

Bienio 2007-2010



UNIVERSITAT DE BARCELONA



Facultad de Psicología

**Departamento de Personalidad, Evaluación
y Tratamientos Psicológicos**

Doctorando: *Anna Brown*

Director: *Prof. Alberto Maydeu-Olivares*

How Item Response Theory can solve
problems of ipsative data

Universidad de Barcelona

Facultad de Psicología

Departamento de Personalidad, Evaluación y Tratamientos Psicológicos

Tesis Doctoral

How Item Response Theory can solve problems of ipsative data

Doctorado de Personalidad, Desarrollo y Comportamiento Anormal

Bienio 2007-2009

Barcelona, 2010

Doctorando: *Anna Brown*

Director: *Prof. Alberto Maydeu-Olivares*

First prompted by the fact of aviation, I have applied the laws of the resistance of air to insects, and I arrived, [with my assistant engineer], at the conclusion that their flight is impossible.

Antoine Magnan, entomologist, 1934

That was a time when we were just beginning to think we understood aerodynamic principles, as applied to fixed-wing aircraft, but scientists recognized their limitations in applying the principles to the birds and insects and other creatures in the natural world. I'm sure no one, including the bees, seriously doubted that insects can fly. Now we're beginning to understand why.

Z. Jane Wang, physicist, 2000

Table of Contents

Overview	1
Abstract.....	3
Introduction.....	4
Single-stimulus response format	4
Response biases affecting single-stimulus items.....	5
Forced-choice format.....	7
Problems of ipsative data.....	8
<i>Relative nature of scores.....</i>	<i>9</i>
<i>Distorted construct validity</i>	<i>10</i>
<i>Distorted reliability estimates</i>	<i>11</i>
Inadequacy of the classical methods of scoring forced-choice items.....	13
IRT approaches to scoring forced-choice items.....	14
The need to model forced-choice dominance items.....	16
Thurstone’s framework for comparative judgment.....	17
Method	20
Binary coding of forced-choice response data	20
Thurstonian factor models for forced-choice items	22
<i>Response model for ranking.....</i>	<i>22</i>
<i>Items as indicators of latent traits</i>	<i>24</i>
Thurstonian Models for forced-choice items as IRT models.....	27
<i>Reparameterized model (first-order Thurstonian IRT factor model).....</i>	<i>27</i>
<i>Identification of Thurstonian IRT models for forced-choice questionnaires</i>	<i>31</i>
<i>Item characteristic function</i>	<i>32</i>
<i>Estimation of Thurstonian IRT models for forced-choice questionnaires.....</i>	<i>34</i>
<i>Latent trait estimation.....</i>	<i>36</i>
<i>Information functions and reliability estimation</i>	<i>37</i>
<i>Response biases and forced-choice format.....</i>	<i>42</i>
Simulation studies.....	45
Simulation study 1. A forced-choice questionnaire measuring 2 traits	46

<i>Design 1. Questionnaire with positively keyed items only</i>	49
<i>Design 2. Questionnaire with both positively and negatively keyed items</i>	53
Simulation study 2. A forced-choice questionnaire measuring 5 traits	58
<i>Design 1. Blocks of 2 items (pairs)</i>	59
<i>Design 2. Blocks of 3 items (triplets)</i>	63
<i>Design 3. Blocks of 4 items (quads)</i>	65
Empirical applications	67
Application 1. A Big Five questionnaire constructed from IPIP items.....	67
<i>Instrument</i>	67
<i>Sample 1 – English version</i>	68
<i>Sample 2 – Spanish version</i>	68
<i>IRT model estimation for forced-choice and single-stimulus responses</i>	68
<i>Correlation patterns</i>	69
<i>Empirical reliability and ordering of respondents</i>	70
Application 2. Customer Contact Styles Questionnaire	72
<i>Instrument</i>	72
<i>Sample</i>	74
<i>Properties of the classical ipsative and normative scores</i>	74
<i>IRT model estimation for forced-choice and single-stimulus responses</i>	77
<i>Empirical reliability and standard error of measurement</i>	79
<i>Ordering of respondents</i>	81
<i>Correlation patterns and principal components</i>	84
<i>Individual profiles</i>	87
Typical profiles	91
Extreme profiles.....	93
Discussion	95
<i>Model and parameter estimation</i>	95
<i>Recommendations for forced-choice questionnaire design</i>	96
Keyed direction of items	96
Number of traits	97
Correlations between traits.....	99
Block size.....	100
<i>Information and test reliability</i>	100

<i>Response biases and application results</i>	101
<i>Future research directions</i>	104
Conclusions	105
References	106
Appendix A: Violation of Alpha’s consistent coding assumption in MFC questionnaires	112
Appendix B: <i>Mplus</i> syntax for the Thurstonian second-order formulation of the example model	113
Appendix C: <i>Mplus</i> syntax for the Thurstonian IRT formulation of the example model	115
Appendix D: Designs involving blocks of 2 items (pairs)	117
Appendix E: Posterior MAP information for a trait in MFC questionnaire	118
Appendix F: The Big Five questionnaire with IPIP items	120

List of Tables

Table 1: <i>True item parameters for the short questionnaire (12 item-pairs) using both positively and negatively keyed items; simulation with 2 traits</i>	48
Table 2: <i>Goodness of fit in the simulation studies with 2 traits</i>	50
Table 3: <i>Parameter estimates in the simulation studies with 2 traits.....</i>	51
Table 4: <i>Test reliabilities in the simulation with 2 traits; questionnaire with positively and negatively keyed items (first replication).....</i>	57
Table 5: <i>True trait correlations in the simulation studies with 5 traits.....</i>	59
Table 6: <i>Goodness of fit in the simulation studies with 5 traits</i>	61
Table 7: <i>Average relative bias for parameter estimates and standard errors in the simulation studies with 5 traits</i>	62
Table 8: <i>Test reliabilities in the simulation studies with 5 traits (first replication).....</i>	63
Table 9: <i>Estimated correlations between the Big Five markers based on the single-stimulus and forced-choice questionnaires in the empirical example.....</i>	70
Table 10: <i>Reliabilities and correlations between the single-stimulus and forced-choice Big Five marker traits in the empirical example.....</i>	71
Table 11: <i>Short descriptions of the 16 traits measured by the Customer Contact Styles Questionnaire (CCSQ).....</i>	73
Table 12: <i>Rotated pattern matrix and component correlations for classical ipsative scores in the CCSQ Application</i>	76
Table 13: <i>Rotated pattern matrix and factor correlations for the classical normative scores in the CCSQ Application</i>	77

Table 14: <i>Reliabilities of the classical scores, IRT-based empirical reliabilities and standard errors in the CCSQ Application.....</i>	80
Table 15: <i>Correlations between classical scores and IRT-based scores in the CCSQ Application</i>	82
Table 16: <i>Estimated correlations between the CCSQ scales based on the single-stimulus and forced-choice responses.....</i>	83
Table 17: <i>Rotated pattern matrix for IRT scored single-stimulus ratings in the CCSQ Application</i>	85
Table 18: <i>Rotated pattern matrix for IRT scored forced-choice ratings in the CCSQ Application</i>	86
Table 19: <i>Average of individual profile correlations and distances for classical and IRT scores in the CCSQ Application.....</i>	90

List of Figures

Figure 1: <i>Thurstonian second-order factor model for a questionnaire with 3 traits and 3 blocks of 3 items</i>	26
Figure 2: <i>Thurstonian IRT model for a questionnaire with 3 traits and 3 blocks of 3 items</i>	30
Figure 3: <i>Item Characteristic Surface (ICS) for the binary outcome $\{i5, i6\}$ for the simulation with 2 uncorrelated traits</i>	35
Figure 4: <i>Item Information Surfaces (IIS) in directions of Trait 1 and Trait 2 for the binary outcome $\{i5, i6\}$ for the simulation with 2 uncorrelated traits</i>	39
Figure 5: <i>Scatterplot of MAP estimated trait scores vs. true latent trait scores for the simulation with 2 uncorrelated traits</i>	54
Figure 6: <i>MAP test information function for the simulation with 2 uncorrelated traits</i>	56
Figure 7: <i>Distributions of individual average profile scores based on IRT and CTT scoring of forced-choice responses in the CCSQ Application</i>	87
Figure 8: <i>Distributions of profile similarity coefficients for IRT-scored single-stimulus and forced-choice responses in the CCSQ Application</i>	89
Figure 9: <i>Sample CCSQ personality profiles based on IRT scores and classical ipsative scores (typical cases)</i>	92
Figure 10: <i>Sample CCSQ personality profiles based on IRT scores and classical ipsative scores (extreme cases)</i>	94

Overview

This dissertation is a result of 4-year research into modeling of preference decisions as applied to forced-choice personality questionnaires. Personality is the main focus of this research; however, most of the theory outlined below will also apply to assessment of motivation, interests, attitudes etc.

The dissertation is structured as follows.

In the **Introduction**, the forced-choice format is introduced and its advantages in reduction of response biases are discussed. The conventional methodology of scoring forced-choice tests is described that results in ipsative data. The psychometric properties of ipsative data are summarized and their implications for psychological assessment are discussed.

Second, new emerging approaches to constructing and scoring forced-choice items under the IRT framework are reviewed. It is shown that none of these approaches may be applied to the existing forced-choice questionnaires.

In the **Method** section, a multidimensional IRT model based on Thurstone's theory of comparative judgments is proposed, which effectively overcomes the limitations of existing approaches and is suitable for most forced-choice questionnaires existing today. It is shown how the Thurstonian IRT model can be embedded in a structural equation modeling framework, and its identification constraints and estimation options are described. The item characteristic function and item information function are given. It is also shown how to compute the test information and estimate the test reliability.

Next, the forced-choice response model is used to investigate necessary conditions for providing resistance to response biases. A class of response biases is identified for which forced-choice format is effective; other classes are briefly discussed and recommendations are given on how to construct the forced-choice questionnaires to maximize resistance against the response biases.

In the **Simulation Studies** section, simulations are performed to investigate the performance of the model across a variety of forced-choice designs. The simulation studies provide important information about the assessment of model fit, and estimation of the model parameters under different conditions. The designs used in the simulation studies are chosen to answer important questions about strengths and limitations of forced-choice questionnaires with dominance items, and their results have important implications on how forced-choice tests should be designed and used in the future.

Two **Empirical Applications** are given to illustrate how the model may be applied in practice. A short Big Five instrument designed specifically for this research is assessed, comparing results derived from the traditional rating scale format and the forced-choice format using the Item Response Theory. Also, a forced-choice personality test used for workplace assessments is scored using the classical methodology (producing ipsative scores) and using the IRT approach. The IRT-derived scores are explored in detail, including their reliability, construct validity and interpretation of individual profiles. It is concluded that when the IRT method is used, the estimated scale scores have normative properties and are similar to the scores derived from the single-stimulus version of the questionnaire.

Finally, the **Discussion** section summarizes the research findings and their implications, makes recommendations for effective forced-choice designs and outlines directions for future research.

Abstract

Multidimensional forced-choice questionnaires can reduce the impact of numerous response biases typically associated with Likert scales. However, if scored with traditional methodology these instruments produce *ipsative* data, which has psychometric problems, such as constrained total test score and negative average scale inter-correlation. Ipsative scores distort scale relationships and reliability estimates, and make interpretation of scores problematic. This research demonstrates how Item Response Theory (IRT) modeling may be applied to overcome these problems. A multidimensional IRT model for forced-choice questionnaires is introduced, which is suitable for use with any forced-choice instrument composed of items fitting the dominance response model, with any number of measured traits, and any block sizes (i.e. pairs, triplets, quads etc.). The proposed model is based on Thurstone's framework for comparative data. Thurstonian IRT models are normal ogive models with structured factor loadings, structured uniquenesses, and structured local dependencies. These models can be straightforwardly estimated using structural equation modeling (SEM) software *Mplus*. Simulation studies show how the latent traits are recovered from the comparative binary data under different conditions. The Thurstonian IRT model is also tested with real participants in both research and occupational assessment settings. It is concluded that when the recommended design guidelines are met, scores estimated from forced-choice questionnaires with the proposed methodology reproduce the latent traits well.

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

Introduction

Personality research relies heavily on self-reported measures (i.e. questionnaires). The typical questionnaire item consist of an *item stem*, which is the stimulus material requiring a response, and a system of recording the *item response* (McDonald, 1999, page 18). Item stems are very often statements describing behavior, state, interest or preference. Respondents have to evaluate the statements in terms of how well they describe them (their typical behavior, their current state, preference etc.).

Single-stimulus response format

The most popular response format used in personality assessment is the so-called *single-stimulus* (SS) format, where respondents are asked to rate each item according to the extent it describes their personality. The distinct feature of the single-stimulus format is that each item is rated separately, therefore *absolute* judgments are made.

Item responses can be given by selecting one of several categories ranging, for example, from “strongly disagree” to “strongly agree”, or from “never” to “always”, or from “very inaccurate” to “very accurate” etc.:

	<i>strongly</i>		<i>neither</i>		<i>strongly</i>
	<i>disagree</i>	<i>disagree</i>	<i>agree nor</i>	<i>agree</i>	<i>agree</i>
<i>I am careful over detail</i>				✓	

This ordered-category response format is commonly referred to as the *Likert* scale. Continuous responses are also possible, and this is accomplished by providing a *graphic rating* scale or *sliding* scale, where the respondent can choose any value between given extremes. One or several anchors can be provided to help the respondent make this choice.

Responses to single-stimulus items are typically coded by assigning a number to each of the ordered response categories, or to the chosen position on the continuous scale. For instance, whole numbers from 1 to 5 are assigned to respective categories of 5-point response scales, such as the “strongly disagree”-“strongly agree” scale presented above.

After responses to items serving as indicators for a particular personality trait have been coded, these codes can be combined in some meaningful way to form a *test score*. Under the Classical Test Theory (CTT) approach, most often item codes are simply added together, forming a sum-score. The Item Response Theory (IRT) provides a more complex way of deriving a test score through finding a trait level that maximizes the likelihood of the given responses to all items (response pattern).

In relation to the single-stimulus personality items, the IRT approach introduced a potential advantage by treating the categorical Likert responses as merely ordinal, where the classical sum-score approach would use whole numbers representing the response categories as they were interval scores. In doing so, the CTT approach makes an assumption of equal distances between the response categories. In reality, however, it is highly unlikely that the difference between “disagree” and “neither agree nor disagree” is exactly the same as the difference between “agree” and “strongly agree” (Baron, 1996). The IRT approach allows for estimating boundaries between the response categories rather than assuming that they are equidistant (e.g. Samejima, 1997).

Response biases affecting single-stimulus items

Fundamentally, scoring of single-stimulus items relies on the assumption that respondents interpret the rating scale (category labels or anchors) in the same way. This assumption, however, is very rarely tested in applications, and research available on the issue suggests that interpretation and meaning of response categories vary from one respondent to another (Friedman & Amoo, 1999). Clearly, individual differences in interpretation of the rating scale can affect the validity of the test score, regardless whether the CTT or IRT approach is used.

In addition to the bias arising from the interpretation of the rating scale itself, there are several other biasing factors associated with the single-stimulus presentation of items. For instance, some respondents may avoid extreme response categories (*central tendency* bias), whereas others may prefer them. Respondents might tend to agree with statements as presented (commonly referred to as *acquiescence* bias or “yea-saying”). The opposite is also possible – the general tendency to disagree with statements. This idiosyncratic use of the rating scale often becomes apparent when both positively and negatively keyed items are used to measure a psychological attribute. The tendency to agree with both positively and negatively keyed items creates a problem with the trait measurement, and requires special modeling (Maydeu-Olivares & Coffman, 2006).

Another problem with the single-stimulus format is getting respondents to differentiate between personal attributes. For example, when asked to rate another person’s workplace behaviors (as in the 360 degree feedback), it is quite common for respondents to give very similar ratings on all behaviors. This bias reflects an over-generalized view of the rated person based on his/her performance on a single important dimension. Depending on whether the rater judges this individual to be generally a good or a poor performer, “halo” and “horn” effects are described (Murphy, Jako & Anhalt, 1993).

Consciously or unconsciously, respondents tend to agree with seemingly desirable items and disagree with undesirable ones, engaging in *socially desirable responding* (for an overview see Zickar & Gibby, 2006). When the stakes are high, as in personality assessment in occupational settings, conscious distortions may occur bringing the responses to a level perceived to be more favored by a potential employer referred to as *faking* (e.g. Griffith & McDaniel, 2006). The single-stimulus format where items are rated separately from each other makes all these types of biases possible. It is hard to estimate the frequency of such distortions, but they will inevitably affect the questionnaire’s validity (Haaland & Christiansen, 1998).

Forced-choice format

Forced-choice (FC) formats were designed to reduce response biases. Instead of evaluating each statement in relation to a rating scale, respondents have to choose between statements according to the extent these statements describe their personality. Therefore, the forced-choice format involves *comparative* judgments.

Forced-choice tests consist of blocks of two or more statements. When there are 2 statements in a block, respondents are simply asked to select one statement that better describes them. For blocks of 3, 4 or more statements, respondents may be asked to rank-order the statements, or to select one statement which is “most like me” and one which is “least like me”. For example:

	<i>Most like me</i>	<i>Least like me</i>
I manage to relax easily		✓
I am careful over detail	✓	
I enjoy working with others		
I set high personal standards		

The example above involves items from different dimensions. It is also possible to have a mixture of statements from the same dimension and different dimensions compared in the same block, and this is considered a special case within a general multidimensional framework.

Direct item comparison overcomes the problems with interpretation of the rating scale altogether. The fact that respondents cannot endorse all items eliminates acquiescence responding (Cheung & Chan, 2002). The forced choice makes it impossible to provide the same response on all items, which will typically result in a greater differentiation of scores within a profile thus reducing the “halo” effects. Bartram (2007) shows that if the forced-choice format is employed in ratings of competencies by line managers, where the “halo” effects are notoriously high, it can increase operational validity of predictor instruments by as much as 50% in comparison to single-stimulus performance ratings.

Forcing to choose between seemingly equally desirable items can reduce socially desirable responding. Recent evidence has shown that forced-choice item formats may be useful in applicant contexts because they are less susceptible to impression management distortion than single-stimulus items (Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen & Hunt, 2001; Vasilopoulos, Cucina, Dyomina, Morewitz & Reilly, 2006). At the group level of analysis, when instructed to “fake good” on single-stimulus measures, respondents can raise their scores by as much as one standard deviation, compared to one third of the standard deviation for forced-choice measures (e.g. Christiansen et al., 1998; Jackson et al., 2000). This is not to say that respondents cannot fake such measures, but that they find it harder, particularly when all traits measured in the questionnaire are perceived to be equally important for the job success. Unsurprisingly, there is evidence that cognitive ability is a positive predictor of successful response distortion of forced-choice instruments (e.g. Vasilopoulos et al., 2006).

For practitioners, the most important measure of questionnaire effectiveness is criterion-related validity. If a reduction in response biases results in more accurate prediction of external criteria, other considerations are less important. Previous research generally found the external validities of single-stimulus and forced-choice formats to be comparable in straight-taking, or honest, conditions (e.g. Bartram et al, 2006; Gordon, 1976; Jackson et al, 2000), and forced-choice formats to be superior in faking conditions, where biases are higher (e.g. Christiansen et al., 1998; Jackson et al, 2000). In sum, comparative judgments made in forced-choice questionnaires can have advantages over absolute judgments made in single-stimulus questionnaires.

Problems of ipsative data

Despite their possible advantages in reducing response biases, forced-choice questionnaires have been controversial because their traditional scoring methodology results in *ipsative* data, very special properties of which can pose threats to construct validity, and score interpretation as well as other substantial psychometric challenges

(e.g. Dunlap & Cornwell, 1994; Johnson, Wood, & Blinkhorn, 1988; Meade, 2004; Tenopyr, 1988).

Data is *ipsative* when the sum of all scores obtained for any individual is a constant. Forced-choice questionnaires represent only one of various ways by which ipsative data is obtained. Because item responses from forced-choice questionnaires are ordinal, the data derived from them is known as *ordinal ipsative* data (Cheung & Chan, 2002). Variations in item keying and scoring produce *fully ipsative* or *partially ipsative* scores, with partially ipsative scores being half-way between the ipsative and the normative scores. To illustrate the controversy behind the use of ipsative scores, the below discussion will concentrate on the most extreme, and therefore the most problematic type, fully ipsative scores.

It is easy to see how this type of data comes about if one considers how the forced-choice format is scored. The usual way of scoring forced-choice items is simply taking their inverted rank-orders (or values derived from them through a linear transformation), and adding them to the respective personality traits. For example, in a block of four statements the most preferred item (ranked first) is given 3 points to add to the trait it measures, and the least preferred (ranked fourth) is given 0 points. Alternatively, if only “most like me”, “least like me” choices are made, the most preferred item adds 2 points to its respective trait, the least preferred item adds 0 points, and the remaining items add 1 point each to their respective traits. Therefore, regardless of the choices made, item scores in the block always add up to the same number, and therefore the total test score (sum of all the blocks) is the same for each individual.

Relative nature of scores

Because the questionnaire allocates the same total number of points for everyone, it is impossible to achieve high (or low) scores on **all** scales in a multi-trait questionnaire. Achieving a high sum-score on one scale will inevitably mean receiving lower scores on other scales. Therefore all ipsative profiles have the same average score regardless if the true scores were overall above or below average.

In idealized conditions, any two individuals with the same ordering of true scores will produce exactly the same ipsative profiles, even when one individual's true scores were on the high end of the distribution, and the other's were at the low end. Therefore, many have argued (e.g. Closs, 1996), ipsative scores make sense for comparison of relative strength of traits within one individual, but they do not provide information on absolute (normative) trait standing, therefore comparisons between individuals are meaningless.

Is this true for all ipsative scores? It is still not well understood by many researchers that the number of measured traits can substantially influence the impact of the constrained overall test score. Baron (1996) shows that with a large number (30 or more) of relatively independent scales, less than one in 100 million of respondents will have all their true trait scores on the same side of the profile, i.e. all high or all low. With a comprehensive range of relatively independent normally distributed traits, most people will have their average profile score around the mean anyway; and in this case, norming of ipsative scores is appropriate and intra-individual comparisons can be performed meaningfully. In carefully designed forced-choice questionnaires with 30 or more measured traits, the ordering of people on each trait largely corresponds to their normative ordering (Baron, 1996; Karpatschhof & Elkjaer, 2000).

With a small number of scales, however, it is more likely that people with the same relative ordering of trait scores can have very different true scores. Therefore, for a small number of scales, the constrained total test score will result in distorted ordering of people (see for example the results of a study with 8-scale forced-choice measure by Meade, 2004). For such tests, ipsativity can have serious implications for the interpretation of scores and selection decisions in applied settings, and remains the most serious limitation in practice.

Distorted construct validity

Constraining the total score to be a constant will lead to zero variance of the total test score. This means that all elements of traits' covariance matrix will sum to zero (Clemans, 1966). It is easy to see that with the covariances summing to zero,

the average off-diagonal covariance is a negative value, and the same is true for correlations. The average off-diagonal correlation in an ipsative measure with d dimensions is:

$$\bar{r} = \frac{-d}{d(d-1)} = \frac{-1}{d-1}. \quad (1)$$

Again, the extent to which the negative average correlation causes a problem depends on the number of traits in the questionnaire, because the average correlation will approach zero as the number of traits increases. If a fully ipsative measure was designed measuring only 2 traits, they would correlate at -1. Therefore one trait's score would completely determine the other trait. With 4 traits, the average off-diagonal correlation is -0.33. In this case, the distortion to the relationships between constructs may be very substantial, particularly when the true trait scores are supposed to be positively correlated. With 30 traits, the average off-diagonal correlation is -0.03, allowing for a wide range of both negative and positive inter-correlations (Bartram, 1996; Baron, 1996). Still, correlations between traits are typically depressed in ipsative instruments as compared with their normative counterparts (Bartram, Brown, Fleck, Inceoglu & Ward, 2006).

Though more interpretable with a large number of measured traits, conventional factor analytic procedures are inappropriate for use with ipsative data. If attempted, the factor analysis extracts bipolar factors, which include contrasting scales from two different substantive factors (Dunlap & Cornwell, 1994; Baron, 1996). In summary, ipsative data clearly compromises the construct validity of forced-choice questionnaires.

Distorted reliability estimates

It is generally agreed among researchers that the ipsative data distorts the internal consistency of instruments, but in which direction and to what degree appears to be highly dependent on specific conditions (see discussion in Baron, 1996). Saville & Willson (1991) found that under the perfect conditions of simulated forced-

choice data, ipsative scores showed reliability of 0.96 for 32 uncorrelated constructs and gradually lower values as the number of scales decreased.

Generally, Cronbach's alpha is an inappropriate statistic for reliability estimation for the forced-choice format because ipsative data violates some important assumptions that the alpha statistic relies on. The first problem is that the most fundamental assumption, independence of error variance, is violated in ipsative data. Indeed, responses to items in the same block are not independent given the latent traits – instead, a response given to one item depends on responses given to all other items in the block (Meade, 2004). The second problem is that the assumption of consistent coding (i.e. high values must have the same meaning across items) is also violated in ipsative data, except in very specific designs. See **Appendix A** for an illustration of this argument.

Some authors have argued that appropriateness of other types of reliability, such as test-retest, is also doubtful with the ipsative data. Classical Test Theory defines the reliability as the proportion of variance due to the true scores. Because ipsative scores violate the CTT assumptions, formulae used to derive various reliability estimates are simply not tenable (Hicks, 1970; Dunlap & Cornwell, 1994; Johnson, Wood, & Blinkhorn, 1988; Tenopyr, 1988; Meade, 2004). In summary, while relying on alpha and other CTT statistics clearly has its limitations when dealing with ipsative data, the question about the real levels of reliability in forced-choice questionnaires remains unanswered.

In spite of these statistical arguments, forced-choice tests have been popular with practitioners over years as with a sufficient number of measured dimensions normative and ipsative versions of the same instrument produced empirically comparable results (Baron, 1996; Karpatschhof & Elkjaer, 2000). For theorists, however, the problems with ipsative data are serious enough to remain concerned by use of the forced-choice format.

Inadequacy of the classical methods of scoring forced-choice items

It is important to understand, however, that the apparent psychometric problems of ipsative data are not inherent to the forced-choice format itself, but originate from the traditional way of scoring.

Despite their obvious presentation differences, single-stimulus and forced-choice items have been scored in pretty much the same way. For positively keyed items, in the single-stimulus format an item adds points to its respective trait score according to the degree it was agreed with; and in the forced-choice format an item adds points to the trait score according to the degree it was preferred to other items. The respondent will receive the highest number of points for the item he/she preferred (ranked first), and the lowest number of points for the item that was least preferred (ranked last). Nevertheless, it is quite clear that forced-choice items are different from single-stimulus items because the item's rank in the block depends not just on the item itself (or more precisely on the trait the item is intended to measure), but also on all other items in the block. When giving the top rank to one item, the respondent does so not because he/she agrees with the statement, but because he/she agrees with that statement *more* than with the other statements in the block. The classical scoring methodology as it stands cannot adequately describe the decision process behind responding to the forced-choice questionnaires. For instance, the fact that forced-choice items are not assessed independently, therefore violating one of the basic assumptions of test theory, independence of error variance, is totally ignored in current ipsative scoring. This scoring assumes that preferring one item to another is the same as to agreeing with one and disagreeing with the other. Meade (2004) shows how responding to one forced-choice item is dependent on *all* traits represented in a block, and argues that the decision process that respondents use to select items "is unknown and inherently alters the psychometric properties at the item level".

The psychological process of responding to forced-choice items is certainly different from single-stimulus items, and understanding this process is the key to making sense of comparative data. With potentially advantageous features of the

forced-choice format and problematic properties of the data it is associated with, an alternative method of scoring is needed for forced-choice questionnaires.

IRT approaches to scoring forced-choice items

Two new approaches have been proposed recently to describe the process of making choices between questionnaire items, and apply this modeling to creating new forced-choice measures. Both approaches make use of a special type of items, so-called *ideal-point* (or *unfolding*) items. These terms were coined by Coombs (1964) based on the original work of Thurstone, who described a process of responding to attitude items (Thurstone, 1929). Thurstone argued that because such items often represent moderate or mid-scale standing for a particular attitude, the probability of agreeing with them is the highest for individuals with this exact level of the attitude, and reduces for persons with extremely strong attitudes in both directions – either toward the top or the bottom of the scale. The likelihood of agreeing with an item in unfolding models peaks at a certain point on the latent trait continuum (the item *location*), and decreases as the person’s trait score departs further from that location. When plotting the likelihood of agreeing with such an item against the latent trait, the item have ideal-point (bell-shaped) response function, as opposed to *dominance* (s-shaped) response function typical for traditional personality items.

Originally suggested for attitude items, the unfolding models have been tested with behavioral items typical for personality questionnaires. Stark, Chernyshenko, Drasgow and Williams (2006) found that a very small proportion of existing personality items they examined had a response function that complied with the dominance model for most of the latent trait continuum, but showed a small downward trend for very extreme positive scores. They have argued that these items would be better served by the ideal-point model. As far as existing personality items are concerned, the occurrence of the ideal-point response functions is very rare; however, it is possible to write items specifically to fit the ideal-point response model. Such items would be designed to represent a moderate standing on the latent trait, for instance: “My attention to detail is about average”. Clearly, the likelihood of

agreeing with this item would be high for respondents with an average score on Conscientiousness, and be lower for respondents with very high or very low scores.

McCloy, Heggstad and Reeve (2005) rely on ideal-point items and relate the likelihood of preferring one item to another to the difference between distances from the item locations to the person's true scores. In this model, a respondent is more likely to prefer the item located closer to one's own true score on the respective trait to the item located further from one's own true score. Based on this theory, McCloy and colleagues suggest a way of creating forced-choice tests by repeatedly presenting blocks of items from different dimensions with locations that vary across the trait continuum. The items' IRT location parameters are established through single-stimulus presentation. By combining items with different locations, it is theoretically possible to find the most likely trait level for an individual. This method is proposed as a way of creating new forced-choice tests but does not offer a solution for scoring most tests existing today because of its restrictions on item properties. It is essential for the method to work that all items have ideal-point response functions and varied locations, and grouped in a very specific way, which is not the case with the existing forced-choice questionnaires.

Stark, Chernyshenko and Drasgow (2005) approximate the probability of preferring one item to another by the joint probability of accepting one statement and rejecting the other. These probabilities of acceptance and rejection are IRT-based and established through single-stimulus trialing. Stark and colleagues show how to create forced-choice tests by assembling pairs of items from different dimensions (using a small proportion of one-dimensional pairings) based on their single-stimulus IRT parameters. They also use the generalized unfolding model as a basis for IRT calibration of single-stimulus items. However, the model is limited to pairs of items and does not deal with blocks of three or more statements, which are popular in existing forced-choice questionnaires. One-dimensional pairings required by the method to set a scale for each trait also limit its applicability to existing forced-choice tests. Furthermore, the methodology involves a numerical solution of systems of equations requiring substantial expertise in using specialist subroutines and setting starting values, which is likely to be a barrier for most researchers.

The need to model forced-choice dominance items

Most existing personality items and questionnaires were created under the classical test theory assumptions where items by design correlate strongly with their respective scales. Any mid-level or double-barreled items that might be acceptable in the unfolding models are typically removed from the classical personality scales. “Good” personality items, in the classical sense, represent strong statements typical of someone with a very high (or very low for the negatively keyed items) true score on the trait. From the ideal-point process logic, such items would have such high/low locations that no respondents would exist beyond that location to disagree with them.

Such items, whether used in the single-stimulus or the forced-choice format, represent a *dominance* response process. In the dominance model, just like in ability testing, the probability of agreeing with an item is monotonically increasing as the score on the underlying trait increases. The dominance model assumes that items are written in a way that they serve as either positive or negative indicators of the latent trait, having ever growing (or decreasing) utility for respondents with higher trait scores. Examples of positively and negatively keyed dominance items are: “I keep my paperwork in order” and “I struggle to organize my paperwork“, respectively. For any two respondents, the utility for the first item will be higher for the individual whose score on Conscientiousness is higher, and this will be reversed for the second item.

Dominance items are by far more prevalent in existing personality questionnaires, either using the single-stimulus or the forced-choice formats. An examination of the popular 16PF questionnaire shows that the vast majority of items fit the dominance model (Stark, Chernyshenko, Drasgow & Williams, 2006). Popular forced-choice questionnaires, such as the Occupational Personality Questionnaire (OPQ; Bartram et al., 2006), the Customer Contact Styles Questionnaire (CCSQ; SHL, 1997), the Survey of Interpersonal Values (SIV; Gordon, 1976) consist of strong statements representing a very high level of the latent trait and strongly correlating with their respective scales. Given the absence of an adequate model for these

popular personality questionnaires, and any future questionnaires utilizing the most widespread type of items, this research is aimed at introducing a model suitable for the multidimensional forced-choice format with dominance items.

Thurstone's framework for comparative judgment

Trying to address the problem of ipsative data, Chan and Bentler (1998) proposed a method for analyzing the covariance structure of ordinal ipsative data, which uses information from comparisons between the object ranked first and all other objects. Such modeling would apply to a single ranking block. Maydeu-Olivares (1999) proposed a method of analyzing mean and covariance structure of comparative data that uses **all** paired comparisons underlying the choices within the ranking block, therefore using more information from the observed ranking patterns. Crucially, he linked preference choices to the Thurstone's theory of latent utilities.

Thurstone (1927, 1931) proposed a theory that attributes the outcome of a comparative judgment to the relative utility value of the objects under comparison. As such, it is based on three assumptions. First, each choice alternative elicits a latent continuous *utility* judgment as a result of a discrimination process. Utility is a concept typically describing the *value* of an object for the respondent; for personality items the utility would describe the extent of how closely the statement resembles the respondent's typical behavior or preference. Second, the respondent chooses the alternative with the largest utility value at the time of comparison. Third, utility values are distributed normally in the population of respondents. By assuming that a factor model underlies the utilities of the choice alternatives, this approach offers a suitable model for forced-choice dominance items.

In Thurstonian factor models for rankings (Maydeu-Olivares, 1999; Maydeu-Olivares and Böckenholt, 2005), pairwise comparisons between all items in a ranking task (that essentially amount to "preferred"- "not preferred" outcomes) are the observed binary variables; and the latent utilities are modeled as the first order factors that determine the binary outcomes. The second order factors are (in this case) the personality traits assumed to cause utilities of items. Therefore, a

Thurstonian factor model is essentially a second-order factor model for dichotomous data.

Although Thurstonian models provide an attractive representation of responding to forced-choice items, they have not been extensively used since no efficient estimation methods were available. Maydeu-Olivares (1999) proposed an estimation approach that embeds Thurstonian models within a more familiar structural equation modeling (SEM) framework and allows complex models for comparative judgments to be estimated and tested efficiently (see also Maydeu-Olivares & Böckenholt, 2005). A Thurstonian factor model works successfully with small forced-choice tasks to estimate parameters of the latent utilities at the sample level. However, the model estimates many latent variables and therefore there are limits on the amount of forced-choice items that can be estimated with today's hardware and software. In the author's experience, at most five scales measured by 15-20 blocks of statements of four items can currently be estimated using the second-order Thurstonian factor model for ranking. Most forced-choice personality tests in use aim at measuring multiple personality traits and they often consist of hundreds of items. Such tests are simply too large to be estimated with current computing capabilities using this approach. Most importantly, however, the Thurstonian second-order factor model does not allow estimating the person scores on the latent traits due to zero error variance of the outcome comparison variables (this will become clear later when the details of this model are described in the Method section).

The approach proposed in this dissertation differs from the second-order Thurstonian factor models in one key aspect. Unlike in marketing applications, where Thurstonian models are most often used, in personality assessment the latent utilities of items are not of interest. Consequently, the model bypasses the latent utilities, directly linking choices made by an individual to the latent traits measured by the test. The resulting model is therefore an IRT model. Crucially, this formulation changes the way the error of the binary outcomes of comparisons between items is modeled, and allows the latent trait estimation (person scores) that are the main focus of the personality assessment. In addition, the IRT formulation contains fewer

latent variables and is suitable for modeling the responses given to large multi-scale forced-choice questionnaires such as the OPQ32 measuring 32 personality traits (Bartram et al., 2006), the CCSQ measuring 16 traits (SHL, 1997), and others.

Method

Binary coding of forced-choice response data

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

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

	Ranking
Item A	–
Item B	–
Item C	–
Item D	–

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

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

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

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

$$y_l = \begin{cases} 1 & \text{if item } i \text{ is preferred over item } k \\ 0 & \text{if item } k \text{ is preferred over item } i \end{cases} \quad (2)$$

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

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

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

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

One consequence of dealing with blocks of statements (each of which is a small ranking task) is that responses made within one block are always *transitive*. For

example, if the respondent rank-orders A above B, and B above C, it automatically follows that A is ranked before C and therefore the outcome of {A,C} pair can be deducted from pairs {A,B} and {A,C}. It then follows that only $n!$ different binary patterns may be observed for a block of n items.

Thurstonian factor models for forced-choice items

Response model for ranking

Thurstone (1927) proposed the following model describing comparative choices, such as ones made in forced-choice blocks. 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. He argued that in a comparative task, 1) each item elicits a utility as a result of a *discriminal process*; 2) respondents choose the item with the largest utility value at the moment of comparison; and 3) the utility is an unobserved (continuous) variable and is normally distributed in the population of respondents.

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

$$y_i = \begin{cases} 1 & \text{if } t_i \geq t_k \\ 0 & \text{if } t_i < t_k \end{cases}, \quad (3)$$

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

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

$$y_l^* = t_i - t_k \quad (4)$$

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

$$y_l = \begin{cases} 1 & \text{if } y_l^* \geq 0 \\ 0 & \text{if } y_l^* < 0 \end{cases} \quad (5)$$

Importantly, the difference of utilities determines the preference response, i.e. there is no error term in Equation (4). This is because in ranking tasks responses are transitive (Maydeu-Olivares & Bockenholt, 2005).

It is convenient to present the response model in a 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 $\tilde{n} = n(n-1)/2$. Then the set of \tilde{n} equations (4) can be written as

$$\mathbf{y}^* = \mathbf{A} \mathbf{t}, \quad (6)$$

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

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

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

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

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & | & 0 & 0 & 0 & | & 0 & 0 & 0 \\ 1 & 0 & -1 & | & 0 & 0 & 0 & | & 0 & 0 & 0 \\ 0 & 1 & -1 & | & 0 & 0 & 0 & | & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & | & 1 & -1 & 0 & | & 0 & 0 & 0 \\ 0 & 0 & 0 & | & 1 & 0 & -1 & | & 0 & 0 & 0 \\ 0 & 0 & 0 & | & 0 & 1 & -1 & | & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & | & 0 & 0 & 0 & | & 1 & -1 & 0 \\ 0 & 0 & 0 & | & 0 & 0 & 0 & | & 1 & 0 & -1 \\ 0 & 0 & 0 & | & 0 & 0 & 0 & | & 0 & 1 & -1 \end{pmatrix}.$$

Thurstone's model assumes that the latent utilities \mathbf{t} are normally distributed in the population of respondents. Thus, we can write $\mathbf{t} \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ denote the mean vector and covariance matrix of the latent utilities \mathbf{t} .

Items as indicators of latent traits

The next important step is to assume that the latent utilities \mathbf{t} are indicators of a set of d common factors (latent traits):

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

This is because questionnaire items are designed to measure some psychological constructs (personality traits, motivation factors, attitudes etc.). Here it will be assumed that every item measures one trait only. In factor analytic terms, this means that the relationship between the items and the common factors is an independent cluster solution.

In Equation (7), $\boldsymbol{\mu}_i$ contains m means of the latent utilities \mathbf{t} , $\boldsymbol{\Lambda}$ is an $m \times d$ matrix of factor loadings, $\boldsymbol{\eta}$ is an d -dimensional vector of common factors (latent traits in IRT terminology), and $\boldsymbol{\varepsilon}$ is an m -dimensional vector of unique factors (residual variances of utilities). This model assumes that the traits are normally distributed, have mean zero, unit variance and are freely correlated (their covariance matrix is $\boldsymbol{\Phi}$). The uniqueness terms are normally distributed, have mean zero and are uncorrelated, so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal.

In this standard factor model, the utility of an item *monotonically* depends on the latent trait, that is, it increases when the latent trait increases (for positively keyed items – those with positive factor loadings), or decreases when the latent trait increases (for negatively keyed items – those with negative factor loadings). This model describes *dominance* response process – it assumes that items are written in a way that they serve as either positive or negative indicators of the latent trait, having ever growing (or decreasing) utility for respondents with higher trait scores.

To illustrate how binary outcomes, their underlying utilities and traits are modeled, in **Figure 1** a Thurstonian factor model is sketched for a very short forced-choice questionnaire. An *Mplus* syntax for this model is given in **Appendix B**. The questionnaire measures $d = 3$ correlated traits; each trait is measured by 3 items. The nine questionnaire items ($m = 9$) are presented in triplets (blocks of $n = 3$ items) so that there are no two items within a block measuring the same trait. There are $p = 3$ such blocks in this simple example. Trait 1 is measured by items 1, 4, and 7; trait 2 is measured by items 2, 5, and 8; and trait 3 is measured by items 3, 6, and 9. Respondents are asked to rank-order the items within each block. The resulting rankings are transformed into 3 binary outcomes per block (9 outcomes in total), which are modeled as differences of underlying utilities using Equation (6). Because each binary outcome is the result of comparing two items, it depends on two latent utilities. Utilities, in turn, are functions of the 3 personality traits. The 9 binary outcomes are measured without error (because responses to ranking blocks are transitive). However, the 9 utilities have disturbance terms accounting for the items' unique variance not explained by the latent traits.

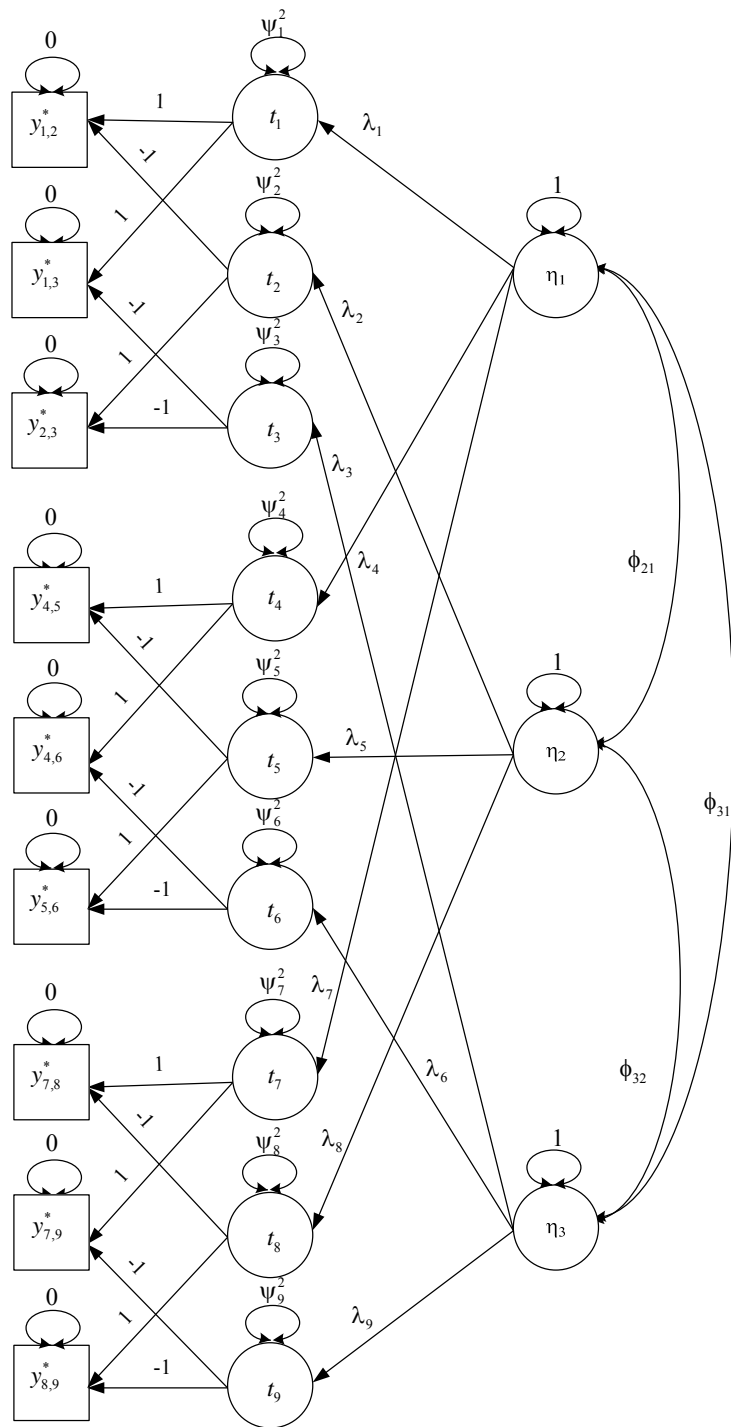


Figure 1. Thurstonian second-order factor model for a questionnaire with 3 traits and 3 blocks of 3 items

Thurstonian Models for forced-choice items as IRT models

In this section it is shown how the Thurstonian factor model, which is a second-order factor model for binary data with some special features, can be equivalently expressed as a first-order model, again, with some special features. The item characteristic and information functions for the model are provided, and item parameter estimation, latent trait estimation, and reliability estimation are discussed.

Reparameterized model (first-order Thurstonian IRT factor model)

There are several reasons for reparameterizing the Thurstonian factor model for forced-choice presented above as a first-order model. First, in psychometric testing applications the first order factors (the latent utilities) are not of interest. Rather, interest lies in estimating the second order factors (the latent traits). Second, and most importantly, since the residual error variances of the latent response variables \mathbf{y}^* are zero in the second-order factor model with latent utilities, latent trait estimates cannot be computed (see Maydeu-Olivares, 1999; Maydeu-Olivares & Brown, 2010). When the model is reparameterized as a first order model, the residual error variances of the latent response variables are no longer zero, enabling latent trait estimation. In addition, the reparameterization provides some valuable insights into the characteristics of the model, and enables formulation of such important descriptors of any IRT model as item characteristic functions and information functions.

The reparameterization involves writing the second-order factor model obtained from Equations (6) and (7)

$$\mathbf{y}^* = \mathbf{A} (\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon}, \quad (8)$$

as the first-order model

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \check{\boldsymbol{\Lambda}}\boldsymbol{\eta} + \check{\boldsymbol{\varepsilon}}, \quad (9)$$

in which the latent traits influence the outcomes of comparisons directly, according to their difference weighted by the factor loadings of the items involved. The reparameterized model (9) involves

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

$$\check{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda}, \quad (10)$$

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

$$\check{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}', \quad (11)$$

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

$$\boldsymbol{\gamma} = -\mathbf{A}\boldsymbol{\mu}_t. \quad (12)$$

That is, restriction (12) is not imposed on $\boldsymbol{\gamma}$. This is because in IRT applications the means $\boldsymbol{\mu}_t$ of the latent utilities are not of interest. Therefore an unrestricted vector of thresholds $\boldsymbol{\gamma}$ will be estimated leading to a considerably less constrained model.

To illustrate the structure imposed by the model on the matrices $\check{\mathbf{\Lambda}}$ and $\check{\boldsymbol{\Psi}}^2$, consider the previous example of a very short forced-choice questionnaire measuring $d = 3$ latent traits with $p = 3$ blocks of $n = 3$ items. For this example,

$$\check{\mathbf{\Lambda}} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \hline \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \hline \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \end{pmatrix}, \quad (13)$$

whereas

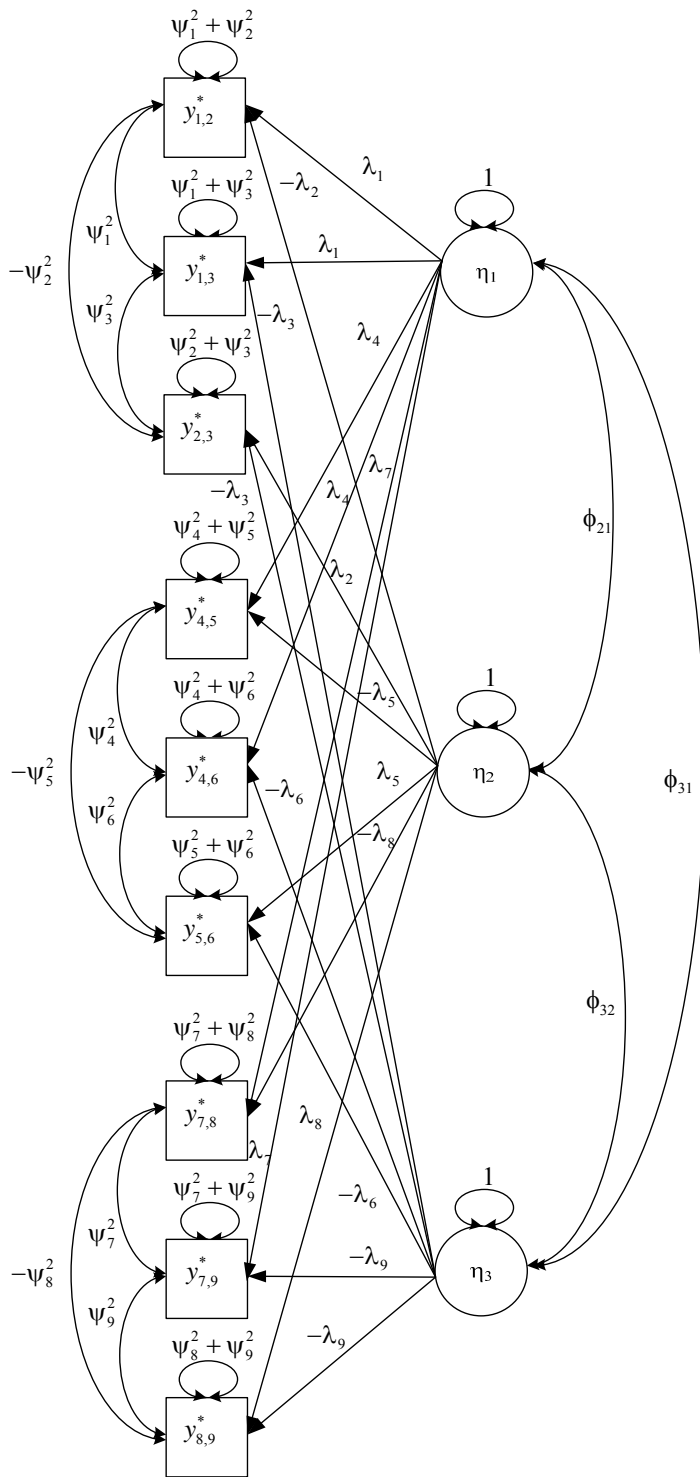


Figure 2: Thurstonian IRT model for a questionnaire with 3 traits and 3 blocks of 3 items

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

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

Identification of Thurstonian IRT models for forced-choice questionnaires

The reparameterized model is algebraically equivalent to the original Thurstonian factor model, thus yielding the same number of parameters, and requiring exactly the same identification constraints. For a single ranking task, Maydeu-Olivares and Brown (2010) suggested the following constraints to identify the model: (a) fixing all factor loadings involving (arbitrarily) the last item to 0 ($\lambda_{ni} = 0$ for all $i = 1, \dots, d$); and (b) fixing the unique variance of the last item to 1, $\psi_n^2 = 1$. These identification constraints are needed to set the scale origin for factor loadings and for the uniquenesses because of the comparative nature of the data. To set the scale for the latent traits, the variances of the latent traits are simply set equal to one.

In the case of a multidimensional model involving several blocks of items each measuring a single trait (i.e. forced-choice questionnaire model), the identification constraints are simpler than in the case of a single block. The model is identified simply by imposing a constraint among the uniquenesses within each block.

Arbitrarily, the uniqueness of the last item in each block can be fixed to 1. For example, to identify the Thurstonian IRT model depicted in **Figure 2**, the uniqueness of the last item in each block is set to 1: $\psi_3^2 = 1$, $\psi_6^2 = 1$, and $\psi_9^2 = 1$. Also, the variances of the latent traits are set to 1. Mplus syntax for testing this model and computing individual trait scores is provided in **Appendix C**.

This general identification rule is valid in all but two special cases: a) when $n = 2$ and $d > 2$ (i.e., items presented in pairs measuring more than 2 traits), and b) when $d = n = 2$ (only two traits are measured using pairs of items). In Case a), no item uniqueness ψ_i^2 can be identified. They can be set equal to 0.5, so that $\tilde{\psi}_i^2 = \psi_i^2 + \psi_k^2 = 1$. Case a) is discussed in more detail in **Appendix D**. Regarding Case b), all item uniquenesses need to be fixed as in the case above. In addition, each binary outcome will depend on both traits involved, the factor loading matrix contains no zero elements, and the model is essentially an exploratory factor model. To avoid the indeterminacy problem in this case (see McDonald, 1999, page 179), it is sufficient to fix the 2 factor loadings of the first pair. For a model with 3 or more traits, no such constraints are needed because there are sufficient numbers of zero elements in each column and row of the factor loading matrix.

Item characteristic function

It follows from (9) and the normality of the latent response variables \mathbf{y}^* that the probability of preferring item i over item k is

$$\Pr(y_i = 1 | \boldsymbol{\eta}) = \Phi\left(\frac{-\gamma_i + \tilde{\boldsymbol{\lambda}}_i' \boldsymbol{\eta}}{\sqrt{\tilde{\psi}_i^2}}\right), \quad (15)$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function evaluated at x , γ_i is the threshold for binary outcome y_i , $\tilde{\boldsymbol{\lambda}}_i'$ is the $1 \times d$ vector of factor loadings, and $\tilde{\psi}_i^2$ is the uniqueness for binary outcome y_i . Because it is assumed that each item only measures one trait (an independent-cluster solution), each binary outcome only depends on two traits. As a result, the item characteristic function for

the binary outcome variable y_l , which is the result of pairwise comparison between items i and k measuring traits η_a and η_b , is

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi \left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}} \right). \quad (16)$$

Here, λ_i and λ_k are factor loadings for traits η_a and η_b , respectively; and $\psi_i^2 + \psi_k^2$ is the variance of the error term for the latent response variable y_l^* . Equation (16) describes the item characteristic function using a threshold/loading parameterization. This is simply a standard two-dimensional normal ogive IRT model for binary data except that (a) factor loadings are structured so that every binary outcome y_l involving the same item will share the same factor loading, (b) uniquenesses of binary outcomes are structured so that they equal the sum of uniquenesses of the 2 items involved, and (c) the item characteristic functions are not independent (local independence conditional on the latent traits does not hold). Rather, there are patterned covariances among the residual variances of the latent response variables.

Now, letting

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_k^2}}, \quad \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_k^2}}, \quad \beta_k = \frac{\lambda_k}{\sqrt{\psi_i^2 + \psi_k^2}}, \quad (17)$$

the item characteristic function (16) can be written in an **intercept / slope** form as

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b). \quad (18)$$

Because the probability of binary outcome l depends on 2 latent traits, this equation describes the Item Characteristic Surface (ICS), an example of which is presented in **Figure 3**. When a respondent is forced to choose between 2 items, his/her standing on the two underlying traits will influence the difference of utilities of the choice alternatives, and therefore, the outcome of the comparison. With an increase in the true score on the first trait and decrease on the second trait, the probability of preferring the first item to the second item is non-decreasing and is influenced by: a) the respondent's scores on the two underlying traits η_a and η_b , b)

the discriminations (slopes) β_i and β_k of the two items on their underlying traits, and c) a threshold α_l governing the combination of the latent traits when the statements' utilities are equal.

The Thurstonian IRT model for forced-choice questionnaires also applies to the case where some (or all) binary outcomes arise from comparing items measuring the same trait. Indeed, a test developer might want to include items measuring the same trait in the same block. In this case equations (16) and (18) are rewritten to include only one latent trait η :

$$\Pr(y_l = 1 | \eta) = \Phi(\alpha_l + (\beta_i - \beta_k)\eta). \quad (19)$$

The one-dimensional case will not be specifically referred to in the present research, however, it is important to point out that this specific case is easily accommodated in the more general model described above. Special features of the one-dimensional case are described in Maydeu-Olivares and Brown (2010). One most obvious observation arising from the equation (19) is that 2 items with similar discrimination parameters (slopes) will provide virtually no information for the estimation of the latent trait, when they are used in a forced-choice block. Therefore, if one wants to present items measuring the same trait in a forced-choice block, items with very different slopes should be used such as positively and negatively keyed items (Maydeu-Olivares & Brown, 2010).

Estimation of Thurstonian IRT models for forced-choice questionnaires

IRT models are most often estimated using full information maximum likelihood (FIML). For models describing forced-choice questionnaires such estimation is not feasible due to the very large number of dimensions involved. However, the Thurstonian IRT models can be straightforwardly estimated using limited information methods. First, the sample thresholds and tetrachoric correlations are estimated. Then, the model parameters are estimated from the first stage estimates by unweighted least squares (ULS) or diagonally weighted least squares (DWLS).

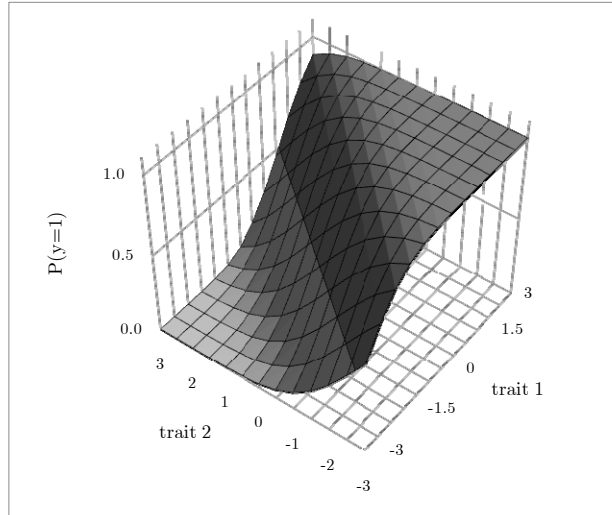


Figure 3: *Item Characteristic Surface (ICS) for the binary outcome $\{i5, i6\}$ for the simulation with 2 uncorrelated traits*

Note: Item parameters for the binary outcome $\{i5, i6\}$ in the intercept/slope form: $\alpha = 0.72$; $\beta_1 = 0.90$; $\beta_2 = 0.72$ (see **Table 1**).

In practice, differences between using ULS or DWLS in the second stage of the estimation procedure are negligible (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009). All models in the present research are tested in *Mplus* using either the DWLS estimator with mean corrected Satorra-Bentler goodness-of-fit tests (Muthén, 1993), or ULS estimator for larger models. Note that the DWLS estimation procedure is denoted as WLSM estimation in *Mplus*.

When the number of items per block is larger than 2, a correction to degrees of freedom is needed when testing model fit. This is because for a ranking block there are $r = n(n-1)(n-2)/6$ redundancies among the thresholds and tetrachoric correlations estimated from the binary outcome variables (Maydeu-Olivares, 1999). For instance, there is $r = 1$ redundancy in every block of 3 items, and there are $r = 4$ redundancies in every block of 4 items. With p ranking blocks in the questionnaire, the number of redundancies is $p \times r$. Thus, when $n > 2$, one needs to subtract $p \times r$ from the degrees of freedom given by the modeling program to obtain the correct p -

value for the test of exact fit. Goodness-of-fit indices involving degrees of freedom in their formula, such as the RMSEA, also need to be recomputed using the correct number of degrees of freedom. When $n = 2$, no degrees of freedom adjustment is needed, the p -value and RMSEA printed by the program are correct.

Latent trait estimation

Once the IRT model parameters have been estimated, scores on the latent traits for individuals can be estimated using their pattern of binary outcome responses. There are 3 popular procedures for latent trait estimation: maximum likelihood (ML), expected a posteriori (EAP), and maximum a posteriori (MAP) estimation (Embretson & Reise, 2000). The focus here will be on the MAP estimator, which maximizes the mode of the posterior distribution of the latent traits, as it is the method implemented in *Mplus*. The posterior distribution is obtained by multiplying the joint likelihood of the binary outcome responses by the density of the population distribution, which is standard multivariate normal in this model. The MAP estimator exists for all response patterns, is more efficient than the ML estimator when a small number of items is involved (and in personality questionnaires the number of items per trait is generally small), but is known to produce estimates biased towards the population mean (see Embretson & Reise, 2000, page 174).

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

Information functions and reliability estimation

In Item Response Theory, unlike in classical scoring, the precision of measurement depends on the latent traits and therefore is not the same for all respondents. The precision of measurement is provided by the test information function $\mathcal{I}(\boldsymbol{\eta})$, which is computed from item information functions $\mathcal{I}_l(\boldsymbol{\eta})$. Recall that in the forced-choice questionnaires, observed variables (“items”) are binary outcomes of pairwise comparisons between the questionnaire items.

The item information function is computed in a manner similar to its one-dimensional IRT counterpart, except that since each binary outcome depends on two dimensions, the direction of the information must be also considered (Reckase, 2009; Ackerman, 2005). The definition of item information in the multidimensional case is generalized to accommodate the change in slope with direction taken from a point in the latent trait space:

$$\mathcal{I}_l^\alpha(\boldsymbol{\eta}) = \frac{[\nabla_\alpha P_l(\boldsymbol{\eta})]^2}{P_l(\boldsymbol{\eta})[1 - P_l(\boldsymbol{\eta})]}, \quad (20)$$

where $\boldsymbol{\alpha}$ is a vector of angles to all d axes that defines the direction from a point $\boldsymbol{\eta}$. In this expression, ∇_α is the gradient (directional derivative) in direction $\boldsymbol{\alpha}$, which is given by (Reckase, 2009):

$$\nabla_\alpha P_l(\boldsymbol{\eta}) = \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_1} \cos \alpha_1 + \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_2} \cos \alpha_2 + \dots + \frac{\partial P_l(\boldsymbol{\eta})}{\partial \eta_d} \cos \alpha_d. \quad (21)$$

Because each binary outcome depends on 2 latent traits, in the above expression directional derivatives for all but the 2 relevant dimensions will be 0. For each binary outcome, the contributions to the information about two underlying traits it is intended to measure, η_a and η_b , are of interest. Therefore for each binary outcome two main directions of information are considered: one coinciding with the axis η_a , and another one coinciding with the axis η_b . When computing the information in direction η_a , the angle to η_a is 0° (and therefore $\cos(\alpha_a) = 1$), and the

angle to η_b is determined by the correlation between η_a and η_b so that $\cos(\alpha_b) = \text{corr}(\eta_a, \eta_b)$ (see Bock, 1975).

Using the intercept/slope parameterization from Equation (18), the directional derivatives by η_a and η_b are simply

$$\frac{\partial P_l(\eta_a, \eta_b)}{\partial \eta_a} = \beta_i \phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b) \quad \text{and} \quad \frac{\partial P_l(\eta_a, \eta_b)}{\partial \eta_b} = -\beta_k \phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b), \quad (22)$$

where $\phi(z)$ denotes a standard normal density function evaluated at z (McDonald, 1999, p. 284). It follows from (20) - (22) that the information provided by one binary outcome about traits η_a and η_b are, respectively:

$$\mathcal{I}_l^a(\eta_a, \eta_b) = \frac{[\beta_i - \beta_k \text{corr}(\eta_a, \eta_b)]^2 [\phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b)]^2}{P_l(\eta_a, \eta_b) [1 - P_l(\eta_a, \eta_b)]}, \quad (23)$$

$$\mathcal{I}_l^b(\eta_a, \eta_b) = \frac{[-\beta_k + \beta_i \text{corr}(\eta_a, \eta_b)]^2 [\phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b)]^2}{P_l(\eta_a, \eta_b) [1 - P_l(\eta_a, \eta_b)]}. \quad (24)$$

Equations (23) and (24) describe the Item Information Surfaces (IIS), examples of which are presented in **Figure 4**. It can be seen from these equations that for binary outcomes involving uncorrelated traits, only the derivative in the direction of the trait itself contributes to the information. However, for binary outcomes involving correlated traits, derivatives in directions of both traits involved will contribute. For positively keyed items, binary outcomes involving **positively** correlated traits will provide *less information* than if the traits were orthogonal (holding the item parameters equal). And, for positively keyed items, binary outcomes involving **negatively** correlated traits will provide *more information* than if the traits were orthogonal. These properties, as will be seen later, have important implications for test design.

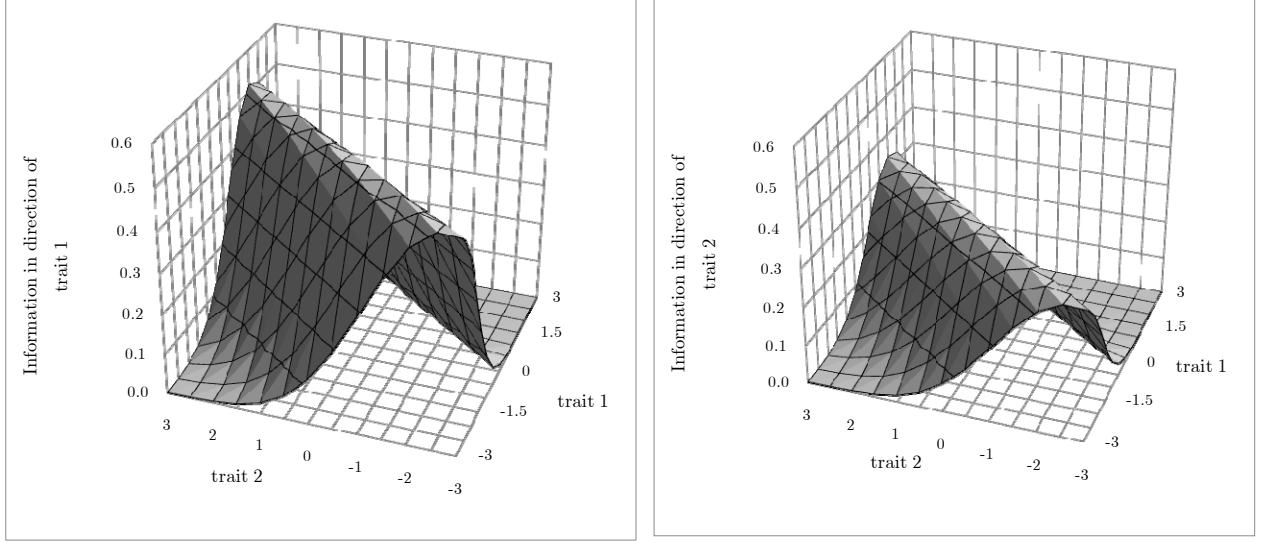


Figure 4: *Item Information Surfaces (IIS) in directions of Trait 1 and Trait 2 for the binary outcome $\{i5, i6\}$ for the simulation with 2 uncorrelated traits*

Note: Item parameters for the binary outcome $\{i5, i6\}$ in the intercept/slope form: $\alpha = 0.72$; $\beta_1 = 0.90$; $\beta_2 = 0.72$ (see **Table 1**).

For the one-dimensional case, i.e. when items measuring the same trait η are compared in the same block, equations (23) and (24) reduce to

$$\mathcal{I}_i(\eta) = \frac{[\beta_i - \beta_k]^2 [\phi(\alpha_l + (\beta_i - \beta_k)\eta)]^2}{P_l(\eta)[1 - P_l(\eta)]}, \quad (25)$$

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

$$\mathcal{I}^a(\boldsymbol{\eta}) = \sum_l \mathcal{I}_l^a(\boldsymbol{\eta}). \quad (26)$$

However, structured dependencies exist between the error terms within blocks of 3 or more items. It has been shown that the test information is overestimated only slightly when these dependencies are ignored (Maydeu-Olivares & Brown, 2010). In this research the simplifying assumption that the binary outcomes are locally

independent is used, and the extent to which the information estimates are sufficiently accurate in applications is investigated.

All the above applies to IRT scores estimated by the maximum likelihood method (ML). When Bayes MAP estimation of the latent traits is used, the posterior test information $\mathcal{I}_P(\boldsymbol{\eta})$ is given by the sum of the ML test information and information given by the prior distribution (see Du Toit, 2003), which is multivariate standard normal:

$$\mathcal{I}_P^a(\boldsymbol{\eta}) = \mathcal{I}^a(\boldsymbol{\eta}) - \frac{\partial^2 \ln(\phi(\boldsymbol{\eta}))}{\partial^2 \eta_a} = \mathcal{I}^a(\boldsymbol{\eta}) + \boldsymbol{\varpi}_a^a, \quad (27)$$

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

$$SE(\hat{\eta}_a) = \frac{1}{\sqrt{\mathcal{I}_P^a(\hat{\boldsymbol{\eta}})}}. \quad (28)$$

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

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

$$\bar{\sigma}_{error}^2(\hat{\boldsymbol{\eta}}) = \frac{1}{N} \sum_{j=1}^N \frac{1}{\mathcal{I}_P^a(\hat{\boldsymbol{\eta}}_j)}. \quad (29)$$

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

$$\rho = \frac{\sigma_P^2 - \bar{\sigma}_{error}^2}{\sigma_P^2}, \quad (30)$$

where σ_P^2 is the observed MAP score variance, and $\bar{\sigma}_{error}^2$ is the mean of the squared errors of the MAP estimates of the trait scores for individuals in the sample. Finally, the correlation between the true latent trait and the estimated latent trait is estimated as follows:

$$\text{corr}(\eta_a, \hat{\eta}_a) = \sqrt{\rho}. \quad (31)$$

This sample-based approach overcomes the main difficulty with the multidimensional information, which is to summarize item information data for a multi-trait forced-choice questionnaire. Since items from the focus trait are compared with items from many other traits, the information in the direction of the target trait for every binary outcome is conditional on different traits. To summarize such contributions for all values of all traits involved in the questionnaire, it is necessary to consider a multidimensional grid, with the number of dimensions corresponding to the number of measured scales. In practice, however, multivariate grids for a large number of dimensions consist of millions of points and are computationally infeasible.

Random sampling of the points on the multidimensional grid can be used instead – and this is exactly what the sample-based approach described above achieves. First, information in two relevant directions is computed for each binary outcome at two given values η_a and η_b , providing two single values (scalars). Second, to obtain the total information for one dimension, single values from all pairs involving this dimension are summed.

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

Response biases and forced-choice format

Forced-choice formats were introduced to reduce response biases. Having established the suitable model for describing the decision process behind responding to forced-choice items, it becomes possible to examine how various biases will influence the responses. The main problem facing such analysis, however, is the lack of clear modeling definition of biases. With the exception of acquiescence or “yea-saying” bias (*uniform* response bias, defined in Cheung & Chan, 2002), no established response models exist for the majority of practically identified and described biases. The task of classification and modeling of response biases is beyond the scope of this research, however, the question about resistance of the forced-choice format to biases can be approached from another perspective. Rather than starting from formulating models for different biases and examining whether the FC format is resistant to them, one can start from the internal workings of the comparative judgments and work back to establish exactly what type of transformations it will be robust against.

The main property of the FC format is its comparative nature. Outcomes of the pairwise comparative judgments are based on the difference of the items’ latent utilities, given by Equation (4), and the binary outcomes are determined by the sign

of this difference as described in (5). It is clear that because the binary outcome only depends on the sign of the difference between the utilities, it is invariant to any transformation of the utilities as long as their difference remains of the same sign:

$$y_l^{*B} y_l^* = (t_i^B - t_k^B)(t_i - t_k) \geq 0 \quad (32)$$

In the expressions above y_l^B denotes the “biased” latent difference, and t_i^B and t_k^B denote “biased” utilities.

It is easy to see why the acquiescence response bias is eliminated by the use of forced-choice items. This is because of the uniform additive nature of this bias, i.e. for a given individual j , “biased” item utilities can be written as $\mathbf{t}_j^B = \mathbf{t}_j + \mathbf{1}\delta_j$ where δ_j is the additive bias uniform across all items within this individual, and $\mathbf{1}$ is a unit vector of dimensionality corresponding to the number of items (Chan, 2003). When the differences of utilities are considered, it can be seen that the additive term simply disappears:

$$y_l^{*B} = t_i^B - t_k^B = (t_i + \delta) - (t_k + \delta) = t_i - t_k = y_l^* \quad (33)$$

Similarly, any uniform multiplicative bias described as $\mathbf{t}_j^B = c_j \mathbf{t}_j$ (Chan, 2003) is also eliminated by the use of the forced-choice format when $c_j > 0$ because

$$y_l^{*B} = t_i^B - t_k^B = ct_i - ct_k = c(t_i - t_k) = cy_l^*, \quad (34)$$

and both biased and unbiased latent differences have the same sign. Extending this logic further, any combination of the uniform additive and multiplicative bias will be eliminated by the comparative nature of forced-choice items:

$$y_l^{*B} = t_i^B - t_k^B = (ct_i + \delta) - (ct_k + \delta) = c(t_i - t_k) = cy_l^* \quad (35)$$

However, in order for such an additive-multiplicative bias to be eliminated, it is not necessary that it is uniform across **all** items. It is sufficient that the additive-multiplicative term **remains constant within each block**, but it can vary across blocks. In fact, any combination of additive and multiplicative transformations to the

original unbiased utilities will preserve the sign of the binary outcome as long as these transformations are **uniform across all items in the block**.

Now, while it is easy to describe and model additive biases (e.g. with the random intercept model; Maydeu-Olivares & Coffman, 2006), multiplicative response biases would suggest a *random slope* – i.e. different item discriminations across respondents. Although conventional common factor models do not allow modeling random slopes, a situation where it might be necessary can be easily imagined. For instance, *extreme responding* distorts the responses in such a way that positive item utilities become more positive and negative ones more negative. Conversely, *central tendency* responding makes the responses less pronounced. This type of bias is multiplicative – there is a certain “magnifying” effect that either stretches the ratings (extreme responding) or shrinks them (central tendency). Such a response style is likely to apply to all items in a questionnaire, i.e. be uniform.

Another common type of bias – *socially-desirable responding* – affects items in a non-uniform fashion. Items that are seen by the respondent as very desirable are likely to get more affected than those that are seen more neutral. Inevitably, the desirable items will change their discrimination on the trait they intend to measure, and this change will be non-uniform across items. Matching items in terms of their desirability inside each ranking block, thus making the multiplicative and additive terms more or less uniform might reduce the impact of socially-desirable responding.

Simulation studies

In this section, random data from several Thurstonian IRT models is generated to assess the model fit, to investigate how the true model parameters are recovered under different conditions, and how well the true trait scores are recovered. Simulation studies are essential to provide benchmarks to which similar real-world applications can be compared. Since the true scores are never known in real applications, it is impossible to judge how well our model can recover them – and to what extent this recovery is affected by the questionnaire design, response bias, or by violation of the assumptions we make in the model.

Another important objective of the simulation studies is to assess the impact of the simplifying assumption of local independence on the latent trait estimation and the test information estimation when blocks of 3 or more items are used.

First, an extremely simplified questionnaire with 2 traits measured by item pairs is considered. This low-dimensionality example provides an opportunity to look at the graphical illustrations of ICS and test information functions. Most importantly, it provides a benchmark for the precision of the latent trait estimation when no local dependencies exist (and no simplifying assumptions for either latent trait estimation or the information estimation need to be made).

Then, a more realistic questionnaire model is considered measuring 5 traits (probably the smallest number of traits one would be interested in measuring in practice). For this model, the block size is manipulated, i.e. blocks of 2, 3 and 4 items are considered.

These examples use the pure multidimensional forced-choice format, i.e. items measuring the same trait never appear in the same block. However, mixed designs involving items measuring the same trait as well as different traits can be easily incorporated in the model as shown in Maydeu-Olivares and Brown (2010).

Simulation study 1. A forced-choice questionnaire measuring 2 traits

The purpose of this simulation study is to give an empirical illustration of the multidimensional Thurstonian IRT model with the smallest number of traits. Let us consider a hypothetical forced-choice questionnaire measuring 2 traits. For example, one can think of measuring global personality factors, such as ‘Dynamism’ and ‘Social Propriety’ also referred to as ‘Getting Ahead’ and ‘Getting Along’ (Hogan, 1983). Alternatively, any narrow traits can also be measured in this fashion.

Despite being somewhat limited, this example is useful for illustrating the model properties before moving on to complex multidimensional models. Particularly, test characteristic and test information functions can be presented graphically, which will not be possible with higher dimensionality. An important feature of this simple example is that blocks of $n = 2$ items (item-pairs) are used, consequently no correlated errors exist and no simplifying assumptions are made.

To set realistic factor loadings and uniqueness parameters for continuous utilities in this model, data from several personality scales were evaluated. It was decided that absolute values of factor loadings should be drawn from a uniform random distribution between the minimum value 0.65 and the maximum 0.95. Obviously, the sign of factor loadings might vary according to whether one chooses to use positively or negatively keyed items to measure the traits. According to the magnitude of the factor loadings, uniqueness terms were set to make the total variance for each utility equal 1 (unobserved utilities are assumed to be standard normal for simplicity); therefore they varied between 0.10 and 0.58. Item intercepts were set to vary between -0.8 and 0.8.

The most common design in existing forced-choice questionnaires is to use positively keyed items only. However, these existing questionnaires typically measure a larger number of traits. Indeed, if 2 traits were measured by positive items only in a traditionally scored forced-choice questionnaire, the ipsative constraint would mean that the score on one trait would be completely determined by the score on the second trait (because the 2 scores should sum to a constant), and therefore correlation between them would be -1. Clearly, 2 traits cannot be meaningfully

measured with positively keyed items only when the traditional ipsative scoring is used. An alternative forced-choice design is sometimes used, whereby both positive and negative items are combined in the same block (as in Jackson, Wroblewski & Ashton, 2000; Heggstad, Morrison, Reeve, & McCloy, 2006). In this case the scoring for the negative item is reversed, giving a possibility for a different number of points to be assigned in some blocks and partially releasing the ipsative constraint. Here the objective is to evaluate how the Thurstonian IRT approach deals with both designs.

First, **only positively keyed items** are used to create 2 questionnaires – one short with 12 items per trait, and one long with 24 items per trait. The item parameters were set as described above, with factor loadings being all positive. Next, both **positively and negatively keyed items** are used to create 2 questionnaires (with 12 and 24 items per trait as above). Item parameters were kept exactly the same as in the questionnaire with positive items, except for factor loadings, which were reversed for some items.

Apart from this difference in the sign of the factor loadings, the same basic design was considered in all simulations. The short questionnaire consists of $p = 12$ item-pairs (thus using 12 items from each trait, $m = 24$ items in total), and the long questionnaire consists of $p = 24$ item pairs (using 24 items per trait, or $m = 48$ items in total). The short questionnaire forms the first half of the long questionnaire. **Table 1** gives the true item parameters for the short questionnaire with positively and negatively keyed items. Items in each pair belong to different traits, and the item order alternates to avoid carry-over effect when responding. Out of each pair, respondents have to select one item that describes them more accurately.

Table 1: True item parameters for the short questionnaire (12 item-pairs) using both positively and negatively keyed items; simulation with 2 traits

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

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

To investigate if trait relationship bears any influence on the model properties, the traits were set to be uncorrelated, positively correlated at 0.5, and negatively correlated at -0.5. For each of the 3 levels of the trait correlations (0, 0.5 and -0.5), 1000 random samples of $N=1000$ cases were generated with 2 normally distributed traits ($\sigma^2 = 1$), and independent uniqueness terms for each questionnaire item with variances as specified above. From these generated variables, $\tilde{n} = 1$ continuous difference of utilities were produced for each block-pair using $\mathbf{y}_j^* = \boldsymbol{\mu}_{y^*} + \mathbf{A}(\boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j)$, where $j = 1, \dots, N$ are individuals in the sample, and dichotomized \mathbf{y}_j^* using (5). Therefore expected binary outcomes for each individual were obtained according to Thurstone's theory of latent utilities.

Having obtained $p \times \tilde{n} = 12$ or 24 binary responses for the short and the long test respectively, the corresponding Thurstonian IRT model was applied to estimate item parameters and trait correlations. As explained in the section on the model identification, models with 2 latent traits essentially amount to exploratory models, and additional constraints are needed in order to identify them. In this model, error variances of all binary outcomes were fixed to their expected values ($\psi_i^2 + \psi_k^2$ for the binary outcome y_j), and also the factor loadings of the first binary outcome to their expected values λ_1 and λ_2 . The model was specified and tested in *Mplus*. The degrees of freedom do not need to be adjusted in this case as there are no redundancies in blocks of 2 items.

Design 1. Questionnaire with positively keyed items only

The number of successful computations varied between 815 and 980 for different questionnaire lengths and trait correlations. Replications that did not yield successful estimations were empirically unidentified. **Table 2** summarizes the goodness-of-fit for different conditions as measured by *Mplus* chi-square statistic (Muthén, 1998-2007). It can be seen that goodness-of-fit tests in this case are off - the model is retained more often than it should.

Table 3 summarizes parameter estimates for different conditions. The trait correlations are estimated relatively accurately and more so for the long questionnaire. Item parameter estimates for the short questionnaire are positively biased (by between 4% and 18% for different conditions), but for the long questionnaire they are accurate. However, very large standard errors were present for several item parameters, indicative of unstable estimates.

Table 2: *Goodness of fit in the simulation studies with 2 traits*

Keyed direction of items	Number of items per trait	Correlation between traits	Number of converged replications	Degrees of freedom	Chi-square mean	Chi-square SD	Chi-square rejection rates
+	12	0	868	43	36.88	7.73	.01 .05 .10 .20
+	12	0.5	853	43	38.26	8.16	.000 .004 .012 .034 .076
+	12	-0.5	815	43	36.39	7.91	.000 .005 .016 .043
+	24	0	974	229	217.84	20.36	.001 .012 .033 .093
+	24	0.5	980	229	221.93	20.55	.006 .021 .049 .115
+	24	-0.5	893	229	215.18	20.44	.000 .011 .027 .071
+/-	12	0	1000	43	43.37	9.51	.016 .049 .105 .206
+/-	12	0.5	999	43	42.66	9.32	.010 .041 .093 .185
+/-	12	-0.5	1000	43	44.00	10.19	.018 .079 .146 .260
+/-	24	0	1000	229	232.09	24.57	.026 .078 .150 .261
+/-	24	0.5	1000	229	233.09	25.25	.032 .074 .137 .261
+/-	24	-0.5	1000	229	232.51	23.01	.027 .071 .134 .240

Table 3: *Parameter estimates in the simulation studies with 2 traits*

Keyed direction of items	Number of items per trait	Correlation between traits	Loadings' average relative bias			Thresholds' average relative bias		
			Estimated mean (SD)	SE mean	Est. SE	Est. SE	Est. SE	
+	12	0	-1.04 (.250)	.232	.126 (.123)	12.17 (5.33)	.161 (.074)	13.33 (6.52)
+	12	0.5	.437 (.184)	.151	.039 (.035)	6.75 (3.90)	.078 (.040)	7.67 (4.62)
+	12	-0.5	-.560 (.252)	.306	.171 (.160)	10.99 (4.46)	.184 (.090)	12.73 (5.89)
+	24	0	-.097 (.218)	.194	-.005 (.044)	.913 (2.038)	.037 (.041)	1.34 (2.48)
+	24	0.5	.438 (.149)	.115	-.035 (.023)	-.094 (.530)	.009 (.037)	.145 (.662)
+	24	-0.5	-.580 (.226)	.258	.016 (.054)	1.60 (1.55)	.028 (.069)	.916 (2.22)
+/-	12	0	.039 (.191)	.188	.023 (.028)	-.029 (.044)	.010 (.007)	-.002 (.024)
+/-	12	0.5	.460 (.184)	.166	.028 (.015)	-.020 (.030)	.007 (.016)	.007 (.029)
+/-	12	-0.5	-.479 (.106)	.106	.011 (.012)	.000 (.025)	.007 (.008)	.001 (.028)
+/-	24	0	.031 (.171)	.166	.018 (.018)	-.013 (.018)	.013 (.023)	-.008 (.023)
+/-	24	0.5	.480 (.109)	.103	.017 (.011)	-.022 (.023)	.009 (.052)	-.002 (.034)
+/-	24	-0.5	-.477 (.099)	.097	.005 (.012)	-.002 (.023)	.000 (.056)	.001 (.021)

Next let us turn to the first sample replication to evaluate the trait recovery. The trait scores computed by *Mplus* using MAP estimation yielded disappointingly low levels of true score recovery overall, which did not depend on the questionnaire length, but strongly depended on the correlations between the traits. For the short questionnaire, the correlations between the MAP estimated and the true scores were for trait 1 and trait 2 respectively: 0.344 / 0.349 in the model with positively correlated traits; 0.601 / 0.629 in the model with uncorrelated traits; and at 0.793 / 0.766 in the model with negatively correlated traits. Similarly, for the long questionnaire these correlations were 0.344 / 0.306 in the model with positively correlated traits; 0.614 / 0.618 in the model with uncorrelated traits; and 0.793 / 0.791 in the model with negatively correlated traits.

Clearly, the recovery of the latent traits is totally unacceptable for the positively correlated traits, is poor for the uncorrelated traits, and only approaches satisfactory levels for the negatively correlated traits. When the correlations for the best model (with negatively correlated traits) are converted into estimates of reliability using Equation (31), they yield $\rho(f_1) = 0.629$ and $\rho(f_2) = 0.625$ for the long questionnaire, and $\rho(f_1) = 0.629$ and $\rho(f_2) = 0.587$ for the short questionnaire. Neither would be considered acceptable levels of reliability for a personality questionnaire.

To understand these results, let us now turn to **Figure 4**, which depicts the ICS for a pair $\{i_5, i_6\}$ from this example. It can be seen that the change in the surface's slope depends on the direction in the trait space. The slope is high in the direction taken from an angle of about 45° towards the positive end of the first trait (η_1) and the negative end of the second trait ($-\eta_2$). It means that such an item-pair will contribute a sizeable amount of information to the trait difference score ($\eta_1 - \eta_2$). Therefore pairs where one has to choose between two positively keyed items will highlight *differences* in the 2 latent traits. At the same time, the ICS appears essentially flat in the direction taken from an angle of about 45° towards the positive ends of both traits. The same item-pair would provide virtually no information on the sum score ($\eta_1 + \eta_2$) of the two latent traits. Therefore, the information provided

by this pair is about the *relative* position of the 2 underlying trait scores, not their *absolute* locations. This is why the recovery of the true score is so poor.

This problem is aggravated even further when the measured traits are positively related to each other. This is because the information provided by the binary outcome is lower in this situation, as can be seen from equations (23) and (24). On the other hand, for the negatively correlated traits, the information provided by the binary outcome is higher than for the uncorrelated traits.

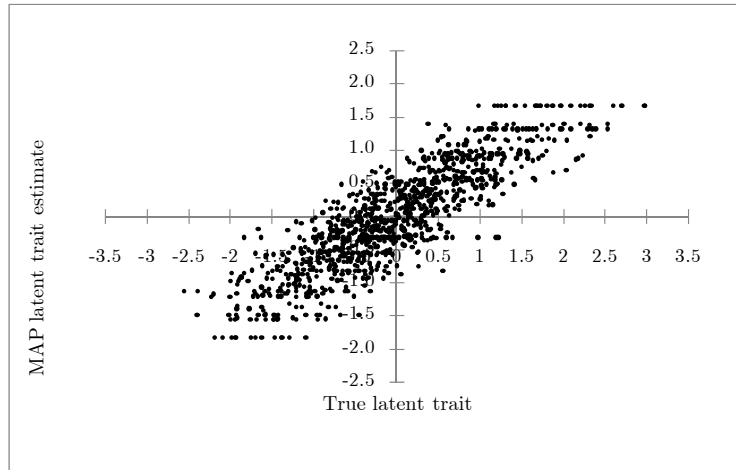
It has to be concluded that when measuring 2 traits, the forced-choice design with items keyed in the same direction is not recommended. When traits are negatively correlated, the recovery of scores is better but still falls short of acceptable levels.

Design 2. Questionnaire with both positively and negatively keyed items

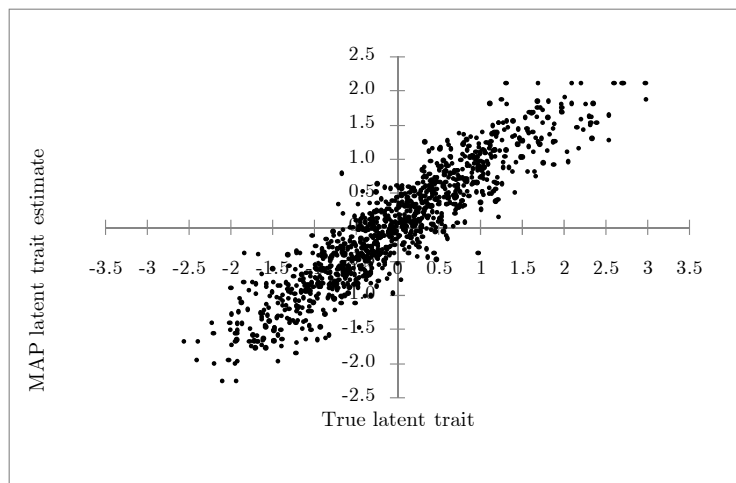
Having learnt from the previous design that positively keyed items alone provide information on the difference score between the latent traits, but not on the sum score, let us turn to a design where positively and negatively keyed items from the 2 traits are combined together in blocks. Item-pairs of positively keyed items should provide information on the difference between the latent traits, and item-pairs of items keyed in opposite directions should add information about the traits' sum, thus locating both traits. In this design, forced-choice blocks are built by combining: 1) positive items from both traits; 2) positive item from trait 1 and negative from trait 2; and 3) negative item from trait 1 and positive from trait 2. There are equal numbers of blocks of each type. Exactly the same item parameters are used here as in the previous design except that the sign of the factor loadings for some items is reversed. True item parameters for the short questionnaire (and the same items form the first half of the long questionnaire) are given in **Table 1**.

The estimation in *Mplus* proceeded successfully for all 1000 replications in all conditions, apart from one condition where 999 replications were successful. Goodness-of-fit statistics are given in **Table 2**, and statistics on parameter estimates are given in **Table 3**. It can be seen that chi-square rejection rates in this case are

very close to the expected values. Trait correlations were estimated accurately, with no more than 5% bias. The item parameters were estimated very accurately with negligible bias.



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



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

Figure 5: *Scatterplot of MAP estimated trait scores vs. true latent trait scores for the simulation with 2 uncorrelated traits*

Next let us consider the first sample replication with uncorrelated traits to evaluate how well the true scores were recovered. **Figure 5** shows plots of the true trait scores versus the estimated MAP scores for the short and the long questionnaires. All in all score recovery is good: true scores and MAP scores correlate at 0.872 for trait 1 and 0.860 for trait 2 in the short questionnaire; and in the long questionnaire they correlate at 0.918 for trait 1 and 0.918 for trait 2. These correlations can be converted into estimates of reliability using Equation (31), yielding figures $\rho(f_1) = 0.760$ and $\rho(f_2) = 0.740$ for the short questionnaire, and $\rho(f_1) = 0.842$ and $\rho(f_2) = 0.843$ for the long questionnaire. Therefore 12 item-pairs provide reliability levels that are considered just acceptable for a personality questionnaire, and 24 item-pairs provide very good reliability indeed. Correlations with the true scores, and therefore the reliability estimates are similar for the models with positively and negatively correlated traits (all reliability estimates for this example are reported in **Table 4**). Clearly, combining positive and negative items in blocks resulted in much more accurate estimation of true scores across all conditions.

Now let us consider the test information and standard errors. **Figure 6** provides the MAP test information functions computed in direction of trait 1 for the short and long questionnaires with uncorrelated traits. It can be seen that with twice as many items in the long questionnaire, approximately 2 times more information is obtained in the middle of the latent distribution.

The MAP test information was evaluated at the estimated individual scores, computing the average error variance across the sample. The empirical reliabilities for the short questionnaire (see **Table 4**) are $\rho(f_1) = 0.691$ and $\rho(f_2) = 0.674$, and for the long questionnaire they are $\rho(f_1) = 0.842$ and $\rho(f_2) = 0.840$. Comparing these figures to the reliabilities obtained by correlating the estimated and the true scores, it can be seen that the information method underestimates the actual reliability for the short questionnaire by about 0.07 or 10%. This is most likely due to the variance of the observed score being low, which is typical when the MAP estimator is used with a small number of items (it is biased towards the population mean).

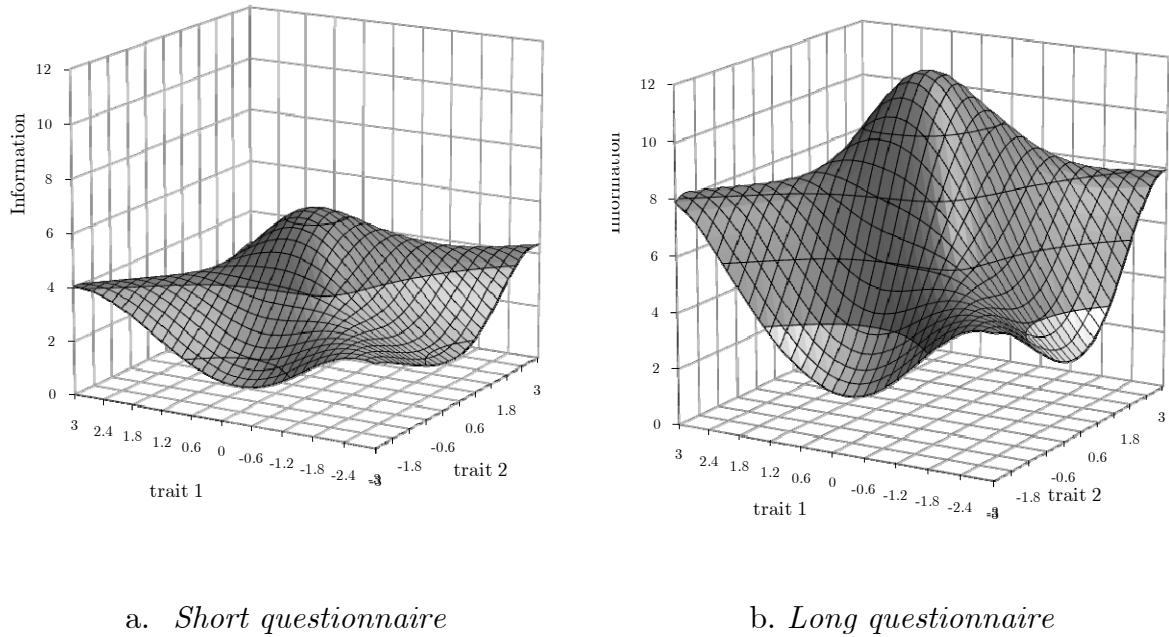


Figure 6: MAP test information function for the simulation with 2 uncorrelated traits

Note: Information is computed in direction of Trait 1. Darker shading on the graphs signifies the information values over 4, corresponding to the test reliability over 0.75.

For the longer questionnaire, however, the information method proves very accurate. Turning to simulations with positively and negatively correlated traits, results were very similar to the uncorrelated case (see **Table 4**). The accuracy of MAP estimation for latent traits remained at approximately the same levels. Also, the information method yielded underestimated reliability levels for the short questionnaire; however, it was very precise when the number of binary outcomes per trait was larger.

To conclude, in a forced-choice application with 2 traits latent trait estimation can be precise when both positively and negatively keyed items are combined in the same blocks. Relationships between the traits do not affect the effectiveness of the IRT score estimates in this case. It is effective to combine positive and negative items (making positive-positive item pairs, positive-negative item pairs, and negative-positive item pairs) in order to locate the absolute trait scores in all applications

with 2 dimensions. Combining negative items with negative provides the same information as positive items. Finally, as few as 12 item-pairs can be used to obtain reliability levels of around 0.75. If higher precision of measurement is required, more item pairs should be used. Also, on the basis of this example, the sample-based empirical estimates of reliability seem to give fairly accurate results, more so for longer questionnaires.

Table 4: *Test reliabilities in the simulation with 2 traits; questionnaire with positively and negatively keyed items (first replication)*

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

Notes: Actual test reliability is computed as squared correlation between MAP estimated and true scores; information-based reliability is calculated using MAP test information evaluated at the MAP estimated sample scores. For the questionnaires with positive items, the information method is not recommended (see text).

Simulation study 2. A forced-choice questionnaire measuring 5 traits

The purpose of this simulation is to investigate how well the latent trait scores can be recovered in a forced-choice questionnaire measuring 5 traits. The 5 traits measured here will mimic the Big Five personality factors (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). Correlations between the traits were set to values reported for the NEO PI-R (Costa & McCrae, 1992), which are given in **Table 5**. Each trait is measured with 12 items. Parameters for the item utilities (factor loadings, intercepts and uniqueness) in this model follow the same rules as in the previous example with 2 traits. Again, it is attempted to create a questionnaire with **positively keyed items** first, and then with **both positively and negatively keyed items**.

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

According to the model, 1000 random samples of $N=1000$ cases were generated with 5 traits ($\sigma^2 = 1$) correlated as per **Table 5**, and 60 independent uniqueness terms with variances $\psi_i^2 = 1 - \lambda_i^2$. For every design, continuous differences of utilities were produced for each pairwise comparison in each block using $\mathbf{y}_j^* = \boldsymbol{\mu}_{y^*} + \mathbf{A}(\boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j)$, where $j = 1, \dots, N$ are individuals in the sample, and dichotomized \mathbf{y}_j^* using (5). Having obtained binary responses according to Thurstone's theory of latent utilities, *Mplus* was used to test the corresponding Thurstonian IRT model using the DWLS estimation, and compute MAP estimates of the latent traits.

Table 5: *True trait correlations in the simulation studies with 5 traits*

	N	E	O	A	C
Neuroticism (N)	1*	-0.21	0	-0.25	-0.53
Extraversion (E)		1*	0.40	0	0.27
Openness (O)			1*	0	0
Agreeableness (A)				1*	0.24
Conscientiousness (C)					1*

Note: (*) Trait variances are set to 1 for identification.

Design 1. Blocks of 2 items (pairs)

In the first questionnaire design with blocks of $n = 2$ items, $d = 5$ traits are measured with $m = 60$ items (12 items per trait), and the number of blocks is $p = 30$. Each block produces $\tilde{n} = 1$ binary outcome, therefore the total number of binary outcomes is $p \times \tilde{n} = 30$, and each trait is measured by 12 binary outcomes.

To identify this model, pairs' uniquenesses have to be fixed as for the previous example with 2 traits. However, no constraints on factor loadings are required (see the section on identification of Thurstonian IRT models above). The degrees of freedom do not need to be adjusted as there are no redundancies in blocks of 2 items.

The model estimation proceeded successfully for 954 replications when positive items only were used; and for all 1000 replications when both positive and negative items were combined in blocks. Both versions yielded correct empirical rejection rates for chi-squares (see **Table 6** for goodness-of-fit statistics). Item parameters and trait correlations were estimated accurately (see **Table 7** for parameter estimation statistics). The correlations between traits were slightly negatively biased for the model with all positive items, but for the model with both positive and negative items they were recovered to a very high degree of accuracy. In the questionnaire with all positive items the standard errors of correlations were negatively biased by about 30%, the SE of item loadings were negatively biased by about 20% and the SE

of item thresholds were negatively biased by about 10%. In the questionnaire with positive and negative items standard errors were accurate.

Let us consider the first replication to evaluate how well the true scores were recovered in this example. The true scores and MAP scores correlated on average at 0.822 for the questionnaire with all positive items, and at 0.889 for the questionnaire combining both positive and negative items. When these correlations are converted into estimates of reliability using Equation (31), they yield reliabilities just below 0.7 for the positively keyed items design, and at around 0.79 for the positive/negative item design (all reliabilities are reported in **Table 8**).

The test information functions and the average squared errors were also computed for this replication, and turned into the reliability estimates using (30). Comparing these information-based estimates presented in **Table 8** to the actual reliabilities, it can be seen that for both designs the information method slightly underestimates the reliability, on average by about 5%. This is likely due to the relatively small number of binary outcomes per trait (12), leading to the substantial “compression” of the MAP score, and consequently small observed score variance.

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

Table 6: *Goodness of fit in the simulation studies with 5 traits*

Keyed Block size	direction of items	Number of successful computations	Degrees of freedom	Chi-square mean	Chi-square SD	Empirical rejection rates of chi-square test
2	+	954	365	360.38	33.60	.023
	+/-	1000	365	367.44	34.23	.047
3	+	1000	1640*	1669.01	106.36	.154
	+/-	1000	1640*	1677.73	101.00	.155
4	+	1000	3830*	3920.31	192.62	.266
	+/-	1000	3830*	3914.01	196.52	.268

Notes: Number of items per trait is 12 for all designs.

(*) Degrees of freedom are adjusted for the number of redundancies in the model.

Table 7: Average relative bias for parameter estimates and standard errors in the simulation studies with 5 traits

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

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

Table 8: *Test reliabilities in the simulation studies with 5 traits (first replication)*

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

Notes: Actual test reliability is computed as squared correlation between MAP estimated and true scores; *empirical* reliability is calculated using posterior test information evaluated at the MAP estimated sample scores. (**) For blocks of 3 or 4 items, the information method makes a simplifying assumption of local independence.

Design 2. Blocks of 3 items (triplets)

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

To identify this model, one item's uniqueness per block has to be fixed, but no constraints on factor loadings are required. The degrees of freedom in this case need

to be adjusted because there are 20 redundancies in 20 blocks of 3 items (1 redundancy per block).

The estimation proceeded successfully for both versions (with positively keyed items only and with positively and negatively keyed items) for all replications. Empirical rejection rates for the chi-square, however, are much higher than the nominal rates (see **Table 6**); therefore models of this kind are rejected more often than they should based on the test of exact fit. All item parameters were estimated very accurately, with negligible bias (see **Table 7**). The correlations between traits were also recovered accurately, particularly for the questionnaire with positive/negative items. In the questionnaire with all positive items the standard errors of correlations were negatively biased by about 15%, and the SE of item loadings were negatively biased by about 10%.

Let us consider the first replication to evaluate how well the true scores were recovered for both versions of the questionnaire. It has been shown that MAP estimation using the simplifying assumption of local independence provides very accurate results even when this assumption is violated in blocks of 3 or more items (Maydeu-Olivares & Brown, 2010). The true scores and MAP scores correlated on average at 0.863 for the questionnaire with all positive items, and at 0.929 for the questionnaire combining positive and negative items. Converting these correlations into estimates of reliability using Equation (31), reliabilities of about 0.75 are obtained for the positive items design, and of about 0.86 for the positive/negative items design (see **Table 8**). The test information functions and the average squared errors were turned into the reliability estimates using (30), yielding figures of about 0.87 for the positively keyed items design, and of about 0.88 for the positive/negative items design. We can see that the information method is very accurate in estimating the reliabilities for the questionnaire with positive and negative items, despite ignoring the correlated uniquenesses in this triplet forced-choice design, and making a simplifying assumption of local independence. The very minor over-estimation of about 2% is totally acceptable in practice.

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

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

Design 3. Blocks of 4 items (quads)

The next questionnaire consists of $p = 15$ blocks of $n = 4$ items (quads), utilizing the same 60 items as in the previous examples. The items are arranged into quads so that all 5 permutations of 4 out of 5 traits are equally represented. This makes each subset of 4 traits appear exactly 3 times in the questionnaire. Each block produces $\tilde{n} = 6$ binary outcomes, therefore the total number of binary outcomes in this model is $p \times \tilde{n} = 90$, and each trait is measured by 36 binary outcomes. Note that here it is assumed that full rankings are performed in each block (not the “most”-“least” incomplete ranking), and therefore all binary outcomes are known.

To identify this model, one item's uniqueness per block has to be fixed, but no constraints on factor loadings are required. The degrees of freedom in this case need to be adjusted because there are 60 redundancies in 15 blocks of 4 items (4 redundancies per block). The estimation proceeded successfully for both positive and positive/negative questionnaire versions for all 1000 replications. Similarly to the model with triplets, empirical rejection rates for the chi-square test are much higher

than they should be (see **Table 6**). The goodness-of-fit test will reject this type of model much more often than it should. All item parameters and the trait correlations were estimated very accurately (see **Table 7**). In the questionnaire with all positive items the standard errors of correlations were negatively biased by about 14%, and the SE of item loadings were negatively biased by about 10%.

Again, consider the first replication to evaluate how well the true scores were recovered. While the trait recovery has not improved compared to the triplet design (the true scores and MAP scores correlated on average at 0.867) for the questionnaire with all positive items, it has improved even further to the impressive average of 0.942 for the questionnaire combining both positive and negative items. Converting these correlations into estimates of reliability using Equation (31), reliabilities of about 0.75 are obtained for the positively keyed items design, and of about 0.89 for the positive/negative item design (see **Table 8**). The information method yielded the reliability figures of about 0.91 for the positively keyed items design, and of about 0.92 for the positive/negative item design. Again, the method assumes that every item improves the prediction and thus yields accurate figures for the positive and positive/negative items designs. However, the design with positive items only has reached its limit in its ability to locate the trait scores (it has not improved the prediction from the triplet design despite increased numbers of binary outcomes). The design with positive and negative items, on the other hand, was able to improve the prediction even further. Therefore, the information method is still accurate in estimating the reliabilities for the questionnaire with positive and negative items, despite ignoring correlated errors in this forced-choice design with quads, and making a simplifying assumption of local independence. The information method makes a very minor over-estimation of around 3%, which would be considered acceptable in practice.

To conclude, in a forced-choice application with 5 traits, the design with 15 quads provides sufficient measurement precision, particularly for the questionnaire combining both positive and negative items within blocks.

Empirical applications

In this chapter, real data applications are considered. In the first application, responses were collected to a short Big Five questionnaire created by the author solely for the purpose of this research. The second application is a popular workplace questionnaire CCSQ (Customer Contact Styles Questionnaire; SHL, 1997) measuring 16 traits relevant to customer service and sales jobs. For both applications, the relationships are provided between the trait scores estimated from the forced-choice items using the Thurstonian IRT model and the trait scores estimated from the single-stimulus versions of the same items.

Application 1. A Big Five questionnaire constructed from IPIP items

Instrument

One of the designs described in the simulation studies with 5 traits was used as a template to create a short forced-choice questionnaire. Items were drawn from the International Personality Item Pool (IPIP), more specifically from its subset of 100 items measuring the Big Five factor markers (Goldberg, 1992). Note that constructs measured by this questionnaire are not the same as in NEO-PIR, and therefore correlations between the five traits are expected to be different from those reported in **Table 5**. Sixty items were selected so that 12 items would measure each of the 5 marker traits. The triplet design was chosen from the simulation study above, with 8 positively and 4 negatively keyed items per trait. The questionnaire items are listed in **Appendix F**. These items were also translated into Spanish, thus creating another language version of the questionnaire.

Each block of 3 in the questionnaire is designed to be presented in 2 formats. First, participants have to rate the 3 items using a 5-point rating scale “very accurate”- “moderately accurate”-“sometimes accurate and sometimes inaccurate”- “moderately inaccurate”-“very inaccurate”. This scale is a modification of the rating scale suggested by Goldberg (1992), where the original middle category “neither inaccurate nor accurate” was replaced with a category more explicitly referring to

the “in-between” possibility, rather than to “neither-nor” meaning often leading to “don’t know” interpretation. Research on the use of intermediate categories has shown that a clear reference to the in-between position has advantages to the scale’s properties (Hernandez, Espejo & González-Romá, 2006).

This single-stimulus presentation was immediately followed by the forced-choice presentation, where the participants were asked to select one item “most like me”, and one “least like me” out of the same block of 3 items. Two formats were used in order to compare trait scores as estimated from the single-stimulus and forced-choice items.

Sample 1 – English version

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

Sample 2 – Spanish version

Four-hundred-and-thirty undergraduate Psychology students from the University of Barcelona completed the Spanish version of the questionnaire online in return for a feedback report. Eighty-three percent were female and 17% were male. Age ranged from 16 to 46 years with a median 19, mean 20.8 and a standard deviation of 4.17 years.

IRT model estimation for forced-choice and single-stimulus responses

First, the single-stimulus version of the questionnaire was analyzed. The multidimensional version of the normal ogive graded response model (Samejima, 1969) was fitted to the item responses for all 5 traits simultaneously, using the ULS estimation in *Mplus*. The five latent traits were allowed to correlate freely. The exact

model fit was relatively poor. On 1700 degrees of freedom, the English version yielded chi-square 3621.59 ($p < 0.001$), RMSEA = 0.051, and the Spanish version chi-square 4375.49 ($p < 0.001$), RMSEA = 0.068. Fitting the model one scale at a time revealed that the scale Openness had its items loading on 2 dimensions (namely imagination, and preference for complex and abstract material). The scale Conscientiousness had 2 items with highly similar content (preference for order) that shared common variance not explained by the main factor. Other scales were broadly one-dimensional and showed good fit indices when tested on their own. However, the Big Five model without any modifications was tested to estimate the model parameters and compute the MAP scores for individuals. The estimated correlations between the five traits are given in **Table 9** (above the diagonal).

Next, the forced-choice questionnaire was analyzed. After coding the forced-choice rankings as binary outcomes, the 5-dimensional IRT model with freely correlated latent traits was fitted to these data in *Mplus*, also using the ULS estimation. One item's uniqueness per block was fixed for identification. The forced-choice model yielded a better fit than the single-stimulus model: on 1640 degrees of freedom the English version yielded a chi-square of 2106.06, RMSEA = 0.025, and the Spanish version a chi-square of 2282.61, RMSEA = 0.035. Degrees of freedom and RMSEA are corrected for the number of redundancies in the model, 20.

Correlation patterns

The estimated correlations between the five dimensions in this model are given in **Table 9**. In the English version, the forced-choice correlations (given below the diagonal) are very similar to the trait correlations estimated from the single-stimulus data (above the diagonal) for all but one correlation. The correlation between traits Agreeableness and Openness is higher for the single-stimulus version (0.41) than for the forced-choice version (0.15). In the Spanish version, trait scores based on the single-stimulus responses were more positively correlated with each other than the forced-choice trait scores.

Table 9: *Estimated correlations between the Big Five markers based on the single-stimulus and forced-choice questionnaires in the empirical example*

	N	E	O	A	C
<i>English version</i>					
Neuroticism (N)	1	-.44 (.04)	-.49 (.04)	-.37 (.05)	-.33 (.05)
Extraversion (E)	-.40 (.06)	1	.52 (.04)	.49 (.04)	.29 (.05)
Openness (O)	-.48 (.07)	.48 (.06)	1	.41 (.05)	.31 (.05)
Agreeableness (A)	-.40 (.08)	.41 (.07)	.15 (.08)	1	.30 (.05)
Conscientiousness (C)	-.30 (.07)	.23 (.07)	.35 (.07)	.31 (.08)	1
<i>Spanish version</i>					
Neuroticism (N)	1	-.20 (.05)	-.30 (.05)	-.11 (.05)	-.22 (.05)
Extraversion (E)	-.15 (.07)	1	.28 (.05)	.49 (.04)	.14 (.05)
Openness (O)	-.27 (.07)	.05 (.08)	1	.27 (.05)	.24 (.05)
Agreeableness (A)	-.24 (.08)	.37 (.06)	.11 (.08)	1	.21 (.05)
Conscientiousness (C)	-.09 (.07)	-.01 (.07)	.04 (.08)	.11 (.07)	1

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

Empirical reliability and ordering of respondents

Scale empirical reliability estimates for the forced-choice data were computed based on the IRT information method described above. Reliability estimates for the single-stimulus data were also computed using equations (26) and (27). The reliability estimates for the English version ranged from 0.775 to 0.844 for the single-stimulus data, and from 0.601 to 0.766 for the forced-choice data (see **Table 10**). For the Spanish version, the reliabilities ranged from 0.783 to 0.889 for the single-stimulus data, and from 0.648 to 0.845 for the forced-choice data. It can be seen that the rank-order of scales in terms of their reliability is the same for both formats, however, the reliabilities are lower by about 0.1 for the forced-choice format. Clearly, responses to 60 items using the ordinal 5-point scale provided more information than 60 binary outcomes of rankings.

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

The MAP estimated trait scores for individuals based on single-stimulus and forced-choice responses correlated strongly (see **Table 10**), with correlations ranging from 0.69 for Agreeableness to 0.82 for Extraversion for the English sample, and from 0.70 for Agreeableness to 0.87 for Neuroticism for the Spanish sample. Interestingly, while the Openness scale was the most problematic in terms of its dimensionality, the weakest correlation between the forced-choice and the single-stimulus formats was observed for the scale Agreeableness.

Table 10: *Reliabilities and correlations between the single-stimulus and forced-choice*

<i>Big Five marker traits in the empirical example</i>					
	N	E	O	A	C
<i>English version</i>					
SS reliability	0.825	0.844	0.824	0.775	0.802
FC reliability	0.704	0.766	0.729	0.601	0.685
corr(SS,FC)	0.804	0.817	0.772	0.692	0.764
<i>Spanish version</i>					
SS reliability	0.858	0.889	0.783	0.829	0.828
FC reliability	0.824	0.845	0.648	0.718	0.721
corr(SS,FC)	0.869	0.840	0.743	0.702	0.800

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

Application 2. Customer Contact Styles Questionnaire

Instrument

The Customer Contact Styles Questionnaire (CCSQ version 7.2) is published by SHL and used in assessment for customer service and sales roles. Its 16 work-related dimensions cover a wide range of behavioral styles, with a strong emphasis on achievement motivation (SHL, 1997). Short descriptions of the scales can be found in **Table 11**.

CCSQ items are presented with 32 blocks of 4 statements (128 statements in total) so that there are no two items within a block measuring the same trait. The number of items measuring each scale varies from 7 to 10. All statements are positively worded and keyed. For each block the respondents have to rate all four statements on a 5-point Likert scale (ranging from “Strongly Disagree” to “Strongly Agree”), and then select one item that is ‘most like me’ and one ‘least like me’. Thus, the test combines both single-stimulus and forced-choice formats in one. Here is a sample block:

I am the sort of person who...

- A. generates imaginative solutions
- B. easily forgets unfair criticism
- C. needs to beat the opposition
- D. is eager to help others out

Scale scores in the questionnaire are derived by adding together the classical normative and ipsative scale scores. In this test, the normative scores are assumed to provide the ‘absolute’ standing on the 16 traits, and ipsative scores to provide additional information on ‘relative’ order of traits. This approach works well in the questionnaire, and the composite scores are shown to have better operational validities than the normative scores alone (SHL, 1997). The ipsative scores provide incremental validity over and above the normative scores.

Table 11: *Short descriptions of the 16 traits measured by the Customer Contact Styles Questionnaire (CCSQ)*

-
1. **Persuasive** - enjoys selling, negotiating and gaining commitment.
 2. **Self-control** - restrained in showing irritation or annoyance; rarely criticizes others openly; remains patient.
 3. **Empathic** - sensitive and understanding towards others; prepared to go out of their way to help.
 4. **Modest** - reserved about personal achievements and disinclined to talk about self.
 5. **Participative** - enjoys team work and wants to develop constructive relationships.
 6. **Sociable** - sociable, talkative and confident with different types of people; livens up group activities.
 7. **Analytical** - enjoys analyzing information; working with data; probing the facts and solving problems.
 8. **Innovative** - comes up with a wide range of ideas and offers imaginative or novel solutions.
 9. **Flexible** - open to new approaches and readily adapts to different circumstances.
 10. **Structured** - plans ahead; considers preparation, priority setting and structure to be important.
 11. **Detail conscious** - ensures accuracy by checking details carefully and by being neat and tidy.
 12. **Conscientious** - willing to persevere, to keep firmly to deadlines and to make sure that tasks are completed.
 13. **Resilience** - copes with external stresses and pressures by being calm, thick skinned and looking on the bright side.
 14. **Competitive** - needs to win at all costs, hates to lose and likes to be the best.
 15. **Results orientated** - sets ambitious personal targets; stimulated by challenging targets; keen to improve own performance.
 16. **Energetic** - enjoys being active; keeps busy; sustains a high level of energy over a long time.
-

In what follows, the psychometric properties of classical normative and ipsative test scores are considered first. These scores are of interest because this is how the questionnaire is currently scored. Then scores estimated from forced-choice ratings using the IRT model are compared to the classical FC and SS scores, and to the IRT-estimated SS scores.

Sample

The CCSQ UK Standardization sample, consisting of $N = 610$ respondents, was collected in 2001 using paper and pencil supervised administration. The sample was drawn from nine different organizations in industry, commerce and the public sector. Approximately half were job applicants, and the rest completed the questionnaire to provide data in return for feedback. Sixty-one percent were males, thirty-nine percent females. Most respondents were currently working in sales (61%) and customer service roles (26%). The average age was 33 years.

Properties of the classical ipsative and normative scores

Ordering of respondents on each of the 16 scales based on the CTT single-stimulus and IRT forced-choice scores was relatively similar (see **Table 15**). Cross-format scale correlations ranged from 0.50 to 0.73 (median 0.68). The scale Resilient yielded the lowest correlation across formats (0.50). The scales' classical reliabilities are given in **Table 14**. The alphas as computed from the single-stimulus ratings range from 0.78 to 0.91 with the median 0.84. The alphas as computed from the traditional forced-choice scores range from 0.57 to 0.80 with the median 0.72. It can be seen that alphas for the forced-choice rankings are substantially lower than for the single-stimulus ratings for every scale.

As we know, ipsative scores produce distorted scale correlations. In this application the average off-diagonal correlation was $r = -0.07$, as it would be the case for any ipsative questionnaire measuring 16 traits, as shown by Equation (1). To perform factor analysis on ipsative scales, it is necessary to remove one of the scales so that the correlation matrix may be inverted. For this reason principal component analysis was used for analyzing the ipsative data, and throughout the application for

consistency and comparability of results. A scree plot suggested extracting only two components explaining 34% of the variance. The emerged components are “contrast” factors typical for ipsative data (see **Table 12**). The first component has strong positive loadings from several scales related to Conscientiousness, and negative from scales related to Sociability. Selecting items from one of these two domains meant rejecting items from the other; i.e. being more Structured, Detail Conscious and Analytical led to earning lower scores on Sociable and Participative. Similarly, the second component has strong positive loadings from scales related to Drive and negative from Agreeableness. Being more Persuasive, Competitive and Results Orientated meant being less Self-Controlled, Empathic and Modest. Though somewhat interpretable, these “contrast” factors present a problem for understanding the true relationships between personality dimensions.

In contrast to the depressed ipsative correlations, the 16 normative scales correlate positively with each other overall (average off-diagonal scale correlation $r = 0.22$). Principal component analysis was performed on the normative scale scores, and a scree plot suggested extracting four components explaining 58.3% of the total variance. The components can be labeled “Conscientiousness”, “Dominance”, “Agreeableness”, and “Adaptability and Dynamism”. The rotated loadings (oblique rotation) and component correlations are presented in **Table 13**.

The average profile scores (average of standardized scores on the 16 dimensions) were distributed almost normally for the scale scores derived from the single-stimulus ratings. They ranged from $z = -1.56$ to $z = 1.44$ with mean 0.00 and standard deviation 0.51. While most people’s average profile score was around zero (68% had their average profile score within $z = 0.51$ of the mean), some people in the sample had above/below average scores on a number of scales. However, standardized ipsative scores’ averages showed no variation (they ranged from $z = -0.13$ to $z = 0.12$ with mean 0.00 and standard deviation 0.04). This extremely limited distribution can be seen in **Figure 7** (two tall columns around zero). Clearly, ipsative scores do not allow for much variation in the profile locations, and despite well-differentiated scale scores within each profile, each profile as a whole is centered on zero.

Table 12: *Rotated pattern matrix and component correlations for classical ipsative scores in the CCSQ Application*

	Conscientiousness versus Sociability	Dominance and Drive versus Agreeableness
Persuasive	-0.30	0.48
Self-control	-0.24	-0.67
Empathic	-0.22	-0.57
Modest	0.02	-0.58
Participative	-0.33	-0.37
Sociable	-0.54	0.21
Analytical	0.63	0.18
Innovative	-0.01	0.35
Flexible	-0.13	0.28
Structured	0.73	-0.05
Detail conscious	0.83	-0.05
Conscientious	0.57	-0.01
Resilience	-0.43	-0.15
Competitive	-0.17	0.45
Results orientated	0.14	0.61
Energetic	-0.41	0.36
Component correlations		
Component 1	1	-0.03
Component 2		1

Note: Loadings above +/-0.4 are set in boldface.

To summarize, the limitations of the ipsative scores obtained from the CCSQ with classical scoring procedures are obvious. Their construct validity is difficult to establish, they do not allow for variability of the profiles' locations, and finally, reliability estimates cannot be trusted due to violation of several assumptions.

Table 13: *Rotated pattern matrix and factor correlations for the classical normative scores in the CCSQ Application*

	1	2	3	4
	Conscientious ness	Dominance	Agreeableness	Adaptability and Dynamism
Persuasive	-.01	.55	.06	.34
Self-control	.10	-.52	.44	.38
Empathic	.18	-.22	.76	.00
Modest	.15	-.67	-.06	.25
Participative	-.03	.11	.69	-.05
Sociable	-.11	.38	.48	.28
Analytical	.68	.01	-.22	.21
Innovative	.22	.37	-.15	.46
Flexible	.24	.05	.17	.47
Structured	.83	.02	.01	-.05
Detail conscious	.89	-.13	-.02	-.08
Conscientious	.80	.02	.23	-.13
Resilience	-.09	-.23	-.01	.89
Competitive	.11	.66	-.02	.05
Results orientated	.47	.38	.20	.22
Energetic	.06	.26	.11	.56
Component correlations				
Component 1	1	.02	.18	.34
Component 2		1	.05	.16
Component 3			1	.25
Component 4				1

Note: Loadings above +/-0.4 are set in boldface.

IRT model estimation for forced-choice and single-stimulus responses

To overcome these limitations, the Thurstonian IRT model was applied to the forced-choice responses. After transforming the rankings into binary outcomes, there

were $6 = 4 \times 3/2$ outcomes per block, making $192 = 6 \times 32$ binary outcomes in total. These were the observed variables to be modeled as a function of the freely correlated 16 latent traits. To identify the model, one uniqueness per block was fixed to 1. Importantly, because the incomplete ranking format with “most”-“least” choices is used here, one of each 6 binary outcomes per block was not known and was treated as missing data.

Using the DWLS estimation in such a large model would lead to a very significant increase in time for establishing diagonal weights. Therefore ULS estimator with theta parameterization (Muthén, 2006) was used in this application. Despite the estimation proceeding very fast, it was necessary to switch off computing goodness-of-fit statistics and standard errors in order for the estimation to finish.

When the estimated forced-choice item parameters were converted into the threshold/slope parameterization, almost all estimated item slopes were over 0.5 in magnitude. However, the scale Competitive had one item with a near-zero slope (“resents others winning”). Unlike in CTT scoring, in IRT scoring this item will make virtually no influence on the scale score.

To estimate latent scores derived from the single-stimulus responses, the normal ogive Graded Response model (Samejima, 1997) was fitted in a multivariate fashion, to all dimensions simultaneously. Again, the ULS estimation was used, with freely estimated factor loadings and the factor variances fixed to 1. For any given item, the item factor loading obtained in the single-stimulus estimation was generally similar in magnitude to the corresponding factor loading from the forced-choice model. The same item “resents others winning” from the scale Competitive that showed a near-zero loading in the forced-choice model also had a near-zero loading in the single-stimulus model. Another item, namely “dislikes working alone” from the scale Participative, had a near-zero loading in the SS model, but discriminated well in the FC model. This is the only item in its scale that is negatively worded (compare to similar in meaning but positively phrased “likes working in a team”), and perhaps idiosyncratic use of the rating scale would explain why agreeing with

this statement did not necessarily follow from agreeing with more positively phrased statements.

After estimating IRT parameters for the SS and FC models, MAP scores on the 16 traits were computed for the standardization sample using *Mplus*. Next, properties of the IRT scores estimated from the forced-choice ratings are discussed, specifically focusing on their similarities and differences to ipsative scores obtained using classical scoring procedures, and to the IRT scores estimated from the single-stimulus presentation.

Empirical reliability and standard error of measurement

To obtain reliability estimates for the single-stimulus data, the empirical MAP information was computed (Muthén, 2006; Du Toit, 2002) independently for each trait. For the forced-choice data, reliability indices were computed using the empirical information method as described above. Item information functions were computed for non-missing responses for each individual based on their estimated trait scores, summed to produce the test information function, and finally the information given by the prior distribution was added, as described in Equation (27). In both formats, the average standard error for the sample was computed and Equation (30) was used to compute the empirical reliability estimates for each scale.

Table 14 summarizes the reliability findings for the two response formats. Standard errors of measurement for the IRT scores for a set of thetas equal 0 for all 16 scales are also given. These values are indicative of the magnitude of the standard errors for average trait scores, where the information function is likely to reach its peak, and the questionnaire reach its maximum precision.

The IRT empirical reliabilities for the single-stimulus questionnaire ranged from 0.70 to 0.87 with median 0.79. These estimates are slightly lower (on average by about 6%) than reliabilities provided by alpha for the classical normative scores. The empirical IRT reliabilities for the forced-choice questionnaire ranged from 0.71 to 0.87 with median 0.79. These are higher than alphas for the classical ipsative scores. While it has been argued that alphas underestimate the reliability in forced-

choice questionnaires with many measured traits (see **Appendix A**), it is likely that the IRT empirical information method slightly over-estimated the test reliability in this case. This is because: 1) ignoring the local dependencies in blocks of four items is likely to over-estimate the reliability; 2) in a questionnaire with positive items, there is a limit to how well the latent traits can be recovered, and the information functions become very peaked thus providing distorted estimates of empirical reliability (see the section on simulation studies with 5 factors).

Table 14: *Reliabilities of the classical scores, IRT-based empirical reliabilities and standard errors in the CCSQ Application*

CCSQ scale	Number of items	Single-Stimulus			Forced-Choice		
		Alpha	IRT empirical reliability	SE at theta=0	Alpha	IRT empirical reliability	SE at theta=0
Persuasive	7	0.80	0.70	0.53	0.68	0.76	0.32
Self-Control	9	0.89	0.83	0.40	0.72	0.80	0.32
Empathic	9	0.83	0.78	0.44	0.74	0.77	0.35
Modest	9	0.88	0.82	0.37	0.75	0.77	0.36
Participative	10	0.90	0.87	0.41	0.80	0.83	0.30
Sociable	8	0.78	0.73	0.46	0.68	0.76	0.36
Analytical	8	0.79	0.73	0.49	0.66	0.83	0.31
Innovative	9	0.91	0.87	0.32	0.78	0.82	0.29
Flexible	7	0.82	0.76	0.47	0.62	0.74	0.36
Structured	8	0.86	0.81	0.42	0.73	0.83	0.32
Detail Conscious	7	0.85	0.78	0.44	0.75	0.87	0.26
Conscientious	7	0.87	0.81	0.38	0.72	0.81	0.31
Resilient	9	0.83	0.76	0.43	0.64	0.71	0.43
Competitive	7	0.82	0.79	0.40	0.71	0.84	0.26
Results Orientated	7	0.82	0.77	0.43	0.57	0.78	0.38
Energetic	7	0.87	0.81	0.41	0.75	0.75	0.38
MEDIAN		0.84	0.79	0.42	0.72	0.79	0.32

Ordering of respondents

Scores produced from the same response format yielded similar ordering of people (see **Table 15**). Correlations between IRT and classical scores based on the single-stimulus format were nearly perfect, ranging from 0.93 to 0.99 (median 0.96). Correlations between classical and IRT scores based on the forced-choice responses ranged from 0.82 to 0.90, with the median correlation 0.87. This is lower than near-perfect correlations between SS scores derived from summing Likert scale responses and estimated with IRT. It is clear that despite being derived from the same responses, differences between IRT forced-choice and classical ipsative scores are not trivial.

Ordering of respondents on each of the 16 scales based on the IRT single-stimulus and IRT forced-choice scores was similar (see **Table 15**). Cross-format scale correlations for 15 scales were 0.61 or above (median 0.66). The scale Resilient correlated across formats at 0.45, consistently with its low correlation across formats when classical scores were considered. A possible reason for this is that item pairs related to the desirable scale Resilient showed higher average thresholds in the forced-choice format than items from other scales. This means that it was more “difficult” to prefer statements from this scale (FC format), than it was to agree with these generally desirable statements on their own (SS format).

As can be seen from **Table 15**, the IRT scoring did not increase correlations between the SS and FC scores compared to the CTT scoring. The correlations across formats remained roughly the same. The most likely reasons for not achieving higher cross-format correlations are: 1) small number of items per scale in this application, 2) many positively correlated scales in this application (positive average scale inter-correlation). These points will be discussed in more detail in the Discussion. However, similar magnitude of cross-format correlations for CTT and IRT scale scores do not mean that the IRT methodology did not deliver any advantages, as we shall see. These advantages will become clear when the test’s covariance structure is considered, and whole personality profiles are considered rather than isolated scales.

Table 15: *Correlations between classical scores and IRT-based scores in the CCSQ Application*

	Cross-method (CTT vs. IRT)		Cross-format (FC vs. SS)	
	SS	FC	CTT	IRT
Persuasive	0.93	0.83	0.69	0.65
Self-Control	0.97	0.86	0.63	0.64
Empathic	0.97	0.84	0.63	0.62
Modest	0.99	0.89	0.58	0.63
Participative	0.95	0.88	0.71	0.69
Sociable	0.95	0.87	0.72	0.66
Analytical	0.95	0.87	0.65	0.64
Innovative	0.98	0.90	0.69	0.70
Flexible	0.97	0.82	0.63	0.61
Structured	0.96	0.89	0.67	0.71
Detail Conscious	0.95	0.90	0.70	0.72
Conscientious	0.96	0.87	0.69	0.72
Resilient	0.98	0.85	0.50	0.45
Competitive	0.96	0.90	0.73	0.77
Results Orientated	0.95	0.86	0.69	0.66
Energetic	0.98	0.85	0.67	0.67
MEDIAN	0.96	0.87	0.68	0.66

Table 16: *Estimated correlations between the CGSQ scales based on the single-stimulus and forced-choice responses*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Persuasive	1*	0.03	0.16	-0.20	0.12	0.44	0.29	0.41	0.22	0.15	0.06	0.15	0.23	0.46	0.44	0.36
2 Self-Control	-0.26	1*	0.64	0.33	0.21	0.19	0.23	0.05	0.33	0.25	0.32	0.33	0.43	-0.06	0.25	0.17
3 Empathic	0.00	0.60	1*	0.18	0.38	0.40	0.22	0.11	0.31	0.27	0.32	0.39	0.18	-0.05	0.32	0.20
4 Modest	-0.09	0.38	0.31	1*	0.03	-0.23	0.09	-0.08	0.02	0.15	0.24	0.21	0.18	-0.20	-0.03	0.00
5 Participative	-0.01	0.27	0.37	0.16	1*	0.28	0.08	0.15	0.24	0.10	0.08	0.18	0.17	0.09	0.28	0.20
6 Sociable	0.35	0.11	0.25	-0.09	0.29	1*	0.08	0.31	0.35	0.15	0.09	0.25	0.35	0.21	0.41	0.45
7 Analytical	0.23	0.14	0.19	0.17	0.10	0.00	1*	0.44	0.40	0.53	0.69	0.48	0.29	0.10	0.52	0.20
8 Innovative	0.32	-0.02	0.02	0.02	0.11	0.13	0.53	1*	0.38	0.27	0.22	0.21	0.27	0.29	0.49	0.37
9 Flexible	0.18	0.21	0.27	0.11	0.23	0.28	0.49	0.46	1*	0.34	0.34	0.43	0.42	0.17	0.60	0.42
10 Structured	0.05	0.17	0.23	0.21	0.05	-0.01	0.57	0.22	0.34	1*	0.75	0.69	0.21	0.21	0.49	0.32
11 Detail Conscious	-0.03	0.18	0.21	0.28	0.07	-0.06	0.72	0.24	0.30	0.75	1*	0.71	0.19	0.12	0.43	0.22
12 Conscientious	0.07	0.21	0.33	0.22	0.16	0.14	0.46	0.18	0.38	0.66	0.66	1*	0.20	0.19	0.62	0.27
13 Resilient	0.01	0.38	0.10	0.28	0.10	0.29	0.13	0.19	0.32	0.05	0.09	0.14	1*	0.08	0.29	0.40
14 Competitive	0.53	-0.26	-0.19	-0.13	-0.01	0.11	0.00	0.12	0.04	0.01	0.05	0.11	0.02	1*	0.49	0.30
15 Results Orientated	0.45	0.07	0.18	0.02	0.26	0.33	0.52	0.46	0.55	0.46	0.43	0.60	0.15	0.42	1*	0.52
16 Energetic	0.22	0.00	0.05	0.02	0.10	0.33	0.17	0.27	0.32	0.28	0.16	0.27	0.29	0.24	0.43	1*

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

Correlation patterns and principal components

Correlations between latent trait scores as estimated from both single-stimulus and forced-choice responses using IRT are given in **Table 16**. The average off-diagonal correlation between the 16 scores was 0.26 for the single-stimulus test, and 0.21 for the forced-choice test. The correlations between IRT scores for both formats are similar, and different from classical ipsative scores, which yielded a negative average scale correlation (-0.07) in accordance with Equation (1).

To investigate the structure underlying the 16 traits, principal component analysis was performed on the MAP estimated single-stimulus scores. A scree plot suggested extracting four components (explaining 71% variance). **Table 17** shows rotated loadings (oblique rotation), and correlations between the components. The components can be labeled as “Conscientiousness”, “Dominance”, “Agreeableness”, and “Adaptability and Dynamism”. This solution is almost identical to the one obtained from the classical normative scores shown in **Table 13**.

Principal component analysis with the IRT estimated forced-choice scores also suggested extracting four components (explaining 63.4% variance). The components can be labeled as “Conscientiousness”, “Dominance and Drive”, “Agreeableness” and “Adaptability and Dynamism” (see rotated loadings and correlations between components in **Table 18**). This solution is strikingly similar to the one derived from the single-stimulus scores. The only difference is that the scale Results Orientated has a stronger loading on the second component, making it more motivation-related.

Table 17: Rotated pattern matrix for IRT scored single-stimulus ratings in the CCSQ

	<i>Application</i>			
	1	2	3	4
	Conscientiousness	Dominance	Agreeableness	Adaptability and Dynamism
Persuasive	.08	.63	.09	.39
Self-control	.11	-.52	.50	.35
Empathic	.19	-.20	.83	-.03
Modest	.21	-.74	-.06	.24
Participative	-.01	.12	.74	-.05
Sociable	-.15	.38	.54	.38
Analytical	.79	.03	-.17	.20
Innovative	.23	.36	-.15	.51
Flexible	.27	.06	.20	.53
Structured	.91	.01	.02	-.06
Detail conscious	.97	-.13	.00	-.09
Conscientious	.87	.01	.24	-.11
Resilience	-.08	-.26	.02	.94
Competitive	.17	.69	-.02	.11
Results orientated	.54	.40	.20	.27
Energetic	.06	.25	.12	.61
Component correlations				
Component 1	1	.04	.25	.39
Component 2		1	.04	.19
Component 3			1	.32
Component 4				1

Note: Loadings above +/-0.4 are set in boldface.

Table 18: *Rotated pattern matrix for IRT scored forced-choice ratings in the CCSQ*

	<i>Application</i>			
	1	2	3	4
	Conscienti- ousness	Dominance and Drive	Agreeableness	Adaptability and Dynamism
Persuasive	.00	.82	.07	.01
Self-control	.03	-.71	.39	.14
Empathic	.16	-.34	.72	-.19
Modest	.22	-.57	.08	-.03
Participative	.00	-.04	.71	-.02
Sociable	-.33	.29	.59	.32
Analytical	.79	.01	-.15	.22
Innovative	.26	.21	-.17	.55
Flexible	.32	-.01	.17	.60
Structured	.86	-.05	-.02	-.05
Detail conscious	.91	-.13	-.09	-.04
Conscientious	.76	.02	.26	-.01
Resilience	-.21	-.44	-.08	.81
Competitive	.05	.71	.02	-.03
Results orientated	.52	.48	.30	.30
Energetic	.09	.24	.15	.51
Component correlations				
Component 1	1	.00	.11	.15
Component 2		1	-.06	.16
Component 3			1	.16
Component 4				1

Note: Loadings above +/-0.4 are set in boldface.

Individual profiles

In this section, the relationship between IRT-scored FC and SS profiles are investigated, as well as the relationships between forced-choice profiles obtained using classical scoring methodology and IRT. A known problem of forced-choice scores based on classical scoring procedures is that the average profile score (average of all standardized scale scores) is bound to be around zero, so that it is impossible to score high or low on all scales. Indeed, this is what was found in this application.

However, IRT scoring overcomes this problem. The IRT-estimated forced-choice average profile scores (the average of all IRT forced-choice scores) are distributed as shown in **Figure 7**. They range from -1.13 to 1.07 (mean is -0.01, and standard deviation 0.37). The IRT-estimated single-stimulus average profile scores range from -1.49 to 1.86 (mean 0.00, $SD = 0.56$). It can be seen that forced-choice IRT scores are indeed considerably more similar to the single-stimulus scores (regardless of whether they have been obtained using classical scoring procedures or IRT) than to the ipsative scores.

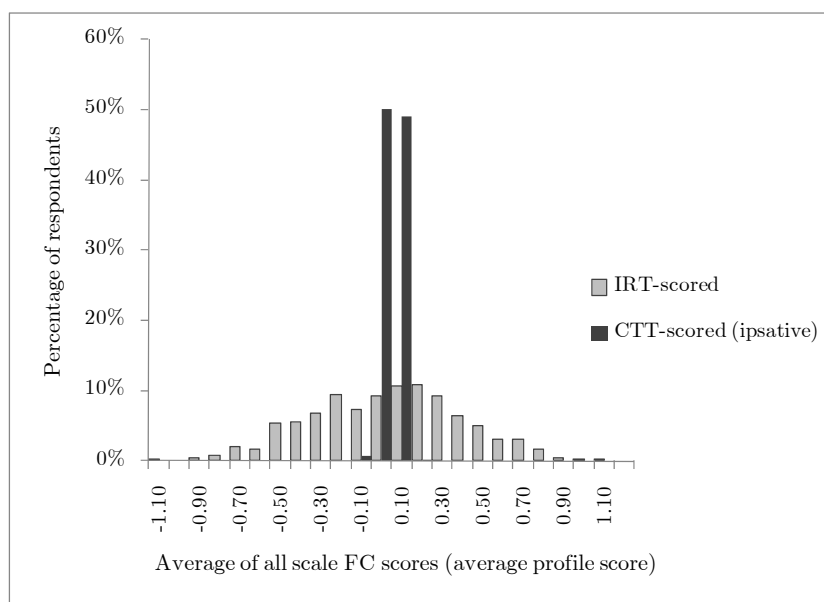


Figure 7: *Distributions of individual average profile scores based on IRT and CTT scoring of forced-choice responses in the CCSQ Application*

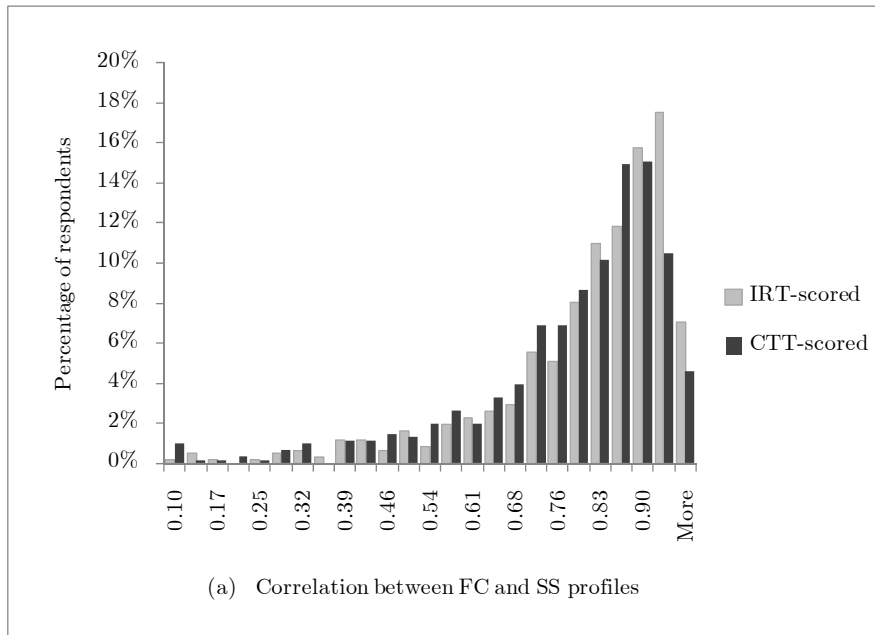
The main difference between FC and SS scores is that there are considerably more positively located profiles in the single-stimulus completion than in the forced-choice completion. Over-rating bias might be responsible for some positively shifted single-stimulus profiles.

Next, the similarity of individual profiles based on SS and FC responses is assessed. The profiles' shapes (indicating *relative* ordering of scales) were compared by correlating 16 IRT scale scores across formats (single-stimulus versus forced-choice) for the same individual (see **Table 19**). Profile similarity coefficients ranged from 0.10 to 0.98 (median 0.84) and were distributed as shown in **Figure 8a**. Most people (79%) had profiles with similarity 0.7 or higher. Only a few people (3%) had profiles with similarity less than 0.4. Those who had dissimilar profiles, tended to have “flatter”, less differentiated SS profiles. It is easy to see why this is the case - those with similar true scores on all scales would find it difficult to make choices between statements of similar utility (e.g. McCloy et al., 2005). Many such choices will be random; therefore, the forced-choice profile will be also “flat”, and therefore largely uncorrelated with the normative profile.

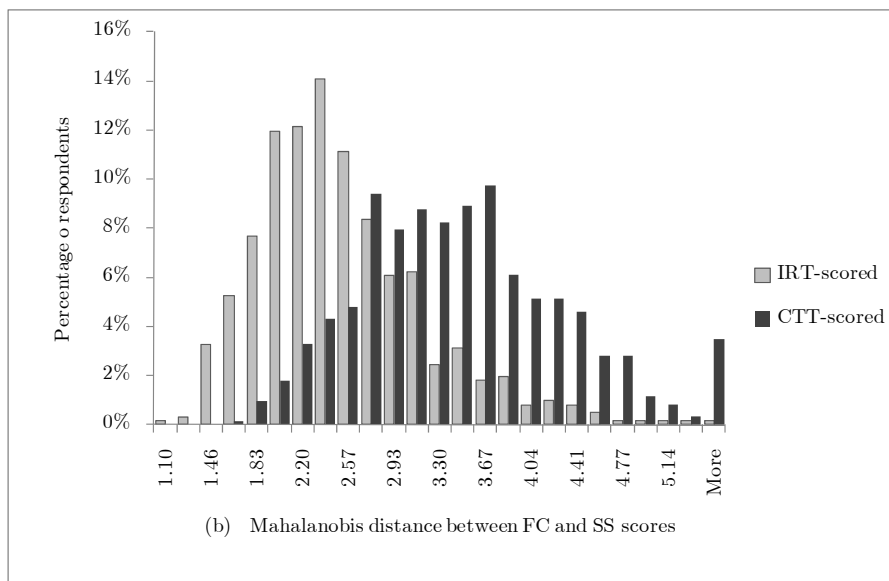
The locations of IRT-scored FC and SS profiles (indicating *absolute* positions of trait scores) were also compared. Similarity of location was measured by computing the Mahalanobis distance between the MAP estimated trait scores for both formats. The Mahalanobis distance measures the distance between 2 points in a multidimensional space, taking to account non-orthogonal nature of the axes (personality traits). This measure is particularly suitable in this Application, where the 16 traits are correlated with each other and the Euclidean distance would provide wrong estimates. For simplicity and consistency of results, the single-stimulus CCSQ correlation matrix given in **Table 16** was used as an indicator of the latent space structure in all computations.

In this approach, each individual's 16 trait scores derived from a particular test format (or a scoring method) are considered as a point in the 16-dimensional space. For the same individual, distances between points representing his/her IRT scores on the FC or SS measure were computed. The distances between the FC and

SS individual locations ranged from 1.10 to 5.51 (median 2.33) and were distributed as shown in **Figure 8b**.



(a) *distribution of correlation coefficients between individual profiles*



(b) *distribution of Mahalanobis distances between individual profiles*

Figure 8: *Distributions of profile similarity coefficients for IRT-scored single-stimulus and forced-choice responses in the CCSQ Application*

Table 19: *Average of individual profile correlations and distances for classical and IRT scores in the CCSQ Application*

	Cross-method (IRT vs. CTT)		Cross-format (SS vs. FC)	
	Single-stimulus	Forced-choice	CTT	IRT
Correlations between profiles				
Minimum	0.82	0.75	-0.17	0.10
Maximum	0.99	1.00	0.97	0.98
Mean	0.96	0.96	0.76	0.79
SD	0.02	0.02	0.17	0.16
Median	0.97	0.97	0.81	0.84
Mahalanobis distances between profiles				
Minimum	0.36	0.65	1.55	1.10
Maximum	3.23	3.35	6.49	5.51
Mean	1.36	2.02	3.39	2.43
SD	0.48	0.43	0.87	0.68
Median	1.28	1.99	3.31	2.33

To help evaluate the magnitude of the profile distances, they can be compared to cross-method results (distances between single-stimulus estimates based on CTT and IRT, and forced-choice estimates based on CTT and IRT), and also to the benchmark distances between CTT scores across the formats (i.e. between the normative and ipsative scores). All results are summarized in **Table 19**. First, it can be seen that although the cross-format distances between IRT estimated FC and SS scores are about 80% greater than the same-format distances between classically scored and IRT estimated single-stimulus scores, they are much smaller than the cross-format distances for the CTT estimated scores (ipsative versus normative, which are about 42% further from each other than the IRT scores are). It is clear that the classical ipsative scores are distorted – not only they are further apart from the single-stimulus scores; they are also far away from the IRT-estimated forced-choice scores.

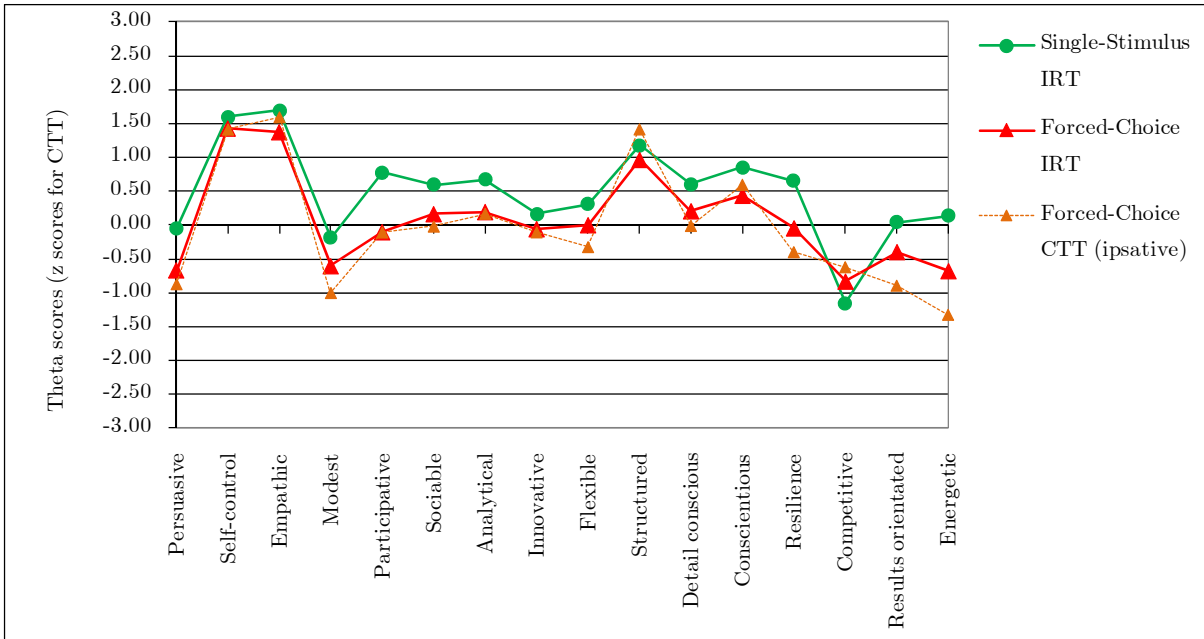
To conclude, while the IRT and classically scored profiles derived from the single-stimulus responses are very similar in both shape and location, the IRT and classically scored profiles derived from the forced-choice responses are very similar in shape but not so much in location. Classically scored forced-choice profiles, as was discussed before, are always centered on zero.

What do these results mean in practice? Let us consider four real participants from the CCSQ standardization sample. Two participants are representative of the majority in this sample, with single-stimulus and forced-choice profiles being similar in shape and location, and two representing very rare, extreme situations.

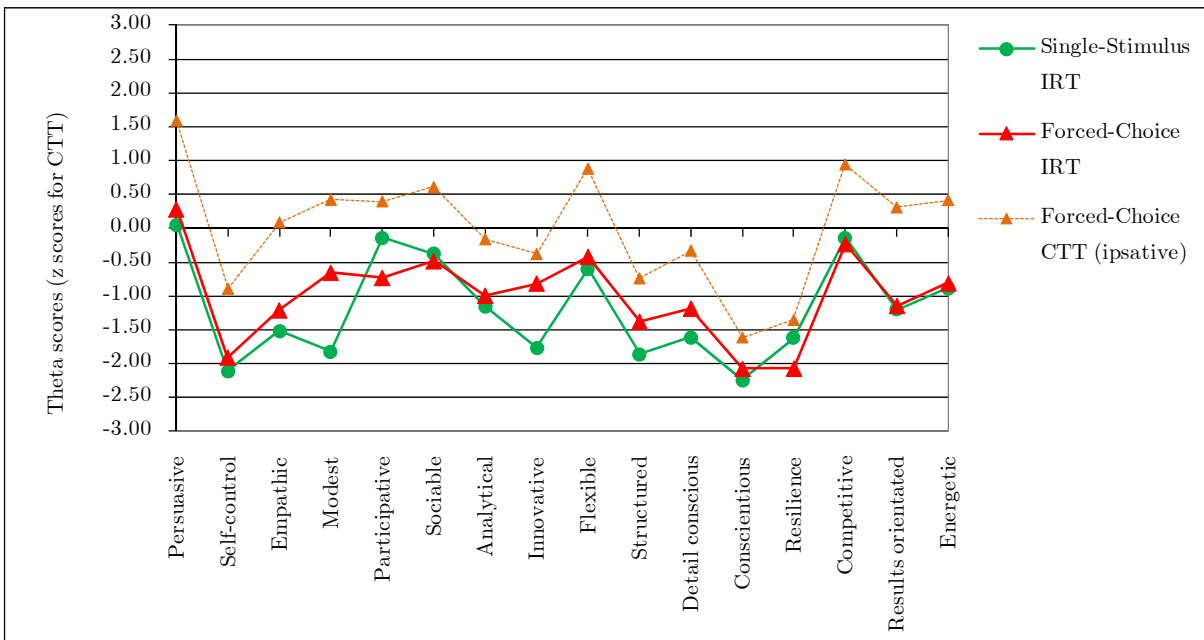
Typical profiles

Respondent A completed the CCSQ as a part of a selection process (see **Figure 9a**). The single-stimulus profile for this individual is slightly elevated (average profile score 0.49 for IRT SS scores). The traditional ipsative scores provided a reasonable approximation to the SS scores (similarity with the classical normative profile $r = 0.81$ and $M.dist. = 2.66$), however, the Thurstonian IRT approach improved the profile similarity ($r = 0.92$) and location ($M.dist. = 1.54$).

Respondent B is a job incumbent, who completed the CCSQ in return for feedback (see **Figure 9b**). Both SS and FC IRT-based profiles are dominated by lower than average scores, and the average profile location is below average (-1.19 for IRT SS theta scores). This respondent's SS and FC IRT-based profiles are similar in shape ($r = 0.81$) and location ($M.dist. = 2.09$). The classical FC (ipsative) profile is given for comparison. Despite being very similar in shape, it is located well above the IRT recovered FC profile and centered on zero; its Mahalanobis distance to the classical normative profile is almost double the value of the IRT-based distance ($M.dist. = 3.78$).



(a) *Highly similar SS and FC profiles (Respondent A, job applicant)*



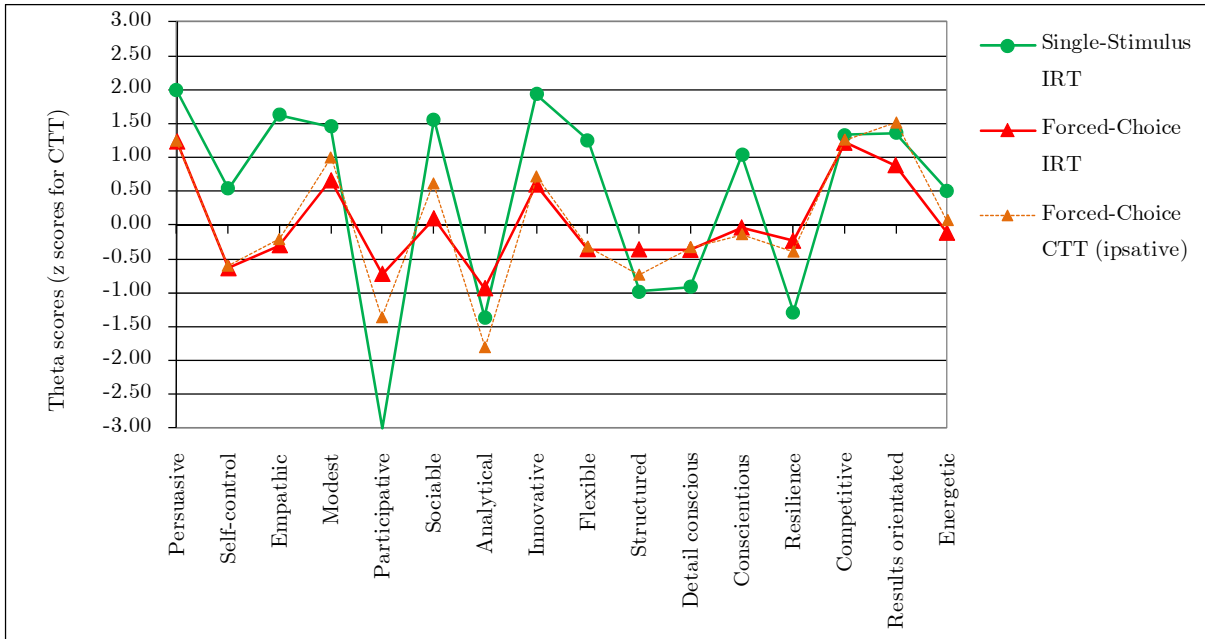
(b) *Low average location for SS and FC profiles (Respondent B, job incumbent)*

Figure 9: *Sample CCSQ personality profiles based on IRT scores and classical ipsative scores (typical cases)*

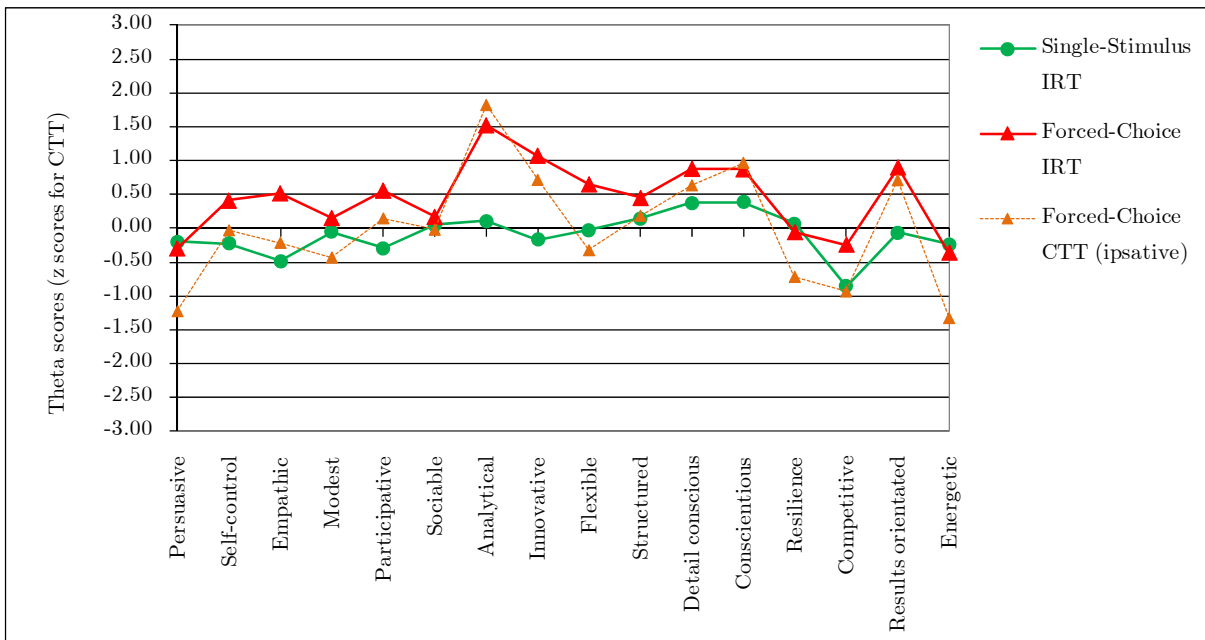
Extreme profiles

Respondent C is a job incumbent who completed the CCSQ in return for feedback (see **Figure 10a**). The single-stimulus profile of this individual is quite striking – scores on all scales are rather extremely high or extremely low. Intentional distortion is unlikely here because very desirable traits such as Participative and Resilient show extremely low scores. The most likely reason for such a differentiated profile is extreme responding style, where the individual either strongly agrees or strongly disagrees with statements. While the forced-choice profile is quite similar in shape to the SS profile ($r = 0.68$), it is clearly less extreme. The distance between the IRT-based SS and FC profile is very large $M.dist = 5.51$. Interestingly, despite the ipsative profile appearing closer to the SS profile on **Figure 10a**, it is actually even further away than the IRT-based FC profile based on the Mahalanobis distance ($M.dist = 5.88$).

Respondent D is a job applicant who completed the CCSQ as a part of a selection process. His single-stimulus IRT scores are largely average (average profile score -0.09 ; see **Figure 10b**). The forced-choice profile is also quite “flat” and located closely to the SS profile for all except three scales, Analytical, Innovative and Results Orientated. When presented with single-stimulus items, this individual tended to select the middle rating options. Being forced to differentiate between statements, he however judged statements related to analysis, originality and drive for results to describe self better than other statements. In this case, forced choice helped identify strength for analytical thinking, which was masked by the central tendency responding in the single-stimulus format.



(a) *Extreme responding on SS profile (Respondent C, job incumbent)*



(b) *"Flat" SS and FC profiles (Respondent D, job applicant)*

Figure 10: *Sample CCSQ personality profiles based on IRT scores and classical ipsative scores (extreme cases)*

Discussion

In this dissertation, an Item Response model suitable for modeling responses to multidimensional forced-choice questionnaires with dominance items was introduced. The model proposed here is an IRT formulation of the Thurstonian second-order factor model for comparative data introduced in Maydeu-Olivares and Böckenholt (2005) applied to the problem at hand.

Model and parameter estimation

The Thurstonian IRT model is suitable for forced-choice instruments composed of many scales and items and is estimated fast, however, current computing capabilities prevent from computing goodness of fit indices and standard errors in large applications. Extant research performed on smaller tests suggests that the model fits well in applications. Also, it is equivalent and hence yields an equivalent fit to the second-order Thurstonian factor model (Maydeu-Olivares and Böckenholt, 2005) when the latter can be computed. Once the model parameters are known, MAP estimation of trait scores is performed very fast for models of any size.

The simulation studies presented here show that the true model parameters (trait correlations, factor loadings, thresholds, and residual errors) are recovered very accurately from the binary outcomes in all reasonable designs. Some forced-choice designs are simply not recommended, such as questionnaires with 2 traits and positively keyed items only. Some empirically unidentified models and large standard errors encountered in this design should not cause any concerns for the Thurstonian IRT model's use. The poor results obtained in these designs do not reflect the limitations of the model or estimation method employed, but rather, the limitations of the forced-choice format. On the other hand, all models that yielded sufficiently accurate recovery of the true score also showed good convergence, no problems with identification, and accurate parameter estimation. Perhaps the most impressive is accurate recovery of true correlations between traits, e.g. in the simulation study with 5 traits. Clearly, the IRT modeling overcomes the distortion of correlations between measured traits typical for ipsative data.

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

Recommendations for forced-choice questionnaire design

Example models in this research were chosen to answer important questions about strengths and limitations of forced-choice questionnaires with dominance items. Despite many discussions in the literature, many of these questions remain controversial as the evidence is largely based on the inadequate classical modeling leading to ipsative scores, and research results are based on specific questionnaires with very different properties to enable meaningful generalization. Results of the simulation studies have important implications on how forced-choice tests are designed and used in the future. The most important points will be discussed here.

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

Keyed direction of items

When the forced-choice design produces binary outcomes from comparing items from different traits keyed in the same direction, and approximately the same

number of binary outcomes from comparing items keyed in opposite directions, the trait recovery is good with any number of traits, and any trait correlations. This is because items keyed in the same direction contribute to the measurement of the *difference* between 2 trait scores; and items keyed in opposite directions measure the *sum* of the 2 traits involved. When sums and differences of all questionnaire traits are known, their absolute values can be immediately deduced. Conversely, when only differences between traits are known, which is the case in forced-choice designs with items keyed in the same direction, the recovery largely depends on the number of traits assessed – and this is the next factor in this discussion.

A recent approach to creating forced-choice questionnaires involved unidimensional pairings (Stark et al., 2005). It has been argued that comparing items from the same scale is essential to set the scale origin (Chernyshenko et al., 2009). While it is clear from the results of the present research that this is not necessary, items keyed in opposite directions that measure the same trait can be used together in blocks to provide information on the latent trait directly, as the binary outcomes of such comparisons will depend on only one trait. The comparative nature of the forced-choice format means that the dominance items measuring the same trait will only provide sizeable amount of information when their factor loadings are very different (Maydeu-Olivares & Brown, 2010), as it is the case with items keyed in opposite directions.

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

Number of traits

When the number of traits is large, and traits are not strongly positively correlated, any forced-choice designs will reliably locate trait scores provided that sufficient numbers of good quality items are used. That is, it is possible to locate absolute trait scores using only positively keyed items when the number of traits

assessed is large. Baron (1996) shows that even ipsative questionnaires with all positive items measuring many relatively independent traits (30 or more) correlate strongly with their single-stimulus counterparts. Why it is important that traits are relatively independent is the next point of the discussion.

The simulation studies show that when assessing only 2 traits, positively keyed items on their own cannot recover the absolute latent trait scores. In the simulation studies with 5 traits, where traits on average correlated weakly, the true score recovery was good for designs with positive items only (except for the blocks of 2, where the number of binary outcomes was not sufficient). How do binary outcomes measuring only *differences* between traits (and this is the case with items keyed in the same direction) provide information on *absolute* trait scores when the number of traits is large?

When only 2 traits are measured, the information about the first trait depends only on the second trait (and vice versa). As can be seen from **Figure 4**, there is sizeable amount of information for scores that are similar (for example (-2,-2) or (2,2)), but virtually no information for scores that are different (for example (2, -2)). There are many combinations of two scores possible that are very different from each other. For instance, assuming normally distributed traits that are uncorrelated with unit variance, trait scores are different by more than 0.5 standard deviations for around 75% of cases. Therefore for most combinations of latent scores, the test information provided by such a questionnaire will be very low.

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

Extending this logic further, for 30 independent traits there is less than 0.03% chance that all trait scores will be different by 0.5 standard deviations or above. In

this case the information about one trait is conditional on 29 other traits, and because many of them will be similar to the target trait the information will be high for most combinations of scores.

For instance, in the CCSQ questionnaire with its 16 traits, the correlations between the single-stimulus and forced-choice scores were quite modest. With more scales cross-format relationships become stronger and single-stimulus and forced-choice profiles more similar. For example, the Occupational Personality Questionnaire (OPQ; SHL, 2006) measuring 32 personality traits, yields stronger cross-format relationships. Correlations between scores derived from single-stimulus and forced-choice test versions are higher than correlations for the CCSQ found here.

Correlations between traits

Given the item parameters, comparing items keyed in the same direction from positively correlated traits is less effective than if the traits are uncorrelated. The same comparison is even more effective if the traits are correlated negatively. This was apparent from the information functions provided by equations (23) and (24). For binary outcomes of comparisons between items measuring uncorrelated traits, only the focus trait contributes to the information. However, for pairs involving correlated traits, the other trait involved will also contribute to the information. It will increase the information if correlated negatively with the target trait, and reduce it if correlated positively.

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

Block size

By using blocks of different sizes in the simulation studies with 5 traits it is shown that the same items can be made “work harder” by simply combining them in larger blocks. This is because, given the same number of items, the number of binary outcomes will increase when the block size increases. For example, 60 items will produce only 30 binary outcomes when put in blocks of 2; 60 binary outcomes when put in blocks of 3; and 90 binary outcomes when put in blocks of 4. In other words, using larger blocks is attractive because it saves producing and trialing new items, which can be time consuming and expensive.

Of course the block size cannot be increased indefinitely, because readability will worsen and cognitive load will increase dramatically as respondents perform $\tilde{n} = n(n - 1) / 2$ mental comparisons for each n items. The item length can also be a problem, particularly when 4 or more items are compared in one block. In practice, blocks of 4 items are probably the upper limit for forced-choice tests.

To summarize, adhering to the above recommendations (i.e. balancing the number of traits and their correlations, the direction of items, the number of items and the block size) is important for the quality and usefulness of a resulting questionnaire. Provided these factors have been taken into account, most personality items can be used in forced-choice questionnaires. Thousands of dominance personality items have been written and translated to different languages over years. This research has shown how these simple items can be effectively used in forced-choice questionnaires.

Information and test reliability

A method of estimating the empirical test reliability is described, which is based on computing MAP information and the average error variance for a scored sample. Reliability figures produced by this information method were compared to reliabilities assessed through correlations between the estimated and the true scores. The extent to which the information estimates under the simplifying assumption of local independence are accurate in forced-choice questionnaires was also investigated.

The general rule is that the information method provides accurate estimates in designs where increasing the number of binary outcomes improves the latent trait estimation. This was the case in the simulation studies with 2 and 5 traits where positively and negatively items were combined together in blocks. In these studies, the information method provided sufficiently accurate estimates of test reliability even for triplets and quads, where the local independence does not hold and the simplifying assumption of local independence had to be made. Ignoring correlated errors led to a very minor overestimation of reliability – for blocks of 3 the reliability was overestimated by about 2%, and for blocks of 4 by about 3%. The researcher must also be aware that for very short questionnaires, the information method might underestimate the reliability due to a greater trait score “compression” by the MAP estimator.

In questionnaires using only positively keyed items, as is shown earlier, the accuracy of the latent trait recovery depends heavily on the number of traits assessed. In such questionnaires, after a maximum possible level of latent trait recovery has been reached, increasing the number of binary outcomes will not improve it further. In this case the information method might overestimate the reliability for blocks of any size. This is not due to ignoring the correlated errors in blocks of 3 or more items, as overestimation of the same magnitude also occurs for longer questionnaires with blocks of 2 (item-pairs). For instance, if the number of item-pairs is doubled in the Big Five simulated study, the latent trait estimation hardly improves but the information grows. However, it does not grow uniformly across the latent trait distribution – instead, it becomes very peaked in areas where latent trait scores are very similar to each other, and is almost zero elsewhere. The information function still “works” in such designs, however, the empirical reliability fails to reflect very varied levels of test information at different trait scores. To conclude, the information method of computing reliability is recommended only when the information is known to be relatively uniform.

Response biases and application results

In absence of true scores in real-world applications, responses to single-stimulus items are often used as a proxy for the true score, as was done in the two applications of the present research. However, this approach assumes that no systematic biases affect the responses. On contrary, research often shows that different types of biases can be present when rating scales are used (Van Herk, Poortinga & Verhallen, 2004; Murphy, Jako & Anhalt, 1993). Using the single-stimulus responses as indicators of the absolute trait standing is particularly dangerous when data is collected in a high stakes situation. The reason to create forced-choice measures was to reduce biases common in rating scales, therefore relying on the latter as appropriate indicators of true scale standing is inconsistent with available knowledge.

For instance, in the CCSQ standardization sample, there are significant differences in mean scale scores between the applicant and research sub-samples, which are very similar in demographic composition. The biggest mean difference is observed for the scale Conscientious (standardized difference is $d = 0.86$), corresponding to a large effect size. Therefore, any real-world application findings have to be interpreted with this limitation in mind. Only simulation studies with known true scores can address the trait recovery question properly. Also, once the Thurstonian IRT model is well established in personality research, it is possible that forced-choice questionnaires will be used as less biased indicators of absolute trait standing, providing it has been shown in simulations that the scores are reliable.

A better model fit obtained in the forced-choice version of the Big Five questionnaire than in the single-stimulus version might be an indicator of certain reduction in response biases. For instance, responding in socially desirable manner is often associated with such personality traits as Agreeableness and Extraversion. In the work settings, Conscientiousness is perceived as very desirable. In the educational settings, Openness is seen as important. It is possible that inflation of responses to items from these traits by some individuals is responsible for several higher correlations in the single-stimulus version than in the forced-choice version.

General similarity of item parameters in the single-stimulus and forced-choice IRT models would be expected, however, they are not guaranteed. For instance, in the CCSQ application item parameters obtained from the forced-choice responses were largely comparable to parameters from the single-stimulus responses. However, if the single-stimulus responses are affected by response biases associated with rating scales, their parameters will reflect such distortions. Extant research shows that while some simpler distortions can be modeled (e.g. random intercept, see Maydeu-Olivares & Coffman, 2006); others such as heterogeneity of item discriminations are well beyond capabilities of current factor-analytic approaches (Ansari, Jedidi & Dube, 2002). The forced-choice responses might see some of the biases eliminated or reduced, therefore yielding different parameters.

It has been long known by the developers of forced-choice measures that, when put in blocks, items can interact with each other in ways that cannot be fully envisaged from the single-stimulus presentation. Sometimes the forced-choice format can change the item discrimination through interaction with other items. One conclusion from this is that, strictly speaking, one cannot rely on single-stimulus item trials to predict the parameters for a forced-choice test fully. This is why it is very useful to have a way of estimating item parameters for the forced-choice test in actual forced-choice trials. Such a method is introduced here.

In the CCSQ application it was shown that single-stimulus and forced-choice profiles are similar in most cases. An interesting question, however, is when the profiles depart from each other, which is the “correct” one? Uniform response bias in single-stimulus responses is one possible reason for the profile shift. Another reason might be less reliable forced-choice scores for some questionnaire scales. In cases where there is a strong reason to suspect the uniform response bias (like in the case presented in **Figure 10a**), a forced-choice test might provide a more accurate profile. Central tendency responding can produce a “flat” profile – in this case, forcing the respondent to choose might increase the profile’s differentiation (like in the case presented in **Figure 10b**). On the other hand, it may prove difficult to get to true scores with the forced-choice format when the individual’s true scores on all measured scales are very close to each other. In this situation, responding to the

forced-choice questionnaire with items of similar difficulty can be very frustrating for this individual. Either the single-stimulus format or the forced-choice format with items of varying difficulties (for instance keyed in different directions) might be preferable.

Future research directions

The proposed model can be used to analyze existing forced-choice data and to aid development of new forced-choice questionnaires. One example is the shortening and rescoring of the ipsative OPQ32i (Brown, 2009; Brown & Bartram, 2009). It is also easy to see how the proposed model can be used in the future to establish methods and criteria for investigating measurement equivalence for forced-choice questionnaires, for example structural equivalence, differential item functioning (DIF) etc. More practical applications using this model can also be easily envisaged, such as investigations of pros and cons of different questionnaire designs, investigations of response distortion, and validity-related research.

Conclusions

The Thurstonian IRT model introduced here describes the decision process of responding to forced-choice personality questionnaires measuring multiple traits. This model can be used with any forced-choice instrument composed of items fitting the dominance response model, with any number of measured traits, and any block sizes (i.e. pairs, triplets, quads etc.). This makes it widely applicable to many existing forced-choice questionnaires such as the Occupational Personality Questionnaire (OPQ; SHL, 2006), the Customer Contact Styles Questionnaire (CCSQ; SHL, 1997), the Survey of Interpersonal Values (SIV; Gordon, 1976), the Kolb Learning Style Inventory (Kolb & Kolb, 2005) and many others. The Thurstonian IRT model can be embedded within a familiar SEM framework to be estimated and scored by general-purpose software (*Mplus* was used throughout this dissertation). The model also provides means of estimating reliability for forced-choice questionnaires, which has been problematic under the classical scoring model (Tenopyr, 1988; Baron, 1996).

Practitioners have long pointed out that even classically scored forced-choice tests, if constructed well, provide results that are valid and comparable with normative data (Karpatschhof & Elkjaer, 2000). The bumblebee should not fly, but somehow it did. Introducing the appropriate IRT modeling of the decision process behind responding to forced-choice items effectively overcomes the limitations of ipsative data. The Thurstonian IRT model allows using the forced-choice format, which reduces certain response biases, while getting the benefits of standard data analysis techniques that users of single-stimulus questionnaires have enjoyed.

As is shown here, creating a forced-choice questionnaire is a much more complicated endeavor than creating a single-stimulus questionnaire, because it requires consideration of many more factors. Provided these factors are carefully taken to account, and sufficient work has been put into combining suitable statements together in forced-choice blocks, the format can deliver significant advantages. By removing the peculiar properties of ipsative data, the author hopes that the theoretical barriers against the use of the forced-choice format will start to fall.

References

- Ackerman, T.A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. McArdle. (Eds.). *Contemporary Psychometrics* (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum.
- Ansari, A., Jedidi, K. & Dube, L. (2002). Heterogeneous factor analysis models: a Bayesian approach. *Psychometrika*, *67* (1), 49-78
- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.
- Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational Psychology*, *69*, 25-39.
- Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*, 263-272.
- Bartram, D., Brown, A., Fleck, S., Inceoglu, I. & Ward, K. (2006). *OPQ32 Technical Manual*. Surrey, UK. SHL Group.
- Bock, R.D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Brown, A. (2009). Doing less but getting more: Improving forced-choice measures with IRT. *Paper presented at the 24th annual conference of the Society for Industrial and Organizational Psychology*, New Orleans, LA.
- Brown, A. & Bartram, D. (2009). *Development and psychometric properties of OPQ32r*. Surrey, UK. SHL Group.
- Brown, A. & Maydeu-Olivares, A. (2009). Improving forced-choice tests with Item Response Theory. *Paper presented at the International Meeting of the Psychometric Society*, 21-24 July, Cambridge, England.

- Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika*, *30*, 99-121.
- Chan, W. & Bentler, P.M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63*, 369-399.
- Chernyshenko, O., Stark, S., Prewett, M., Gray, A., Stilson, F. & Tuttle, M. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: empirical comparisons with other formats. *Human Performance*, *22*, 105-127.
- Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 55-77.
- Christiansen, N, Burns, G., & Montgomery, G. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance*, *18*, 267-307.
- Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, *14*.
- Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational Psychology*, *69*, 41-47.
- Coombs, C.H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57*, 145-158.
- Costa, P.T. & McCrae, R.R. (1992). *NEO-PI-R Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*, *29*, 115-126.
- Du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincoln-wood, IL: Scientific Software International.

- Embretson, S. & Reise, S. (2000). *Item Response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 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.
- Friedman, H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, *9*, 114-123.
- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26-42.
- Gordon, L.V. (1984). *Survey of personal values: Examiner's manual*. Chicago, IL: Science Research Associates.
- Griffith, R. & McDaniel, M. (2006). The nature of deception and applicant faking behavior. In Griffith, R. & Peterson, M. (Eds). *A Closer Examination of Applicant Faking Behavior*. Greenwich, CT: Information Age Publishing.
- Haaland, D., & Christiansen, N. (1998). Departures from linearity in the relationship between applicant personality test scores and performance as evidence of response distortion. *Paper presented at the 22nd Annual IPMAAC Conference*, Chicago, IL.
- Heggestad, E., Morrison, M., Reeve, C., & McCloy, R. (2006). Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, *91*, 9–24.
- Hernández, A., Espejo, B. & González-Romá, V. (2006). The functioning of central categories Middle Level and Sometimes in graded response scales: Does the label matter? *Psicothema*, *18*, 300-306.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167-184.

- Hogan, R. (1983). A socioanalytic theory of personality. In M.M. Page (Ed.), *Nebraska Symposium on Motivation* (pp. 336-355). Lincoln: University of Nebraska Press.
- International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences* (<http://ipip.ori.org/>). Internet Web Site.
- Jackson, D., Wroblewski, V., & Ashton, M. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? *Human Performance*, *13*, 371–388.
- Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriousness and spuriousness: The use of ipsative personality tests. *Journal of Occupational Psychology*, *61*, 153-162.
- Karpatschof, B., & Elkjaer, H. K. (2000). *Yet the bumblebee flies: The reliability of ipsative scores examined by empirical data and a simulation study*. Department of Psychology, University of Copenhagen: Research Report no. 1.
- Kolb, A. & Kolb, D. (2005). *The Kolb Learning Style Inventory—Version 3.1. Technical Specifications*. HayGroup, Boston.
- Martin, B. A., Bowen C.C., & Hunt, S. T. (2001). How effective are people at faking on personality questionnaires? *Personality and Individual Differences*, *32*, 247-256.
- Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64*, 325-340.
- Maydeu-Olivares, A. & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*, 285-304.
- Maydeu-Olivares, A. & Brown, A. (2010). IRT modeling of paired-comparison and ranking data. *Manuscript submitted for publication*.

- Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*, 344-362.
- McCloy, R., Heggstad, E., Reeve, C. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222-248.
- McDonald, R.P. (1999). *Test theory. A unified approach*. Mahwah, NJ: Lawrence Erlbaum.
- Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organisational Psychology, 77*, 531-552.
- 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, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Newbury Park, CA: Sage.
- Muthén, B.O. (2006). IRT in Mplus. Retrieved from www.statmodel.com
- Muthén, L.K. & Muthén, B.O. (1998-2007). *Mplus User's guide. Fifth edition*. Los Angeles, CA: Muthén & Muthén.
- Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.
- Samejima, F. (1997). Graded response model. In W.J. van der Linden and R. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.
- SHL. (1997). *Customer Contact: Manual and user's guide*. Surrey, UK. SHL Group.
- Stark, S., Chernyshenko, O. & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different

- dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement*, 29, 184-203.
- Stark, S., Chernyshenko, O., Drasgow, F. & Williams, B. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91, 25-39.
- Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology*, 73, 749-751.
- Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, 79, 281-299.
- Thurstone, L.L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology*, 14, 187-201.
- Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, 19, 175-199.
- 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.
- Zickar, M. & Gibby, R. (2006). A history of faking and socially desirable responding on personality tests. In Griffith, R. & Peterson, M. (Eds). *A closer examination of applicant faking behavior*. Greenwich, CT: Information Age Publishing.

Appendix A: Violation of Alpha’s consistent coding assumption in MFC questionnaires

Reliabilities as measured by alpha would be depressed in a classically scored forced-choice questionnaire (i.e. ipsative) with a large number of dimensions. This is because classical scoring procedures rely on the assumption that higher item rankings correspond to higher true scores on the traits. To achieve high internal consistency, the respondent should prefer (or give top rank order to) an item from a scale on which he/she has the highest true score; this response pattern should be consistently observed across all blocks. If every block of statements involved comparisons between the same scales, such consistency in response would be possible. However, in a questionnaire with many dimensions, the number of all possible comparisons between scales becomes very large (for 30 scales it is $30 \cdot 29/2 = 435$). With such a large number of comparisons to perform, there cannot be more than one or two occasions when items from any given pair of scales “meet” in the same block. Therefore, items from a given dimension are compared to items from different dimensions in every new block of statements.

Let us imagine that an individual’s true trait scores are ordered as follows:

$$\dots < \textit{trait A} < \textit{trait B} < \textit{trait C} < \textit{trait D} < \textit{trait E} < \textit{trait F} < \textit{trait G} < \dots$$

Then in a block of four statements, including items from the first four traits (trait A – trait D), the respondent would select an item from ***trait D*** as “most like me”, because his/her true score on this trait is the highest. However, in a block including the last four traits (trait D – trait G) the respondent would rate an item from ***trait D*** as “least like me”. The same trait will receive the highest rank (maximum points) in one block, and the lowest in another. This response, completely consistent with the true scores, will appear to be inconsistent from the rank-ordering perspective. This apparent “inconsistency” is in fact implicit to the current scoring methodology of multidimensional forced-choice items.

Appendix B: *Mplus* syntax for the Thurstonian second-order formulation of the example model

```
TITLE:   Example forced-choice questionnaire with 3 triplets measuring 3 traits (This is the
second-order Thurstonian factor model depicted in Figure 1)
DATA:   FILE IS ExampleTest.dat;
VARIABLE: NAMES ARE i1i2 i1i3 i2i3 i4i5 i4i6 i5i6 i7i8 i7i9 i8i9;
          USEVARIABLES ARE ALL;
          CATEGORICAL ARE ALL;
ANALYSIS:
          ESTIMATOR IS wlsm;
          PARAMETERIZATION IS theta;
MODEL:
!these are the utilities (first-order factors)
!block 1
      u1 BY i1i2@1 i1i3@1;
      u2 BY i1i2@-1 i2i3@1;
      u3 BY i1i3@-1 i2i3@-1;
!block 2
      u4 BY i4i5@1 i4i6@1;
      u5 BY i4i5@-1 i5i6@1;
      u6 BY i4i6@-1 i5i6@-1;
!block 3
      u7 BY i7i8@1 i7i9@1;
      u8 BY i7i8@-1 i8i9@1;
      u9 BY i7i9@-1 i8i9@-1;
! binary outcomes are measured without error (this is ranking)
      i1i2-i8i9@0;

! these are the latent traits (second-order factors)
      Trait1 BY u1* u4 u7;
      Trait2 BY u2* u5 u8;
      Trait3 BY u3* u6 u9;
! variances for all traits are set to 1
```

```
Trait1-Trait3@1;  
! traits are freely correlated  
Trait1 WITH Trait2* Trait3*;  
Trait2 WITH Trait3*;  
!fixing unique variances of one utility per block to identify the model  
u3@.5;  
u6@.5;  
u9@.5;  
! trait scores for individuals cannot be estimated due to zero error variances
```

Appendix C: *Mplus* syntax for the Thurstonian IRT formulation of the example model

```
TITLE: Example forced-choice questionnaire with 3 triplets measuring 3 traits (The model
is depicted in Figure 2)
DATA: FILE IS ExampleTest.dat;
VARIABLE: NAMES ARE i1i2 i1i3 i2i3 i4i5 i4i6 i5i6 i7i8 i7i9 i8i9;
          USEVARIABLES ARE ALL;
          CATEGORICAL ARE ALL;
ANALYSIS:
          ESTIMATOR IS wlsm;
          PARAMETERIZATION IS theta;
MODEL: ! latent traits are indicated by binary outcomes directly
Trait1 BY i1i2*1 i1i3*1 (11)
          i4i5*1 i4i6*1 (14)
          i7i8*1 i7i9*1 (17);
Trait2 BY i1i2*-1 (12_m)
          i2i3*1 (12)
          i4i5*-1 (15_m)
          i5i6*1 (15)
          i7i8*-1 (18_m)
          i8i9*1 (18);
Trait3 BY i1i3*-1 i2i3*-1 (13_m)
          i4i6*-1 i5i6*-1 (16_m)
          i7i9*-1 i8i9*-1 (19_m);
! variances for all traits are set to 1
          Trait1-Trait3@1;
! traits are freely correlated
          Trait1 WITH Trait2* Trait3*;
          Trait2 WITH Trait3*;
! pairwise errors are free; parameters are declared here to impose constraints later
          i1i2*1 (e1e2);
          i1i3*1 (e1e3);
          i2i3*1 (e2e3);
```

```
i4i5*1 (e4e5);
i4i6*1 (e4e6);
i5i6*1 (e5e6);
i7i8*1 (e7e8);
i7i9*1 (e7e9);
i8i9*1 (e8e9);
```

! errors related to the same utility are correlated, some are with minus sign

```
i1i2 WITH i1i3*.5 (e1);
i1i2 WITH i2i3*-.5 (e2_m);
i1i3 WITH i2i3*.5 (e3);
i4i5 WITH i4i6*.5 (e4);
i4i5 WITH i5i6*-.5 (e5_m);
i4i6 WITH i5i6*.5 (e6);
i7i8 WITH i7i9*.5 (e7);
i7i8 WITH i8i9*-.5 (e8_m);
i7i9 WITH i8i9*.5 (e9);
```

MODEL CONSTRAINT:

!loadings relating to the same item are equal in absolute value

```
l2=-l2_m; l5=-l5_m; l8=-l8_m;
```

! errors of pairs are equal to sum of 2 utility errors

```
e1e2=e1-e2_m;
e1e3=e1+e3;
e2e3= -e2_m+e3;
e4e5=e4- e5_m;
e4e6=e4+e6;
e5e6= -e5_m+e6;
e7e8=e7- e8_m;
e7e9=e7+e9;
e8e9= -e8_m+e9;
```

!fixing unique variances of one utility per block to identify the model

```
e3=.5; e6=.5; e9=.5;
```

SAVEDATA: ! trait scores for individuals are estimated and saved in a file

FILE IS ExampleTestResults.dat;

SAVE=FSCORES;

Appendix D: Designs involving blocks of 2 items (pairs)

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

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

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ \hline 0 & 0 & 1 & -1 & & 0 & 0 \\ \hline \vdots & & & & \ddots & & \\ \hline 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}, \quad \mathbf{\Lambda}' = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & \lambda_6 & 0 & 0 & \lambda_9 & 0 & 0 & \lambda_{12} \\ 0 & \lambda_2 & 0 & \lambda_4 & 0 & 0 & 0 & \lambda_8 & 0 & \lambda_{10} & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & \lambda_5 & 0 & \lambda_7 & 0 & 0 & 0 & \lambda_{11} & 0 \end{pmatrix}.$$

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

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

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

Appendix E: Posterior MAP information for a trait in MFC questionnaire

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

$$\mathcal{I}_P^a(\boldsymbol{\eta}) = \mathcal{I}^a(\boldsymbol{\eta}) - \frac{\partial^2 \ln(\phi(\boldsymbol{\eta}))}{\partial \eta_a^2}. \quad (36)$$

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

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

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

$$\frac{\partial \ln(\phi(\boldsymbol{\eta}))}{\partial \eta_a} = \frac{\partial}{\partial \eta_a} \left(-\frac{1}{2} \boldsymbol{\eta}' \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}\right) = -\frac{1}{2} \cdot \frac{\partial}{\partial \eta_a} (\boldsymbol{\eta}' \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}). \quad (37)$$

Let $\boldsymbol{\varpi}_i^j$ be an element of the inverted trait covariance matrix $\boldsymbol{\Phi}^{-1}$ in i^{th} row and j^{th} column; and the matrix form can be expanded as follows:

$$\begin{aligned} \frac{\partial}{\partial \eta_a} (\boldsymbol{\eta}' \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}) &= \frac{\partial}{\partial \eta_a} \left(\eta_1 \left(\sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^1 \right) + \dots + \eta_a \left(\sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^a \right) + \dots + \eta_d \left(\sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^d \right) \right) = \\ &= \eta_1 \boldsymbol{\varpi}_a^1 + \dots + \sum_{j=1, j \neq a}^d \eta_j \boldsymbol{\varpi}_j^a + 2\eta_a \boldsymbol{\varpi}_a^a + \dots + \eta_d \boldsymbol{\varpi}_a^d = 2 \sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^a \end{aligned}$$

Now, it follows from (37) and the expansion above that the second derivative by η_a of the logarithm of the density function is

$$\frac{\partial^2 \ln(\phi(\boldsymbol{\eta}))}{\partial \eta_a^2} = -\frac{1}{2} \cdot \frac{\partial}{\partial \eta_a} \left(2 \sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^a \right) = -\frac{1}{2} \cdot 2 \frac{\partial}{\partial \eta_a} \left(\sum_{j=1}^d \eta_j \boldsymbol{\varpi}_j^a \right) = -\boldsymbol{\varpi}_a^a, \quad (38)$$

and the above expression substituted into Equation (36), arriving at Equation (27).

Appendix F: The Big Five questionnaire with IPIP items

- 1 I am relaxed most of the time
- 2 I start conversations
- 3 I catch on to things quickly

- 4 I show my gratitude
- 5 I do things according to a plan
- 6 I am not easily bothered by things

- 7 I have difficulty understanding abstract ideas
- 8 I am the life of the party
- 9 I inquire about others' well-being

- 10 I like order
- 11 I am good at many things
- 12 I get upset easily

- 13 I sympathise with others' feelings
- 14 I worry about things
- 15 I feel at ease with people

- 16 I love to think up new ways of doing things
- 17 I am quiet around strangers
- 18 I often forget to put things back in their proper place

- 19 I keep in the background
- 20 I have frequent mood swings
- 21 I feel others' emotions

- 22 I follow a schedule
- 23 I am full of ideas
- 24 I don't talk a lot

- 25 I love to read challenging material
- 26 I get overwhelmed by emotions
- 27 I am not interested in other people's problems

- 28 I waste my time
- 29 I get irritated easily

- 30 I talk to a lot of different people at parties
-
- 31 I feel comfortable around people
- 32 I love to help others
- 33 I get jobs done right away
-
- 34 I seldom feel blue
- 35 I know how to comfort others
- 36 I avoid difficult reading material
-
- 37 I find it difficult to approach others
- 38 I panic easily
- 39 I neglect my duties
-
- 40 I make time for others
- 41 I am always prepared
- 42 I can handle a lot of information
-
- 43 I make friends easily
- 44 I have excellent ideas
- 45 I get stressed out easily
-
- 46 I make plans and stick to them
- 47 I rarely get irritated
- 48 I am indifferent to the feelings of others
-
- 49 I leave a mess in my room
- 50 I make people feel at ease
- 51 I am quick to understand things
-
- 52 I feel little concern for others
- 53 I don't mind being the centre of attention
- 54 I lack imagination
-
- 55 I have difficulty imagining things
- 56 I like to tidy up
- 57 I often feel blue
-
- 58 I love order and regularity
- 59 I am not really interested in others
- 60 I am skilled in handling social situations
-

