAP). The former can be used in applications with one to three measured attributes; the latter is recommended in applications with many measured attributes. The software we use to fit the graded-preference model, Mplus, conveniently provides MAP scores. When blocks are of size $n = 2$, factor scores cannot be estimated using the ordinal factor model with covariance structure (9) because $\boldsymbol{\Omega}^2 = \mathbf{0}$ (responses cannot be intransitive). In this case, the second-order factor structure (9) needs to be reparameterized as a first-order structure by using

$$\breve{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda} \quad \text{and} \quad \breve{\boldsymbol{\Psi}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\,\mathbf{A}' + \boldsymbol{\Omega}^2, \tag{12}$$

resulting in the Thurstonian IRT model for ordinal data

$$\boldsymbol{\Sigma}_{y^*} = \breve{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\breve{\boldsymbol{\Lambda}}' + \breve{\boldsymbol{\Psi}}. \tag{13}$$

Information, Standard Errors and Reliability

Information and standard errors. In questionnaires measuring personal attributes, it is of interest to evaluate the amount of information that every graded comparison contributes to the measurement of the attributes, and the amount of information that the questionnaire provides as a whole. Because the graded blocks are typically designed to compare items measuring different attributes, the outcomes of comparisons are multidimensional by design, even when items under comparisons are unidimensional. Because test developers typically employ balanced designs in which numbers of comparisons between items measuring different attributes are approximately equal, any subset of graded comparisons indicating a particular attribute will also indicate other attributes. In such *inseparable* designs, no single attribute can be estimated without estimating the whole model. Inevitably then, the measurement errors of all attributes are correlated – and likely highly correlated – therefore not only their variances (as reciprocals to test information functions) but also covariances must be considered (McDonald, 1999). To complicate things further, the outcomes of graded pairs arising from blocks of size $n \geq 3$ indicate not only the common factors (i.e. attributes), but also the unique factors (i.e. utility errors). And since some of the unique factors and common factors are indicated by the same graded pairs, their measurement errors are also correlated. In this situation, covariances of measurement errors for all the independent variables defining the latent space (the common factors and the unique factors) must be considered.

In Appendix A, we provide the item characteristic functions for graded preference models, which are necessary for computation of item information. In Appendix B, we provide a complete solution for computing information and standard errors for graded-preference questionnaire data, which obviously also applies to binary preferences (i.e. forced choice). Past solutions for computing information in forced-choice questionnaires (Brown & Maydeu-

Olivares, 2011; Maydeu-Olivares & Brown, 2010) were incomplete as they only partially accounted for multidimensionality by controlling for relationships between traits using directional information. Moreover, they did not take into account the correlated measurement error in questionnaires using multidimensional comparisons, and in unidimensional blocks of three or more items. The solution proposed in Appendix B computes the item and test information functions as Fisher information matrices, fully accounting for the inherent multidimensionality in the data, and can be applied to both graded and binary comparative designs. To enable implementation of this solution in practice, as an online supplement to this paper, we provide R functions for computing item and test information from the model parameters and MAP scores estimated in Mplus, as well as a sample R code for estimating standard errors (SEs) for these scores.

Reliability. While the availability of SEs for the estimated trait scores of each person is an advantage for individual diagnostics, summarizing the precision of measurement of the questionnaire for a range of trait values may also be of interest. However, if in unidimensional IRT models a curve depicting either the test information function (or the SE function) is a good summary, in the inherently multidimensional Thurstonian models with non-separable designs, trait information may be conditional on all other measured traits (and on some utility errors when the block size is $n \geq 3$). In this case, instead of exact functions, sample-based scatter plots of SEs against the trait of interest, such as one illustrated in Figure 2 panel b, can be helpful.

Another common method of summarizing SEs is the *empirical reliability* index, which is the ratio of true score variance to the sum of true and error variance estimated in a sample. As suggested in Du Toit (2003) for Bayesian EAP or MAP scores[2], which are

[2] This method for computing reliability differs from previous published works on Thurstonian IRT model, where the true score variance was estimated as the difference between the observed MAP score variance and the error variance. Simulations studies show that the method presented here yield results closer to the true values; the improvement is more noticeable when the Bayesian estimator shrinks score estimates significantly.

regressed estimates of latent traits with the shrunken distribution, the true score variance is best estimated directly from the variance of the EAP or MAP score, say $\text{var}\left(\hat{\eta}_{MAP}\right)$, which is conveniently printed in Mplus output. The error variance is the mean of the squared standard errors estimated for the sample (for example, using the supplied R code), yielding

$$\hat{\rho} = \frac{\text{var}\left(\hat{\eta}_{MAP}\right)}{\text{var}\left(\hat{\eta}_{MAP}\right) + \overline{SE^2}\left(\hat{\eta}_{MAP}\right)}. \quad (14)$$

Empirical example: Measuring the Five Factors of Personality Using Graded Preferences

Participants and Materials

Five-hundred-and-ninety-five undergraduate psychology students from the University of Barcelona completed a questionnaire measuring the Five Factors of personality online in return for a comprehensive feedback report. The sample comprised 71.4% female, with average age of 22.8 years (standard deviation of 7.9).

For this study, we modified the Spanish version of the Forced-choice Five Factor markers questionnaire (FCFFM; Brown & Maydeu-Olivares, 2011a) with respect to the response format only. The FCFFM consists of 60 items selected from the International personality Item Pool (IPIP), more specifically from the subset measuring the Five Factor markers (Goldberg, 1992). Each factor is measured with 12 items. The items are organized in $b = 20$ blocks of three items, with the restriction that within a block no two items measure the same factor. We presented the items from each block as $\tilde{n} = 3$ separate paired comparisons, and respondents had to indicate their preference for the item on the left or on the right using five graded options: “much more – a little more – equal – a little more – much more”. To counteract the carry-over effect in paired comparisons with repeated items, we randomized

the presentation of pairs, so that the pairs from the same block did not appear sequentially. In total, respondents were presented with $b \times \tilde{n} = 60$ graded paired comparisons.

After completing the graded preferences, participants were presented with the same 60 items using a standard Likert format, in which they rated the items according to the extent they represented their personality using a 5-point rating scale "very well – quite well – sometimes well, sometimes badly – quite badly – very badly".

Analysis

Likert format. A confirmatory factor model with five latent correlated factors illustrated in Figure 1 (panel a) was fitted to 60 observed *item ratings* coded from 5 ("very well") to 1 ("very badly"). An ordinal factor analysis model was fitted to these data. Every one of the Five Factors was indicated by 12 items and no item was loading on more than one factor. Thus, one factor loading and four thresholds were estimated per item. In total, this model estimated 60 loadings, $4 \times 60 = 240$ thresholds, and 10 inter-factor correlations. This model is equivalent to a five-dimensional Samejima's (1969) normal ogive Graded Response Model.

INSERT FIGURE 1 ABOUT HERE

Graded-block format. The ordinal factor model for graded-block preferences illustrated in Figure 1 (panel b) was fitted to the 60 observed outcomes of paired *graded preferences*, coded from 5 ("much more" preference for first item in the pair) to 1 ("much more" preference for second item). Since the observed variables were results of comparisons of two items, each ordinal outcome was linked to two latent utilities of items under comparison; the first utility positively influencing the outcome, and the second utility

negatively with the effects fixed to unity as per contrast matrix $\mathbf{A_3}$ in (6). The utilities, in turn, were indicators of five latent correlated factors (the Five Factors of personality). The same factorial structure as in the model for Likert ratings was applied to the utility variables: each factor was measured by 12 utilities and no utility was loading on more than one factor. Thus, one factor loading (pertaining to the item utility) was estimated per item and four thresholds were estimated per graded pairwise outcome. Since every block of three items was presented as 3 paired comparisons, transitivity of preferences could not be guaranteed (as it would be in rankings), necessitating an error term for every observed preference outcome. Because it is reasonable to assume an approximately equal degree of intransitivity in all paired comparisons (Maydeu-Olivares & Böckenholt, 2005), all 60 variances of the pairwise errors $\mathbf{e}$ (the diagonal elements of $\mathbf{\Omega}^2$) were constrained equal. To set the metric of the unique factors, we fixed the uniqueness of the last item in each block to one (thus fixing 20 of the 60 diagonal elements of $\mathbf{\Psi}^2$). In total, this model estimated 60 loadings, $4 \times 60 = 240$ thresholds, 10 inter-factor correlations, $60 - 20 = 40$ uniquenesses, and 1 intransitivity variance parameter common to all pairs.

Estimation. Both the Likert and graded-block models were estimated from polychoric correlations in Mplus 7.2, using the Unweighted Least Squares estimator with robust standard errors (denoted ULSMV). To assess goodness of fit, we considered the chi-square statistic (χ^2), and the Root Mean Square Error of Approximation (RMSEA) with values less than .06 indicating good fit (Hu & Bentler, 1999). Recently, it has been suggested to reverse the role of the null and alternative hypotheses when assessing model fit. This is termed a test of not-close fit (MacCallum, Browne, & Sugawara, 1996) and equivalence testing (Yuan, Chan, Marcoulides, & Bentler, 2015), where significant results provide strong support for good fit. With this approach, claims can be made regarding an upper bound on the size of misspecification (*T*-size) as measured by the RMSEA; specifically, the upper limit of

the 90% RMSEA confidence interval printed by Mplus corresponds to 95% confidence in the maximum size of misspecification (Yuan et al., 2015). In addition to these statistical tests of model fit, we also considered a direct measure of discrepancy between the observed and model-implied polychoric correlations – the Standardized Root Mean Square Residual (SRMR[3]) with values less than .08 indicating good fit (Hu & Bentler, 1999).

Person scores and their standard errors. Mplus produced two sets of MAP scores on the Five Factors – one based on the Likert responses, and the other based on the graded-block responses – for each participant. For the graded-block responses, Mplus produced not only the trait scores (second-order factors) but also the utility scores (first-order factors). At the time of writing, Mplus does not compute SEs for MAP scores. SEs for MAP scores for Likert and graded-block formats using respective multivariate normal priors were computed using R functions supplied with this article according to the formulas provided in the Appendix. (Note that the supplied R functions can also be used to compute SEs of MAP scores in the multidimensional ordinal model for Likert items, as it is a special case of our graded preference model when $n = 1$ and the contrast matrix A set to identity matrix).

We estimated the empirical reliabilities of the Five Factor scores measured in the Likert and graded-block models using (14), with the error variance of the MAP scores estimated by squaring and averaging the respective SEs across the whole sample. All these steps are included in the sample R code supplied with this article.

Results

Model fit and parameter estimates. The ordinal factor model applied to the Likert ratings yielded $\chi^2 = 5239$ on 1700 df, $p < 0.001$), a poor fit according to the SRMR = .092, and a barely acceptable approximate fit according to the RMSEA = .059 (90% confidence

[3] To obtain the SRMR in Mplus, MODEL=NOMEANSTRUCTURE setting must be used in the ANALYSIS command.

interval for RMSEA .057-.061). Under the equivalence testing framework, we can be 95% confident that the population RMSEA is no more than .061. Exploring potential reasons for misfit, we examined the model's modification indices (MI). Only five modification indexes exceeded 100; all of them pertained to cross-loadings. For example, the largest MI ($\chi^2 = 197$) was for item "I am always prepared" ("Siempre estoy preparado"), which was designed to measure Conscientiousness, suggesting a cross-loading on Openness. Judging that allowing the suggested cross-loadings would not radically change the model fit or interpretation, we retained the original model. The factor loadings of all the Likert items on the personality factors were in the expected directions and statistically significant. The model-based correlations of the five personality traits for Likert data are given in Table 1, above the diagonal.

The second-order ordinal factor model applied to the graded-block comparisons yielded $\chi^2 = 3874$ (df = 1659, $p < 0.001$), a good fit according to SRMR = .072 and RMSEA = .047 (90% confidence interval .045-.049). Under the equivalence testing framework, we can be 95% confident that the population RMSEA is no more than .049. The a-priori model appeared to fit better to graded-block comparisons than the counterpart model to Likert ratings. The factor loadings of all the first-order utilities on the second-order personality factors were in the expected directions and statistically significant. The model-based correlations between the five personality dimensions in the graded-preference model are given in Table 1 (below the diagonal).

INSERT TABLE 1 ABOUT HERE

It can be seen from Table 1 that the correlations yielded by the Likert and graded-preference models were largely similar; however, the small differences were systematic. The

correlations in the Likert model were always stronger (except the Agreeableness-Neuroticism correlation, which was weaker in the third decimal place for the Likert data – a clearly negligible outlier from this trend). If we reverse the direction of trait Neuroticism, presenting it as Emotional Stability, all inter-trait correlations become positive, yielding the average correlation of .195 in the Likert model and .137 in the graded-preference model. Interestingly, all inter-correlations except those involving Agreeableness, are uniformly larger by about .09 in the Likert model. The correlations involving Agreeableness are very close in the two models.

Standard errors and reliability of factor scores. The standard errors and reliabilities of the MAP Five Factor scores in the Likert and graded-preference models are summarized in Table 2. For comparison, coefficients alpha for sum scores obtained from the Likert items are also provided. We see in Table 2 that the MAP scores in both formats were highly reliable in the range of .8–.9; all scores were slightly more reliable when the Likert format was used (differences in reliabilities around .05). Given the same number of observed variables and the same number of graded categories in both response formats, the slightly more reliable scores with ratings are to be expected since each rating loaded on one factor only, hence providing independent contributions to the reduction of measurement error.

Figure 2 shows the SEs of MAP scores on Neuroticism, plotted against the actual MAP scores, for both response formats. It can be seen that while the Likert data with its separable measurement yield a curve, the graded-preference data with its inseparable measurement yield a scatter, with a curvilinear tendency but significant dispersion reflecting dependencies of the Neuroticism SEs on other variables in the model.

INSERT FIGURE 2 ABOUT HERE

Convergent validity of the factor scores. MAP estimated scores from the Likert and graded-preference measurement models were used to explore the relationships between corresponding personality constructs (hetero-method mono-trait correlations), which are given in Table 2. The estimated trait scores for the same construct correlated highly, and were similar in magnitude to their respective reliability coefficients. The correlation coefficients corrected for unreliability (using the empirical reliability coefficients) are provided in parentheses after the observed value. Except the trait Agreeableness, for which the corrected correlation was .937, the rest of the traits correlated nearly perfectly, suggesting that the same psychological constructs were measured regardless of the response format.

INSERT TABLE 2 ABOUT HERE

Conclusions and Discussion

The present paper introduces an ordinal factor analysis model of graded preferences among pairs of items, where the extent of preference for one or another item can be quantified in terms of ordered categories such as “much more”, “slightly more”, “about the same”, etc. Questionnaires using graded comparisons can be used to assess personality traits, motivations, attitudes, and similar constructs. Items designed to measure different constructs can be combined to create multidimensional graded pairs. Pairs can be formed by simply splitting a pool of items into blocks of two items, in which case no graded pairs have overlapping content. However, the pool of items can also be split into blocks of 3 or more items from which all possible pairs are then drawn (we call this “graded-block” design). In this latter case, graded pairs drawn from the same block have overlapping content with known patterns of dependence. The model we propose for these data is equivalent to an

extension of the Thurstonian IRT model to ordered categorical outcome data. The new contribution of the present paper beyond extending the family of Thurstonian factor and IRT models is the complete solution to the item and test information functions, which are now computed as Fisher information matrices and can be applied to both binary and graded comparative designs.

We believe that when used in the right context, grading of preferences can be superior to both Likert ratings and binary rankings (forced choice). Graded preferences could replace Likert ratings when finer differentiation between judgements is needed, for instance in organizational appraisals where halo effects are common and impact the validity of inferences (Bartram, 2007; Brown, Inceoglu, & Lin, 2017); or in settings where respondents may acquiesce. Graded preferences could also replace forced-choice rankings when the test reliability needs to be increased without increasing the number of item-pairs administered. Indeed, given a fixed number of items, and all other factors held constant, the use of a graded scale over a binary scale is known to increase the amount of information the test provides (Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009).

Graded Preferences versus Likert Ratings

To illustrate the potential advantages and disadvantages of graded preferences as compared to Likert ratings, consider our empirical example where we compared measurement of the Five Factors of personality using the two formats. Both designs had the same number of items (60), the same number of observed variables (60), and the same number of graded options per observed outcome variable (5). As can be seen in Table 2, the empirical reliabilities were over 0.8 in both formats; with ratings still slightly outperforming graded preferences (loss in reliability for each scale was around 0.05). This small loss was due to the multidimensionality inherent to comparative response formats. As explained in the

section on Information, since every graded pair contributes to measurement of more than one common factor (and more than one unique factor), the measurement errors are correlated. To accommodate for this, we evaluate information contribution of every pair to measurement of all the relevant common and unique factors using the Fisher item information matrix, a procedure common in computerized adaptive testing (CAT) applications using multidimensional IRT models. When the measurement errors are correlated, the standard errors of the trait scores are generally larger than in the counterpart Likert questionnaires with factorially pure items, and reliabilities are consequently smaller.

However, the slight loss of information in graded pairs compared to Likert ratings may well be outweighed by potential benefits in reducing unwanted effects such as acquiescence, halo or socially desirable responding. Comparing the inter-trait correlations in the Likert and graded-preference models, we noted that the Likert ratings yielded a slightly stronger positive manifold of correlations among the five personality traits (with Neuroticism reversed to represent Emotional Stability). At the item level, the average model-based correlation between utilities (suitably reversed to measure the desirable poles) was again greater in the Likert model (.168) than in the graded-preference model (.144). It appears that the Likert items elicited utility judgments that were slightly less differentiated than the judgements elicited by the graded pairs. Specifically, the Likert ratings of items indicating the desirable poles of personality traits were more similar to each other, and so were the ratings of items indicating the undesirable poles. This similarity in ratings could not be attributed to acquiescence since it adjusted for item polarities. We believe the more likely reason for less differentiated utility judgements in the Likert version of the FCFFM was socially-desirable responding. The lack of fit and the required cross-loadings in the Likert model also point to an additional source of common variance in the ratings, which our a-priori model did not take into account. It is outside of scope of the present paper to examine alternative models for

Likert ratings, but models exist that incorporate biases as latent "method" variables acting at either the item level as in the random intercept model (Maydeu-Olivares & Coffman, 2006), or at the response category level as in the scoring functions approach (Falk & Cai, 2016). Such models could be used in future research to explore the source and extent of response biases.

We believe that although detectable, response distortions were small in the empirical study presented here because by providing participants with personalized feedback report, we tried to ensure sufficient motivation not to engage in acquiescence and inattentive responding on one hand (Meade & Craig, 2012), and present the true picture of themselves without managing impression on the other hand. The high degree of similarity between the results obtained from absolute and comparative response formats corroborate findings reported in similar low-stakes conditions, for instance in a validation study reported by Brown and Maydeu-Olivares (Brown & Maydeu-Olivares, 2013). However, this degree of similarity is by no means guaranteed, and is actually unusual in medium- or high-stakes assessments (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Brown, Inceoglu, & Lin, 2017; Schmit & Ryan, 1993). Comparing the internal and external validities of scores derived from graded preferences to both Likert ratings and rankings in such contexts would be a good topic for further research.

Graded Preferences versus Binary Preferences (aka Forced Choice)

To illustrate potential advantages of graded preferences in comparison to binary preferences, although we did not collect them in the empirical example presented here, we collapsed the first 3 and the last 2 categories in our graded-pairs data to emulate binary-pairs data. The resulting empirical reliabilities computed in the way described in the present paper but based on two response categories were $\rho_N = .780$, $\rho_E = .811$, $\rho_O = .711$, $\rho_A = .731$ and ρ_C

= .755. Comparing these estimates to their counterparts in Table 2, we can see that the binary choice yielded the reliability loss of between .07 and .10 compared to the graded preferences. This degree of information loss is greater than the loss we observed in using graded comparisons instead of Likert ratings.

Although the information increase is undoubtedly an advantage of graded over binary preferences, the use of ordinal categories to grade one's preferences could potentially open the door to response biases we typically associate with Likert scales. In theory, idiosyncratic uses of the response categories are possible in the graded-preference format – for example, preferring the extreme categories or the middle categories regardless of the item content. However, these styles would influence the judgements of utility differences rather than utilities themselves. Whether this type of distortion will prove problematic in certain contexts, for example cross-cultural research notoriously vulnerable to systematic differences in response styles, and how it will compare to the Likert scales remains to be seen and is also a good topic for future studies. To conclude, when designed well and used in the right context, graded preferences can be an attractive alternative to either Likert ratings or rankings. They can have the benefits of rankings in differentiating well between responses, and the benefits of ratings in allowing respondents to express the extent of preference, thus increasing information and measurement precision. In the present article, we provided tools for fitting factor analysis models to graded preference data, estimating person scores that are free of problems of ipsative data, and assessing the measurement precision of these scores. Equipped with these tools, researchers and test developers can evaluate the performance of various questionnaire designs and select the best one for their required assessment context. We are looking forward to new developments in this area.

Acknowledgments

Work on this paper was partly supported by the National Science Foundation grant #1659936 awarded to the second author.

References

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*(3), 263–272. http://doi.org/10.1111/j.1468-2389.2007.00386.x

Bartram, D., & Brown, A. (2003). *Test-taker reactions to online completion of the OPQ32i*. Thames Ditton.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. http://doi.org/10.1111/j.1468-2389.2006.00354.x

Böckenholt, U., & Dillon, W. (1997). Modeling within-subject dependencies in ordinal paired comparison data. *Psychometrika*, *62*(3), 411–434. http://doi.org/10.1007/bf02294559

Brown, A. (2016a). Item Response Models for Forced-Choice Questionnaires: A Common Framework. *Psychometrika*, *81*(1), 135–160. http://doi.org/10.1007/s11336-014-9434-9

Brown, A. (2016b). Thurstonian Scaling of Compositional Questionnaire Data. *Multivariate Behavioral Research*, *51*(2–3), 345–356. http://doi.org/10.1080/00273171.2016.1150152

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. *Organizational Research Methods*, *20*(1), 121–148. http://doi.org/10.1177/1094428116668036

Brown, A., & Maydeu-Olivares, A. (2011). Forced-choice Five Factor markers questionnaire. http://doi.org/10.1037/t05430-000

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. http://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–47. http://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires. *Psychological Methods*, *18*(1), 36–52. http://doi.org/10.1037/a0030641

Cheung, M. W.-L., & Chan, W. (2002). Reducing Uniform Response Bias With Ipsative Measurement in Multiple-Group Confirmatory Factor Analysis. *Structural Equation Modeling*, *9*(1), 55–77. http://doi.org/10.1207/S15328007SEM0901_4

Clemans, W. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monograph No. 14*.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, *69*(1), 41–47. http://doi.org/10.1111/j.2044-8325.1996.tb00598.x

du Toit (Ed.), M. (2003). *IRT from SSI*. Lincolnwood, IL: SSI International.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor Analysis of Ipsative Measures. *Multivariate Behavioral Research*, *29*(2), 115–126. http://doi.org/10.1207/s15327906mbr2901_4

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. http://doi.org/10.1037/met0000059

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*(4), 625–641. http://doi.org/10.1080/10705510903203573

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. http://doi.org/10.1037/1040-3590.4.1.26

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. http://doi.org/10.1080/10705519909540118

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. http://doi.org/10.1037//1082-989X.1.2.130

Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data. *Mathematical Social Sciences*, *43*(3), 467–483. http://doi.org/10.1016/S0165-4896(02)00017-3

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*(3), 285–304. http://doi.org/10.1037/1082-989X.10.3.285

Maydeu-Olivares, A., & Brown, A. (2010). Item Response Modeling of Paired Comparison and Ranking Data. *Multivariate Behavioral Research*, *45*(6), 935–974. http://doi.org/10.1080/00273171.2010.531231

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–62. http://doi.org/10.1037/1082-989X.11.4.344

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D.

(2009). The effect of varying the number of response alternatives in rating scales: experimental evidence from intra-individual effects. *Behavior Research Methods*, *41*(2), 295–308. http://doi.org/10.3758/BRM.41.2.295

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

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. http://doi.org/10.1037/a0028085

Mulder, J., & van der Linden, W. J. (2009). Multidimensional Adaptive Testing with Optimal Design Criteria for Item Selection. *Psychometrika*, *74*(2), 273–296. http://doi.org/10.1007/s11336-008-9097-5

Muthén, L. K., & Muthén, B. O. (2016). Mplus. Version 7.2 [Computer program]. Los Angeles, CA: Muthén & Muthén.

Samejima, F. (1969). Estimation of latent ability using a response pattern of granded scores. *Psychometrika Monograph*, *17*(17).

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*(6), 966–974. http://doi.org/10.1037/0021-9010.78.6.966

Takane, Y. (1989). Analysis of covariance structures and probabilistic binary choice data. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 139–160). Elsevier B.V.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. http://doi.org/10.1037/h0070288

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2015). Assessing Structural Equation Models by Equivalence Testing With Adjusted Fit Indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *5511*(November), 1–12. http://doi.org/10.1080/10705511.2015.1065414

Appendix A. Item Characteristic Functions in Graded-Block Models

When block size $n = 2$, the intransitivity errors of pairwise preferences $y^*_{\{i,k\}}$ described by (8) are zero, and the only sources of error in measuring the attributes are the utility unique factors ε_i and ε_k with variances ψ_i and ψ_k respectively. According to the ordinal factor analysis model (or equivalently, the normal ogive Graded Response Model, Samejima, 1969), which we assume for the threshold process (3), the probability of selecting **category above *c*** in graded comparison $\{i, k\}$ is conditional on the common factors only,

$$\begin{aligned} \mathrm{P}^*_{\{i,k\}c}(\boldsymbol{\eta}) &= \Pr\left(y_{\{i,k\}} \geq c \middle| \boldsymbol{\eta}\right) = \Phi\left(\frac{-\tau_{\{i,k\}c} + \mathbf{A}_{\{i,k\}}\mathbf{t}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) = \\ &= \Phi\left(\frac{-\tau_{\{i,k\}c} + (\mathbf{A}\boldsymbol{\Lambda})_{\{i,k\}}\boldsymbol{\eta}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) = \Phi(x_c). \end{aligned} \tag{15}$$

In the above expression, $\Phi(x_c)$ is the normal distribution function evaluated at x_c, $\mathbf{A}\boldsymbol{\Lambda}_{\{i,k\}}$ and $\mathbf{A}_{\{i,k\}}$ indicate the row in matrices $\mathbf{A}\boldsymbol{\Lambda}$ and $\mathbf{A}$ corresponding to pair $\{i, k\}$, and ψ_i^2 and ψ_k^2 are the variances of the utility unique factors ε_i and ε_k. As explained in the model identification section, when n =2, the uniquenesses cannot be identified independently, and have to be all set to arbitrary constants, typically $\boldsymbol{\Psi}^2 = \mathbf{I}$. The threshold $\tau_{\{i,k\}c}$ separates category c from category $c + 1$, and because the categories are bounded between 1 and C, we have $\mathrm{P}^*_{\{i,k\}0} = 1$ and $\mathrm{P}^*_{\{i,k\}C} = 0$. With this, the probability of selecting category c is

$$\mathrm{P}_{\{i,k\}c} = \mathrm{P}^*_{\{i,k\}c-1} - \mathrm{P}^*_{\{i,k\}c}. \tag{16}$$

When items i and k are factorially simple, measuring attributes η_a and η_b respectively, the probability (15) simplifies to

$$\mathrm{P}^{*}_{\{i,k\}c}\left(\eta_a,\eta_b\right)=\Pr\left(y_{\{i,k\}}\geq c\middle|\eta_a,\eta_b\right)=\Phi\left(\frac{-\tau_{\{i,k\}c}+\lambda_{ia}\eta_a-\lambda_{kb}\eta_b}{\sqrt{\psi_i^2+\psi_k^2}}\right). \tag{17}$$

When block size n ≥ 3, the response tendency variable $y^{*}_{\{i,k\}}$ described by (8) is determined by the utility common and unique factors, and has the error attributed to intransitive preferences. The probability of selecting **category above *c*** in graded comparison $\{i, k\}$ is then conditional on both the common and unique factors,

$$\begin{aligned}\mathrm{P}^{*}_{\{i,k\}c}\left(\boldsymbol{\eta},\boldsymbol{\varepsilon}\right)&=\Pr\left(y_{\{i,k\}}>c\middle|\boldsymbol{\eta},\boldsymbol{\varepsilon}\right)=\Phi\left(\frac{-\tau_{\{i,k\}c}+\mathbf{A}_{\{i,k\}}\mathbf{t}}{\sqrt{\omega^2_{\{i,k\}}}}\right)=\\&=\Phi\left(\frac{-\tau_{\{i,k\}c}+\left(\mathbf{A\Lambda}\right)_{\{i,k\}}\boldsymbol{\eta}+\mathbf{A}_{\{i,k\}}\boldsymbol{\varepsilon}}{\sqrt{\omega^2_{\{i,k\}}}}\right)=\Phi\left(z_c\right)\end{aligned}. \tag{18}$$

In the above expression, $\Phi\left(z_c\right)$ is the cumulative normal function evaluated at z_c, $\mathbf{A\Lambda}_{\{i,k\}}$ and $\mathbf{A}_{\{i,k\}}$ indicate the row in matrices $\mathbf{A\Lambda}$ and $\mathbf{A}$ corresponding to pair $\{i, k\}$, and $\omega^2_{\{i,k\}}$ is the variance of the intransitivity error $e_{\{i,k\}}$. In a test with factorially simple items, the utilities of items involved in comparison $\{i, k\}$ are indicators of one factor each – let us call them η_a and η_b respectively – and the probability (18) simplifies to

$$\mathrm{P}^{*}_{\{i,k\}c}\left(\eta_a,\eta_b,\varepsilon_i,\varepsilon_k\right)=\Pr\left(y_{\{i,k\}}>c\middle|\eta_a,\eta_b,\varepsilon_i,\varepsilon_k\right)=\Phi\left(\frac{-\tau_{\{i,k\}c}+\lambda_{ia}\eta_a-\lambda_{kb}\eta_b+\varepsilon_i-\varepsilon_k}{\sqrt{\omega^2_{\{i,k\}}}}\right). \tag{19}$$

Appendix B. Item and Test Information in Graded-Block Models

Item information. To evaluate the amount of information each observed variable (graded pair) supplies about latent factors in the Thurstonian model for graded comparisons,

we use Fisher information matrices. The Fisher information matrix for a graded-response pair $\{i, k\}$ in a generic measurement model with r factors $\mathbf{F}$ is an $r \times r$ matrix

$$\mathbf{I}_{\{i,k\}}(\mathbf{F}) = \sum_{c=1}^{C} \frac{\partial^2 \mathrm{P}_{\{i,k\}c} \big/ \partial\mathbf{F}\partial\mathbf{F}^{\mathrm{T}}}{\mathrm{P}_{\{i,k\}c}} = \sum_{c=1}^{C} \frac{\partial^2 \left(\mathrm{P}^*_{\{i,k\}c-1} - \mathrm{P}^*_{\{i,k\}c}\right) \big/ \partial\mathbf{F}\partial\mathbf{F}^{\mathrm{T}}}{\mathrm{P}^*_{\{i,k\}c-1} - \mathrm{P}^*_{\{i,k\}c}}. \tag{20}$$

The latent factor spaces defined by independent latent variables necessary to model block sizes $n \geq 3$ and $n = 2$, however, differ. While the block size $n = 2$ requires the factor space with the common factors $\boldsymbol{\eta}$ only as shown in (15), the block size $n \geq 3$ defines the space with the common factors $\boldsymbol{\eta}$ and unique factors $\boldsymbol{\varepsilon}$ as shown in (18). Therefore, below we provide expressions for Fisher information matrices according to the relevant model.

When block size $n = 2$, the latent factor space $\mathbf{F}$ in (20) includes only d common factors $\boldsymbol{\eta}$ representing the attributes. Denoting x_c the category-dependent argument of the cumulative category probability in (15), the partial derivative with respect to any common factor η_a is

$$\partial \mathrm{P}^*_{\{i,k\}c}(\boldsymbol{\eta}) \big/ \partial\eta_a = \partial\Phi(x_c) \big/ \partial\eta_a = \frac{\lambda_{ia} - \lambda_{ka}}{\sqrt{\psi_i^2 + \psi_k^2}} \phi(x_c) \tag{21}$$

where $\phi(x_c)$ is the normal density function evaluated at x_c, and λ_{ia} and λ_{ka} are the respective factor loadings of items i and k on factor η_a. These factor loadings may of course be zero if neither i nor k measure η_a; only one of these loadings are non-zero if only one of the items i or k measure η_a; and both loadings are non-zero if both i and k measure η_a; (i.e. in a unidimensional comparison).

Using (15) and (21), the Fisher information matrix for pair $\{i, k\}$ is a $d \times d$ matrix

$$\mathbf{I}_{\{i,k\}}(\boldsymbol{\eta}) = \frac{(\mathbf{A}\boldsymbol{\Lambda})_{\{i,k\}}(\mathbf{A}\boldsymbol{\Lambda})_{\{i,k\}}^{\mathrm{T}}}{\psi_i^2 + \psi_k^2} \sum_{c=1}^{C} \frac{\left(\phi(x_{c-1}) - \phi(x_c)\right)^2}{\Phi(x_{c-1}) - \Phi(x_c)}, \tag{22}$$

where $(\mathbf{A}\boldsymbol{\Lambda})_{\{i,k\}}$ is the vector (column) corresponding to the $\{i, k\}^{\text{th}}$ row of matrix $\mathbf{A}\boldsymbol{\Lambda}$. For example, consider a pair with factorially pure items $i = 1$ and $k = 2$ measuring traits η_1 and η_2, with factor loadings λ_1 and λ_2 respectively. The information matrix for this pair will have only four non-zero entries:

$$\mathbf{I}_{\{1,2\}}^{n=2} = \frac{1}{\psi_1^2 + \psi_2^2} \begin{pmatrix} \lambda_1^2 & -\lambda_1\lambda_2 & \cdots & \\ -\lambda_1\lambda_2 & \lambda_2^2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \end{pmatrix} \quad \frac{\left(\phi(x_{c-1}) - \phi(x_c)\right)^2}{{}_{c-1}) - \Phi(x_c)} \tag{23}$$

When block size $n \geq 3$, the latent factor space **F** in (20) includes d common factors $\boldsymbol{\eta}$ representing the attributes and bn unique factors $\boldsymbol{\varepsilon}$ representing the utility errors. Denoting z_c the category-dependent argument of the cumulative category probability in (18), the partial derivative with respect to any common factor η_a is (McDonald, 1999)

$$\partial \mathrm{P}_{\{i,k\}c}^{*}(\boldsymbol{\eta}, \boldsymbol{\varepsilon}) \big/ \partial \eta_a = \partial \Phi(z_c) \big/ \partial \eta_a = \frac{\lambda_{ia} - \lambda_{ka}}{\sqrt{\omega_{\{i,k\}}^2}} \phi(z_c), \tag{24}$$

where $\phi(z_c)$ is the normal density function evaluated at z_c; and λ_{ia} and λ_{ka} are the respective factor loadings of items i and k on factor η_a. Again, any or both of these factor loadings may be zero depending on whether items i and k measure η_a. The partial derivatives with respect to the unique factors of items not involved in the comparison are zero, and the only non-zero derivatives are with respect to ε_i and ε_k:

$$\partial \mathrm{P}^{*}_{\{i,k\}c}\left(\boldsymbol{\eta},\boldsymbol{\varepsilon}\right)\big/\partial\varepsilon_{i} = \frac{1}{\sqrt{\omega^{2}_{\{i,k\}}}}\phi\left(z_{c}\right) \quad \text{and} \quad \partial \mathrm{P}^{*}_{\{i,k\}c}\left(\boldsymbol{\eta},\boldsymbol{\varepsilon}\right)\big/\partial\varepsilon_{k} = \frac{-1}{\sqrt{\omega^{2}_{\{i,k\}}}}\phi\left(z_{c}\right). \tag{25}$$

Using (18) and (25) the Fisher information matrix for graded pair $\{i, k\}$ in (20) simplifies to the $(d + bn) \times (d + bn)$ matrix

$$\mathbf{I}^{n\geq 3}_{\{i,k\}}\left(\boldsymbol{\eta},\boldsymbol{\varepsilon}\right) = \frac{1}{\omega^{2}_{\{i,k\}}}\begin{pmatrix} \mathbf{A}\boldsymbol{\Lambda}_{\{i,k\}}\mathbf{A}\boldsymbol{\Lambda}^{\mathrm{T}}_{\{i,k\}} & \mathbf{A}\boldsymbol{\Lambda}_{\{i,k\}}\mathbf{A}^{\mathrm{T}}_{\{i,k\}} \\ \mathbf{A}_{\{i,k\}}\mathbf{A}\boldsymbol{\Lambda}^{\mathrm{T}}_{\{i,k\}} & \mathbf{A}_{\{i,k\}}\mathbf{A}^{\mathrm{T}}_{\{i,k\}} \end{pmatrix}\sum_{c=1}^{C}\frac{\left(\phi\left(z_{c-1}\right)-\phi\left(z_{c}\right)\right)^{2}}{\Phi\left(z_{c-1}\right)-\Phi\left(z_{c}\right)}. \tag{26}$$

The matrix consists of four distinct blocks: (1) a $d \times d$ top left block corresponding to the products of pair loadings on the common factors, (2) a $bn \times bn$ bottom right block corresponding to the products of pair loadings on the unique factors, and (3) and (4) symmetrical blocks of sizes $d \times bn$ and $bn \times d$ corresponding to the cross-products of pair loadings on the common and unique factors.

As the counterpart example to (23), below is the Fisher information matrix for factorially pure items $i = 1$ and $k = 2$ measuring traits η_1 and η_2, with factor loadings λ_1 and λ_2 respectively. As can be seen, the matrix has only four non-zero entries in each block:

$$\mathbf{I}^{n\geq 3}_{\{1,2\}} = \frac{1}{\omega^{2}_{\{1,2\}}}\left(\begin{array}{cccccccc} \lambda_1^2 & -\lambda_1\lambda_2 & \cdots & & & & \cdots & \\ -\lambda_1\lambda_2 & \lambda_2^2 & \cdots & & & & \cdots & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & & & & \cdots & \\ \lambda_1 & -\lambda_2 & \cdots & & & & \cdots & \\ -\lambda_1 & \lambda_2 & \cdots & & & & \cdots & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & & \cdots & \cdots & \end{array}\right) \qquad \frac{-\phi\left(z_{c}\right)\big)^{2}}{\big)-\Phi\left(z_{c}\right)}. \tag{27}$$

Test information. Since for all block sizes, the response variables $y_{\{i,\,k\}}$ are fully conditioned on all contributing influences including the utility errors in (18), the item information functions are additive. This is a major advantage of the graded block design over

the binary choice designs described in Brown and Maydeu-Olivares (2011)[4]. To evaluate the amount of information for the whole questionnaire, we sum the Fisher information matrices for all the graded pairs. This summation takes care of zero and non-zero loading patterns on all latent factors, therefore providing a convenient summary for the whole test.

When Bayesian methods such as MAP are used for score estimation, the prior information must be added to the Fisher (maximum likelihood) test information to compute the posterior test information. For the block size $n = 2$, the factorial space is defined only by the common factors $\boldsymbol{\eta}$, so the posterior information matrix is derived by adding their inverted covariance matrix $\boldsymbol{\Phi}^{-1}$ (Du Toit, 2003)

$$\mathbf{I}_P^{n=2}(\boldsymbol{\eta}) = \sum_{\{i,k\}} \mathbf{I}_{\{i,k\}}^{n=2}(\boldsymbol{\eta}) + \boldsymbol{\Phi}^{-1}. \tag{28}$$

For blocks of size $n \geq 3$, the factor space includes also the unique factors $\boldsymbol{\varepsilon}$, but since the common and unique factors are uncorrelated, the prior covariance matrix is a block-diagonal matrix with $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}^2$ on the diagonal. The inverted covariance matrix is a block-diagonal matrix with $\boldsymbol{\Phi}^{-1}$ and $\left(\boldsymbol{\Psi}^2\right)^{-1}$ on the diagonal, and the posterior information matrix is

$$\mathbf{I}_P^{n\geq 3}(\boldsymbol{\eta},\boldsymbol{\varepsilon}) = \sum_{\{i,k\}} \mathbf{I}_{\{i,k\}}^{n\geq 3}(\boldsymbol{\eta},\boldsymbol{\varepsilon}) + \begin{pmatrix} \boldsymbol{\Phi}^{-1} & 0 \\ 0 & \left(\boldsymbol{\Psi}^2\right)^{-1} \end{pmatrix}. \tag{29}$$

Rank of Fisher Test Information Matrix. While the Fisher item information matrix always has rank 1 (Mulder & van der Linden, 2009), the maximum likelihood test

[4] In binary choice designs using rankings, the intransitivity errors are zero and the ICC cannot be conditioned on the utility errors as in (18). Instead, the expression (15) is used, despite local dependencies existing between pairs involving the same utilities.

information matrix for block size $n = 2$, which is the sum of the item information matrices described by (22) generally has the full rank d, and therefore is invertible. This is because the matrix $\mathbf{A\Lambda}$ is of full rank, d, unless the test items have the discrimination parameters with the same proportional relationship (Brown, 2016a). Adding the posterior information matrix $\mathbf{\Phi}^{-1}$ preserves the full rank. However, in blocks of size $n \geq 3$, the maximum likelihood test information matrix, which is the sum of the matrices (26) is not of full rank. This is the result of the reduced column-rank of blocks (6) in the contrast matrix $\mathbf{A}$ (Maydeu-Olivares, 1999), which determines the bottom-right block $\mathbf{A}_{\{i,k\}}\mathbf{A}_{\{i,k\}}^{\mathrm{T}}$ of the Fisher information matrix. For instance, matrix $\mathbf{A}_3$ in (6) has rank 2 rather than 3 (the number of columns, also the number of utilities). Because the contrast matrices are identical for every block, the sum of all item information matrices $\mathbf{I}^{n=3}$ has a reduced rank and is not invertible, and the SEs of the maximum likelihood factor scores cannot be computed using this method. However, the posterior test information matrices for any block size, $\mathbf{I}_P^{n=2}$ and $\mathbf{I}_P^{n\geq 3}$, are generally of full rank, therefore they can be inverted to compute the SEs of MAP scores.

Table 1

Correlations between the latent Five Factors underlying Likert ratings and graded preferences in the empirical example

	N	E	O	A	C
Neuroticism (N)		-.239**	-.253**	-.091*	-.157**
Extraversion (E)	-.156**		.290**	.426**	.119*
Openness (O)	-.167**	.208**		.163**	.083
Agreeableness (A)	-.097*	.422**	.138**		.129**
Conscientiousness (C)	-.065	.027	.010	.080	

Note: The mono-method hetero-trait latent correlations from the Likert model are **above** the diagonal, from the graded-preference model are **below** the diagonal. ** Correlations are significant at the .01 level, two-tailed. * Correlations significant at the .05 level, two-tailed.

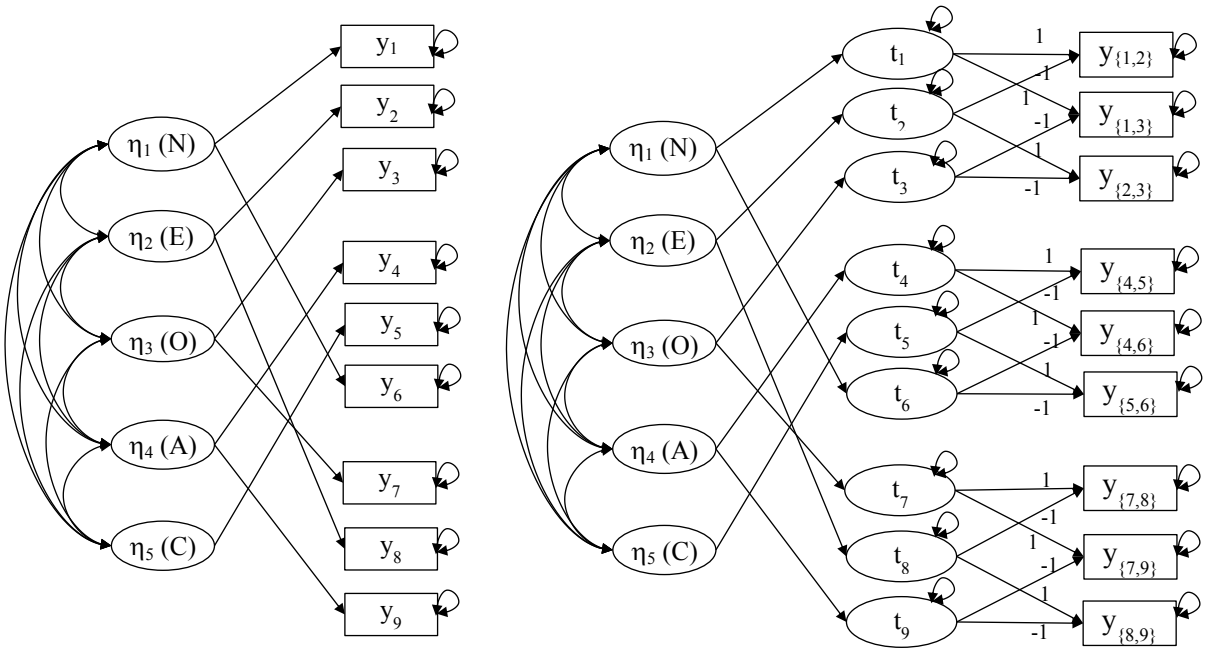
Table 2

Standard errors of MAP scores, reliabilities and mono-method hetero-trait correlations of the Five Factor scores based on Likert items and Graded Preferences in the empirical example

		Likert ratings			Graded preferences			
	α	$\overline{SE^2}(\hat{\eta})$	$\text{var}(\hat{\eta})$	Emp. reliability	$\overline{SE^2}(\hat{\eta})$	$\text{var}(\hat{\eta})$	Emp. reliability	$\text{corr}(\hat{\eta}_L, \hat{\eta}_{GP})$
Neuroticism (N)	.901	.082	.915	.918	.134	.793	.855	.891 (1.005)
Extraversion (E)	.920	.072	.923	.928	.112	.823	.880	.893 (.988)
Openness (O)	.859	.119	.900	.883	.188	.753	.800	.837 (.996)
Agreeableness (A)	.906	.095	.870	.902	.159	.770	.829	.810 (.937)
Conscientiousness (C)	.893	.099	.909	.902	.161	.799	.832	.852 (.984)

Note: L = Likert; GP = Graded Preferences. Observed correlations between the estimated factor scores in the two measurement models are shown; these correlations corrected for unreliability of both measures are in parentheses.

Figure 1. A fragment of measurement models for data collected in the Five Factor questionnaire example (only data from first 9 items are shown)

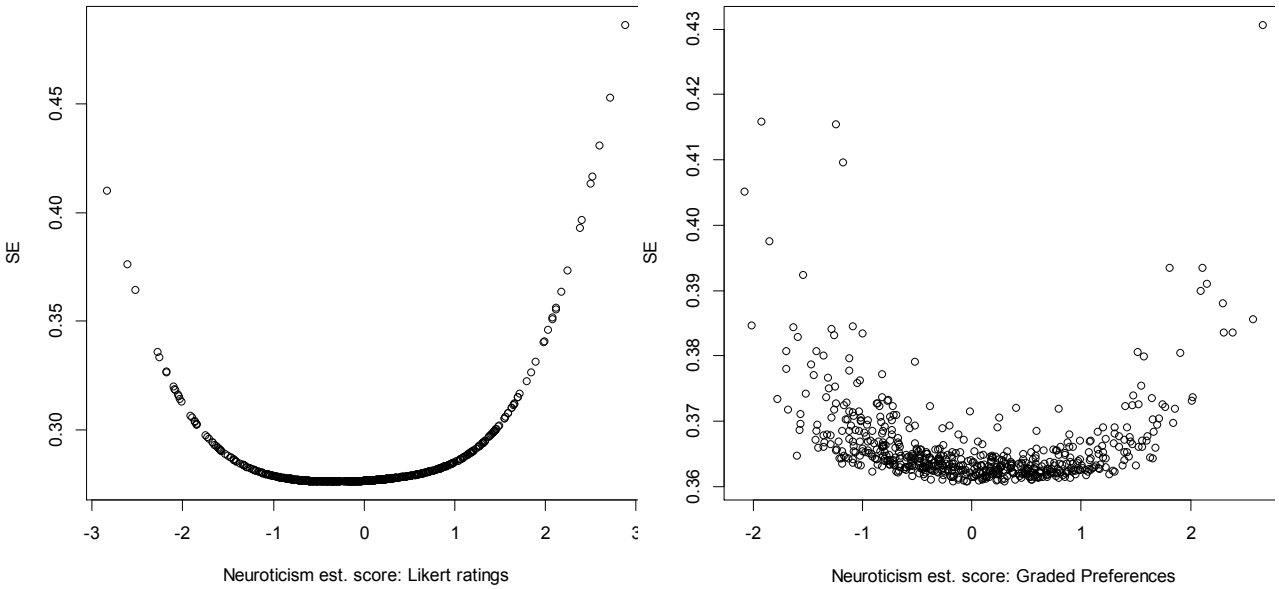


a. Model for Likert data

b. Model for graded-preference data

Note. N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness.

Figure 2. Standard errors of the Neuroticism MAP scores for data collected in the Five Factor questionnaire example



a. Likert data

b. Graded-preference data