



**BERGISCHE  
UNIVERSITÄT  
WUPPERTAL**

## **Advances in Thurstonian Forced-Choice Modeling**

Inaugural-Dissertation im Fach Psychologie  
zur Erlangung des Doktorgrades  
Doktor der Naturwissenschaften (Dr. rer. nat.)

durch die Fakultät für Human- und Sozialwissenschaften  
der Bergischen Universität Wuppertal

vorgelegt von  
Markus Thomas Jansen  
aus Köln

*Prüfungskommission:*

Prof. Dr. Anna Baumert (Vorsitz)  
Prof. Dr. Ralf Schulze (erster Gutachter)  
Prof. Dr. Heinz Holling (zweiter Gutachter)  
Prof. Dr. Philipp Doebler (dritter Gutachter)

Termin der Disputation: 18.09.2023

---

## **Acknowledgments**

There are a number of people that I would like to thank for their support to make this work possible. On the scientific side, I thank Prof. Dr. Ralf Schulze for his support, giving me the opportunity and freedom to work on the projects and especially for his advice and comments that improved this work considerably. Also, I would like to thank my colleagues, notably Maïke Pisters and Susan Hellwig, who helped me probably more than they know, just by listening and by the discussions of some critical points. On the offside of science, I also thank my friends for their social and emotional support, including parts of the dissertation, but also on life in general. Most grateful I am to my wife Roja, who supported me on this journey in uncountable ways, but especially by being always enthusiastic about my work, just because I am enthusiastic about it.

March 2023

Markus Thomas Jansen

---

## Contents

Abstract.....	iv
1. Introduction.....	1
1.1 Trait Assessment, Response Scales, Response Bias and Faking .....	1
1.2 Thurstonian Factor Analytic and IRT Forced-Choice Approaches .....	7
1.3 Interim .....	15
1.4 The (Multidimensional) Block Design.....	17
1.5 Summary .....	21
1.6 Overview of Manuscripts .....	24
2. Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues .....	25
2.1 Summary .....	25
2.1.1 Review about Thurstonian Forced-Choice Models.....	25
2.1.2 Problems in Thurstonian Forced-Choice Modeling.....	25
2.1.3 Discussion and Conclusion .....	26
2.2 Manuscript.....	27
2.2.1 Abstract.....	28
2.2.2 Linear Factor Analytic Thurstonian Forced-Choice Models: Current Status and Issues.....	29
2.2.3 Binary Coding of Forced-Choice Responses .....	34
2.2.4 Thurstonian Models .....	36
2.2.5 Problems with Thurstonian Forced-Choice Models .....	53
2.2.6 Discussion .....	68
2.2.7 References.....	71
3. Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data .....	77
3.1 Summary .....	77
3.1.1 Thurstonian Models and Non-Comparability Between Independent Blocks.....	77
3.1.2 The Thurstonian Linked Block Design and Simulation Study .....	77
3.1.3 Results.....	78
3.2 Manuscript.....	79
3.2.1 Abstract.....	80
3.2.2 The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data .....	81
3.2.3 Binary Coding of Responses of Paired Comparisons and Rankings .....	85
3.2.4 Thurstonian Models .....	87

---

3.2.5 The Thurstonian Linked Block Design.....	97
3.2.6 Simulation Studies .....	104
3.2.7 Overall Discussion .....	121
3.2.8 References.....	127
3.2.9 Appendix A.....	130
4. Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying.....	136
4.1 Summary .....	136
4.1.1 The Information Obtainable from Forced-Choice Responses.....	136
4.1.2 Simulation Study.....	137
4.1.3 Results.....	137
4.1.4 Recommendations for FC Test Construction.....	137
4.2 Manuscript.....	138
4.2.1 Abstract.....	139
4.2.2 Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying .....	140
4.2.3 Thurstonian Models .....	142
4.2.4 Simulation Studies .....	158
4.2.5 Overall Discussion .....	172
4.2.6 References.....	179
5. Thurmod: An R Package for Thurstonian Modeling.....	182
6. General Discussion .....	183
References.....	193
List of Figures.....	204
List of Tables .....	206
Nomenclature.....	207
CRedit Authors Statement.....	209
Statement of Originality.....	210

---

## Abstract

The use of self-reports in the form of rating scales is common in many social sciences. Many traits of interest are accessible only by interoception, for which rating scales are suited. However, there are many downsides to the use and analysis of rating scales, including several response biases and faking, especially in high-stakes situations. To overcome these potential disadvantages, the forced-choice (FC) format has been proposed repeatedly, as it eliminates many response biases. Additionally, the FC format makes it more challenging to fake responses, particularly when the statements between which a respondent has to choose are similar in terms of social desirability. However, FC data is inherently ipsative, which leads to the incomparability of scores from different respondents. More recently, some Thurstonian models, notably the Thurstonian multidimensional forced-choice (MFC) format and an IRT model, were proposed to overcome the ipsativity, by using multidimensional blocks of statements to be ranked.

The current work investigates the Thurstonian models and the MFC format thoroughly, coming to the conclusion that the current use of the Thurstonian FC models is ill-suited to accurately estimate item or person parameters. While the Thurstonian models themselves are mathematically sound, they are hardly usable in many real-world applications. The FC blocks used in the MFC format greatly enhance usability, but lose some of the mathematical precision and enable faking of responses. Additionally, the Thurstonian models could be used for both, item scaling and person scaling, however, the MFC format prevents good item scaling.

To overcome these shortcomings, an adaptation of the block design is proposed by linking all FC blocks. This procedure allows for the estimation of item parameters on the same scale instead of different scales per block and it is shown that parameter estimation is accurate under practical circumstances. Furthermore, the use and usefulness of mixed-keyed blocks, that are blocks including positively and negatively keyed items, are studied and discussed. Beyond

---

that, practical guidelines for the construction of FC tests are given under consideration of the current advances in Thurstonian FC modeling.

In summary, the current work extends the discussion of Thurstonian FC modeling, proposes a solution to some shortcomings of previous applications that is mathematically fitting and applicable in real-world situations. This includes guidelines for FC test construction. Overall, the use of the FC method for the reduction or elimination of several response biases and faking should enhance the reliability and validity of the assessed data, which is crucial for applications and research in psychology, and social sciences in general.

## **1. Introduction**

### **1.1 Trait Assessment, Response Scales, Response Bias and Faking**

The assessment of information about individuals is one of the most prominent tasks in various fields, including politics, healthcare, education, business, and research. While some information, for example somewhat more objective data such as demographics, is readily available through observation or self-report, there is often interest in information that is not directly observable or of a subjective nature. Whenever the interest is about people's opinions, interests or personality traits, there is almost no way around asking people about these directly. In these and similar cases, the most prominent strategy to gather information about people is the use of self-report, often in the form of response scales. A typical response scale is given by a number of verbal categories that describe the extent to which a person agrees or disagrees with a statement (e.g., strong, medium, slight agreement or disagreement). Nowadays, it appears that response scales are extensively used, as they are very easy to implement and can be made available to many people over internet resources with low effort and in a small amount of time, evidenced by the large amount of (online) self-report studies. While non-laboratory assessments go with their own category of caveats, despite the wide use, response scales and the results and interpretations about an individual's traits based on response scales can be troublesome. Therefore, a considerable proportion of research is devoted to the categorization and elimination of problematic features of response scales. Additionally, one strand of research is about the adaptation of response formats in general, to find suitable assessment strategies for the respective situation.

The subjective nature of self-reports in form of response scales gives way to a multitude of response biases, which can generally be described as systematic influences on the responses that are not correct or precise given an underlying theory (Paulhus, 1991). The reasons for such response biases can be diverse. Examples include the desire to conform to

the perceived examiners' expectation (e.g., Rosenthal-effect; Rosenthal & Fode, 1963), a tendency to generally strongly agree or disagree with statements (extreme responding), a tendency to agree with statements (acquiescence bias; Cronbach, 1946) or other forms of showing preference for a specific category (Henninger & Meiser, 2020). Another reason for biased responses can be misinterpretation or a lack of insight in oneself (Paulhus, 1986) or different interpretations of the given scale by different respondents (e.g., King et al., 2004; King & Wand, 2006). While these systematic and unsystematic influences need to be limited for accurate measurement, in some cases it is important to take the deliberate distortion by individuals into account. The most prominent example is the social desirability bias, meaning agreeing with statements that are socially accepted and disagreeing with socially not accepted statements (Edwards, 1953; Paulhus, 2002; Ziegler et al., 2012). The social desirability bias has two interlinked components: the desirability of the statement or stimulus in general and the distorted response of the person (Edwards, 1953). The agreement or disagreement with a statement can result in positive or negative judgement by others. Statements on which a person's agreement results in more positive judgement by others are socially more desirable (Edwards, 1953). On the other hand, responses that are distorted in the direction of agreement with a desirable statement contribute to socially desirable responding (SDR; Edwards, 1953; Paulhus, 2002). In extreme cases, the social desirability bias can be seen as faking, the deliberate deception of others, to alter self-presentation in a way that helps to achieve personal goals (Ziegler et al., 2012). While all response biases reduce the reliability and validity of tests, for example by increased scores for some traits (Birkeland et al., 2006; Viswesvaran & Ones, 1999) or the change of correlations between trait scores (Moors, 2012), socially desirable responding and faking are perceived as more severe due to their intentional nature (Ziegler et al., 2012). However, as faking is linked to a specific goal, it is generally



acknowledged to be prevalent only in specific situations that are important for the individual (Ziegler et al., 2012).

Faking and its prevention are relevant and common topics in psychological research and application. Psychological tests are commonly used for selection, diagnosis, and categorization. If the situation is of importance to the test taker, they are likely to fake at least some responses. Often, it is differentiated between faking good and faking bad (Ziegler et al., 2012). Faking bad describes a distortion of responses and behaviors in a less favorable way, or to be perceived as less able. Faking bad is most common in situations where a person has an external gain such as compensation, payout, early retirement, or to escape punishment (e.g., Ziegler et al., 2012). Classic examples come from clinical or neuropsychological assessment, where individuals sometimes fake psychological or cognitive disorders (e.g., Aronoff et al., 2007; Merten, 2013). On the other hand, faking good describes the distortion of responses and behaviors in a more favorable direction, or to be perceived as more able in a certain trait, which often coincides with more desirable traits (Ziegler et al., 2012). Typical examples are situations such as job applications, test situations, dating, and other situations in which an individual is evaluated based on their intelligence or personality traits. Faking bad can occur in almost all situations, including ability measurement, for example, intelligence assessment, and trait measurement, for example, personality assessment. Faking good is more prevalent in situations such as personality assessment; in (classic) intelligence assessment, test takers are instructed to perform as good as possible anyway. The aforementioned examples from various fields of psychology show the importance of procedures and tests to reduce fakeability.

Research has produced strong evidence that many traits are easily fakeable. For example, in personality psychology, two influential meta-analyses show that the big five factors of personality, especially neuroticism and conscientiousness, are prone to faked

responses (Birkeland et al., 2006; Viswesvaran & Ones, 1999). Any interpretation and decision based on responses that are exposed to response biases and faking, reduce the corresponding validity (see for example Douglas et al., 1996; Lanyon, 1993; Velicer & Weiner, 1975). Therefore, the reduction and elimination of the influence of bias and faking is an important goal of psychological research.

For each of the potential response biases, especially for faking and SDR, attempts were made to produce data that are free from unwanted responses. For faking and SDR, these can be categorized into demand reduction strategies, covariance techniques, and rational techniques (Paulhus & Vazire, 2007). The idea behind demand reduction is to assure anonymity and confidentiality to the respondents. To further support compliance, respondents are offered feedback on their own test results. It was shown that for SDR, demand reduction techniques can reduce desirable responding (Paulhus & Vazire, 2007). However, considering faking, demand reduction seems not to be a promising strategy, as the demand reduction would only be achieved by changing the goal of the assessment, which could render the assessment useless. Covariance techniques focus on statistical control of biases; however, it is argued, that this way also valid variance is controlled for, reducing the amount of information about the traits of interest (Paulhus & Vazire, 2007, Viswesvaran & Ones, 1999). Rational techniques are broader and focus on the prevention of bias and faking in the first place. Some rational techniques include changing the wording of items to appear in a more neutral way or the presentation of only neutrally desired items (Bäckström et al., 2009; Bäckström et al., 2011). These techniques, however, have delicate drawbacks: they can be implemented only with high effort and, more importantly, the universe of items from which can be chosen is reduced to neutrally desirable items. However, if highly desirable or undesirable traits are of interest, items with neutral desirability are hardly suitable for test construction. An example for such undesirable traits can be seen in so-called dark traits (e.g., Moshagen et al., 2018).

A more promising approach, which is central to the present thesis and could be named as a rational technique, is the idea of balancing the desirability in a forced-choice (FC) task. The idea of a FC task is to change the response format entirely, instead of giving the degree of (dis)agreement to a statement on a scale, respondents are asked to rank several statements with respect to the fit to the own person or a criterion (e. g., Christiansen et al., 2005, Dunnette et al., 1962, Edwards, 1953). An easy example for a FC task would be the comparison between statements “I am orderly” and “I am punctual”. A person must choose between the two statements, therefore, they cannot (dis)agree to both statements. If both statements are comparable considering the desirability, the choice should reflect only the traits and behaviors described by the statements. The influence of desirability and especially faking should then be reduced or eliminated. By balancing the desirability of the statements within a FC task, the item universe is not restricted, as equally (un)desirable statements can be matched and used. Also, with the FC method many response biases such as the acquiescence bias and extreme responding are eliminated by design, there is no response scale in the first place.

In spite of the many advantages that the FC method offers, there is one overshadowing disadvantage: FC responses are by design ipsative, making it difficult or even impossible to compare between respondents (Baron, 1996). The main goal behind assessment strategies often is to compare respondents with respect to certain traits of interest. Responses and scores that are comparable *across* respondents are called normative. Responses within a FC task are, however, only comparable *within* the respondent, such responses are called ipsative. For example, one person may describe themselves as more orderly than punctual, while another person chooses to be more punctual than orderly. The comparison between these people is not possible, as the second person could be unpunctual in general, but even less orderly, and the first person could be less punctual than orderly, but still punctual in

general. This between-person information is not directly retrievable. No matter if comparisons are between statements of one or multiple traits, the rank scores per respondent are equal across all respondents (e.g., Clemans, 1966). Therefore, any correlational or factor analytical analyses (which is based on correlation) is distorted (Clemans, 1966; Hicks, 1970) and classical psychometric analysis is impossible (Baron, 1996). The ipsativity in trait scores resulted in limited usefulness of the FC method as an assessment strategy and a solution for response biases and faking.

However, some advances in the past twenty years have revived the discussion about the FC method, as they include the possibility to estimate normative scores, and thus, compare scores of individuals (see Maydeu-Olivares & Brown, 2010; Stark et al., 2005). These advances can roughly be separated by the underlying theory of choice behavior, that could be the ideal-point model or Coombs's unfolding models (Coombs, 1960; McCloy et al., 2005; Thurstone, 1928; Stark et al., 2005) or the dominance response model (e.g., Brown, 2016; Thurstone, 1927). For the ideal point model, it is argued that some statements, for example "Firearms should not belong in private hands", have the highest psychological value (utility) on a scale, is the highest for a person with the exact level of the trait of Militarism (Thurstone, 1928). Statements that represent higher or lower attitudes to the trait should have a lower utility. The point of maximum utility between a person and the attribute continuum is called the ideal point (Coombs, 1960). In contrast, for the dominance response process, it is assumed that each item represents only one trait, and the higher a respondent scores on the trait, the stronger is the degree of agreement to the item, which in turn contributes to a higher utility of the statement. A typical example would be "I keep my desk orderly". The higher a respondent scores on conscientiousness, the more they should agree to the statement. While both processes are important and interesting, the focus of the current work is on the dominance response process. For FC in the context of the dominance response process, the

Thurstonian FC models are prominent. Notably, Brown and Maydeu-Olivares (2011) introduced the multidimensional FC (MFC) block format and before that (Maydeu-Olivares & Brown, 2010) a corresponding model based on the item response theory (IRT). A MFC test has several blocks (for example three statements per block, that is a triplet), and a respondent has to rank the statements within a block, based on which statement fits most as a descriptor of the own person. From the IRT perspective, the estimation of normative scores is possible (Maydeu-Olivares & Brown, 2010). While these advances and the discussion about the corresponding methods are fruitful, there are also some highly problematic features that, up to this point, have not been discussed, or where the discussion led to misconceptions for constructing FC assessments.

The present work will focus on the FC method based on the dominance response process. First, the overall idea behind the FC method and Thurstonian factor analytic and IRT approaches are presented. Then, three studies are presented. The first includes a review of the methods and a discussion of shortcomings in Thurstonian FC modeling, some of which have not been discussed before. Additionally, the model assumptions are discussed and generalized. In the second study, the generalized assumptions are used to present an adaptation of the FC block design that solves many of the shortcomings discussed previously. A third study extends the discussion on how to construct a FC test and provides guidelines for the FC test construction based on the theory and simulation study results. Before concluding, some strengths and limitations are discussed and future directions are proposed.

### **1.2 Thurstonian Factor Analytic and IRT Forced-Choice Approaches**

The general idea of the Thurstonian FC method, first described by Thurstone (1927), is focused on item scaling. It posits, that any statement (in general: any stimulus) evokes a discriminative process that corresponds to the utility of that statement for the specific respondent. This utility value is denoted by  $t_i$  for item  $i$ . When two statements are compared

against each other, the utility values are compared, and the statement with the larger utility is chosen. Moreover, the evoked utilities are not constant, they vary with expected value  $\mu_i$  (vector  $\boldsymbol{\mu}_i$ ) and variance  $\sigma_i^2$  (covariance matrix  $\boldsymbol{\Sigma}_i$ ) for repeated presentation of the same statement and are assumed to be multivariate normal distributed (Maydeu-Olivares & Böckenholt, 2005; Thurstone, 1927)

$$\mathbf{t} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (1.1)$$

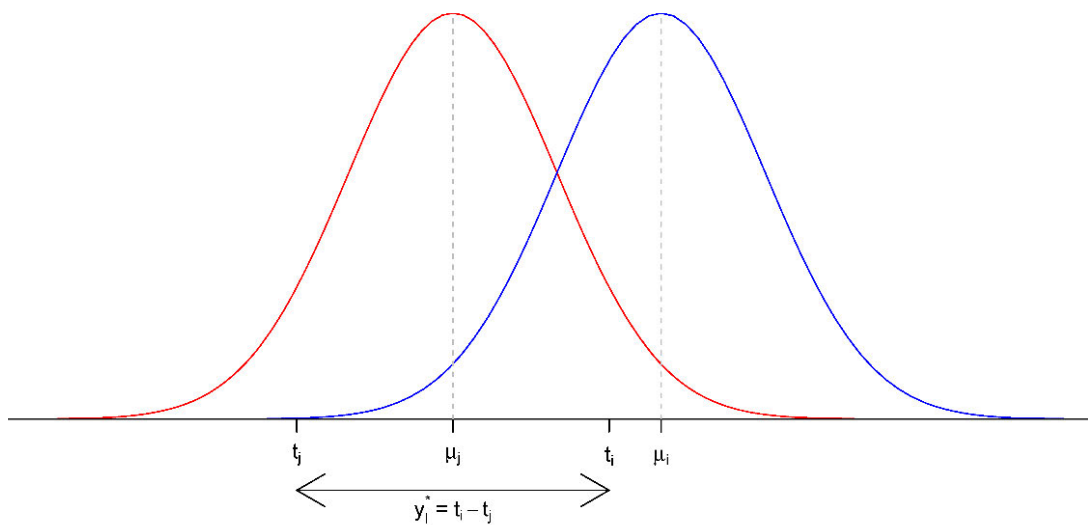
The observable response on a comparison  $l$  denoted by  $y_l$  is determined by the latent discriminative process  $y_l^* = t_i - t_j$  which is unobservable. If the difference between the utilities of the two statements being compared is positive, statement  $i$  is chosen and the comparison is coded with 1, otherwise it is coded with 0. That is

$$y_l = \begin{cases} 1 & \text{if } y_l^* + e_l \geq 0 \\ 0 & \text{if } y_l^* + e_l < 0 \end{cases} \Leftrightarrow y_l = \begin{cases} 1 & \text{if } t_i + e_l \geq t_j \\ 0 & \text{if } t_i + e_l < t_j \end{cases}. \quad (1.2)$$

Figure 1.1 shows a graphical example of a comparison.

**Figure 1.1**

*The Discriminative Process Between Stimuli  $i$  (blue) and  $j$  (red).*

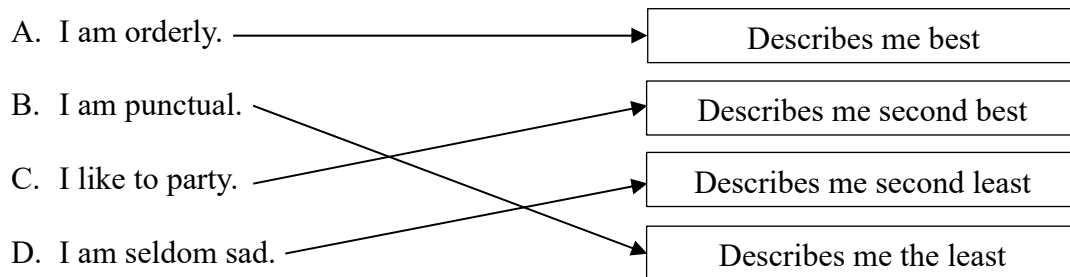


*Note.* This Figure is identical to Figure 2.2 from Study 1.

Now, if instead of two,  $n$  statements are to be compared, there are

$$\tilde{n} = \frac{n(n-1)}{2} \tag{1.3}$$

possible paired comparisons. A common assumption in Thurstonian FC modeling is transitivity, which states that if statements  $i, j$  and  $h$  are to be compared, and  $i$  is preferred over  $j$  and  $j$  is preferred over  $h$ , then  $i$  is preferred over  $h$ . This corresponds to a ranking task and to the assumption, that  $e_i = 0$ . If a block of  $k$  statements is to be ranked, then the number of derivable paired comparisons per block follows equation (1.3) by substituting  $n$  by  $k$  and  $\tilde{n}$  by  $\tilde{k}$ . The binary coding is the same as before. For what follows, a paired comparison task does not assume transitivity, while a ranking task does. For an example of a ranking task, assume four statements with the following instruction: Please rank the following statements in the order they are a fitting description of you:



The arrows represent the person’s ranking, indicating their order of preference for the items. In this case, the respondent’s order of the items is A, C, D, B, so A is preferred over C, D, and B. This results in  $y_{\{A,B\}} = 1, y_{\{A,C\}} = 1, y_{\{A,D\}} = 1$ . Similarly, statements C and D are both preferred over B, resulting in  $y_{\{B,C\}} = 0, y_{\{B,D\}} = 0$ , and C is preferred over D resulting in  $y_{\{C,D\}} = 1$ .

Based on these ideas, Maydeu-Olivares and Böckenholt (2005) derived an important linear factor analytic approach to estimate and test several of Thurstone’s cases with

respective assumptions (Thurstone, 1927). The  $\tilde{n}$  latent utility differences  $y_i^*$  can be rewritten into vector-matrix form for a clean presentation, that is

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{\tilde{n}} \end{pmatrix} = \begin{pmatrix} t_1 - t_2 + e_1 \\ t_1 - t_3 + e_2 \\ \vdots \\ t_{n-1} - t_n + e_{\tilde{n}} \end{pmatrix}. \quad (1.4)$$

Matrix  $\mathbf{A}$  is central, as it defines the paired comparisons that are included within the design.

For the previous example, it would be

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_{n6} \end{pmatrix} = \begin{pmatrix} t_1 - t_2 + e_1 \\ t_1 - t_3 + e_2 \\ t_1 - t_4 + e_3 \\ t_2 - t_3 + e_4 \\ t_2 - t_4 + e_5 \\ t_3 - t_4 + e_6 \end{pmatrix} \quad (1.5)$$

with a total of  $6 = 4 \times 3 / 2$  paired comparisons. The rank of  $\mathbf{A}$  is not full, but  $n - 1$ , which is the result of a location indeterminacy based on the relative nature of these comparisons. To solve this indeterminacy (and identify the later described models), one parameter  $t_i$  has to be fixed to an arbitrary value, often zero (Maydeu-Olivares & Böckenholt, 2005). Given the multivariate normal distribution of the utilities, the mean and covariance structure of the latent utilities can be described by a structural equation model with binary indicators as

$$\boldsymbol{\mu}_{\mathbf{y}^*} = \mathbf{A}\boldsymbol{\mu}_t, \text{ and } \boldsymbol{\Sigma}_{\mathbf{y}^*} = \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}' + \boldsymbol{\Omega}^2 \quad (1.6)$$

where  $\boldsymbol{\mu}_{\mathbf{y}^*}$  is the vector of means and  $\boldsymbol{\Sigma}_{\mathbf{y}^*}$  is the covariance matrix of the  $\tilde{n}$  latent

differences. Finally,  $\boldsymbol{\Omega}^2$  is the covariance matrix of the error terms, in ranking designs it is

$\boldsymbol{\Omega}^2 = \mathbf{0}$ . To estimate models of this type, the thresholds and tetrachoric correlations of the

variables must be estimated (Muthén, 1978). To identify the model, it is convenient to

standardize the latent differences to  $\mathbf{z}^* = \mathbf{D}(\mathbf{y}^* - \boldsymbol{\mu}_{\mathbf{y}^*})$  with  $\mathbf{D} = \text{diag}(\boldsymbol{\Sigma}_{\mathbf{y}^*})^{-1/2}$  (see Maydeu-



Olivares & Böckenholt, 2005; Muthén, 1978). The relationship between observed responses and standardized latent differences is one of the  $\tilde{n}$  thresholds  $\tau_l$

$$y_l = \begin{cases} 1 & \text{if } z_l^* \geq \tau_l \\ 0 & \text{if } z_l^* < \tau_l \end{cases} \quad (1.7)$$

where the vector of thresholds is  $\boldsymbol{\tau} = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}_t$  and the tetrachoric correlation matrix is

$$\mathbf{P}_{z^*} = \mathbf{D}(\boldsymbol{\Sigma}_{y^*})\mathbf{D} = \mathbf{D}(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}' + \boldsymbol{\Omega}^2)\mathbf{D} \quad (1.8)$$

(Maydeu-Olivares & Böckenholt, 2005; Muthén, 1978). Also, the intercepts  $\boldsymbol{\gamma}$  are constrained, to identify the latent utility means to be

$$\boldsymbol{\gamma} = -\mathbf{A}\boldsymbol{\mu}_t. \quad (1.9)$$

Simple Thurstonian models allow for the testing of Thurstone cases such as Case II where all parameters are free except for identification, Case III where utilities are uncorrelated, and Case V where, in addition, all variances of utilities are equal (Thurstone, 1927). For identification, one mean has to be fixed to solving the location indeterminacy. For the covariance matrix  $\boldsymbol{\Sigma}_t$  one variance must be set to unity for Case III and Case V, and in Case II, two variances to unity and all covariances of one item whose variance was fixed must be set to zero. All thresholds are fixed to zero in order to identify the utility means. A graphical representation of an example with four items is given in Figure 1.2.

While simple Thurstonian models are suitable for testing and estimating parameters for the Thurstone cases, often a latent structure of the items, and therefore their utilities, is assumed. One main goal is to compare respondents based on their scores on specific traits. Especially in psychological science, a latent structure of the items of interest is assumed, such as the big five in personality psychology. Imposing such a structure of  $m$  common factors or traits leads to a Thurstonian factor model (Maydeu-Olivares & Böckenholt, 2005),

$$\mathbf{t} = \boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (1.10)$$

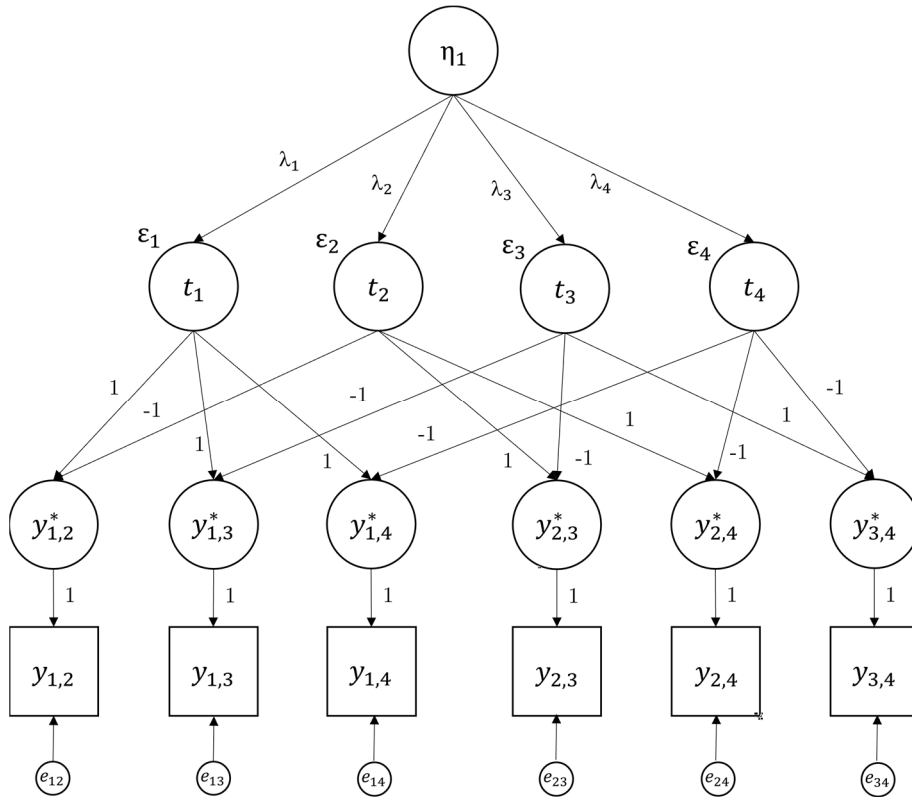
where  $\Lambda$  is the  $n \times m$  matrix of factor loadings of the latent utilities on the latent traits,  $\boldsymbol{\eta}$  is a  $m \times 1$  vector of the latent traits, and  $\boldsymbol{\varepsilon}$  is a  $n \times 1$  vector of unique factors (error term). More importantly, the mean and covariance structure changes to

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t, \text{ and } \boldsymbol{\Sigma}_{y^*} = \mathbf{A}(\mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{A}' + \boldsymbol{\Omega}^2 \quad (1.11)$$

where  $\boldsymbol{\Phi}$  is a  $m \times m$  covariance matrix of the factors and  $\boldsymbol{\Psi}^2$  is the covariance matrix of latent uncorrelated and unbiased error terms. The same identification constraints apply, as for simple Thurstonian models, in addition, all variances of the  $m$  traits need to be fixed to 1. For more technical details see Maydeu-Olivares and Böckenholt (2005), Maydeu-Olivares and Brown (2010) or the manuscripts within the present work. A graphical representation of the example with four items is given in Figure 1.2.

**Figure 1.2**

*Covariance Structure of a Thurstonian Factor Model for  $n = 4$  and  $m = 1$ .*



*Note.* This Figure is identical to Figure 2.4, 3.2 and 4.2 from studies 1, 2 and 3.

Given a paired comparison task, not assuming transitivity, the diagonal of the error covariance matrix is generally nonzero, which allows for the estimation of latent trait scores. These latent trait scores are perceived as normative, and comparisons between respondents are possible, as the latent traits are assumed to be multivariate normal. However, presenting participants with  $\tilde{n}$  paired comparisons is ponderous even for small item sets. For example, a small item set of 15 items, which can be considered as a short scale in many psychological fields, would result in presenting 105 paired comparisons. This is the main reason why a ranking block approach is often used, as it reduces the workload for respondents and the complexity of the design. However, the use of ranking designs comes with the disadvantage that  $\mathbf{\Omega}^2 = 0$ , for which latent traits scores cannot be estimated anymore, due to the lack of variation (Maydeu-Olivares & Brown, 2010).

To finally overcome the ipsative nature of scores and be able to estimate normative scores, a powerful reparameterization of the model in (1.11) was proposed (Maydeu-Olivares & Brown, 2010). If the latent trait scores are of interest, and the utilities are not, the thresholds are free to be estimated. Furthermore, the intercepts  $\boldsymbol{\gamma}$  need not to be constraint anymore. This allows for the mean and covariance structure to be

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}\mathbf{t} + \mathbf{e}, \quad \mathbf{t} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (1.12)$$

with  $\boldsymbol{\tau} = \mathbf{D}\boldsymbol{\gamma}$ , which is unconstrained as it is a rescaling by  $\mathbf{D}$ . By freeing the intercepts and thresholds, the aforementioned reparameterization is

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}(\mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = -\boldsymbol{\gamma} + \mathbf{A}\mathbf{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\boldsymbol{\gamma} + \check{\mathbf{\Lambda}}\boldsymbol{\eta} + \check{\boldsymbol{\varepsilon}} \quad (1.13)$$

with  $\check{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}$  and  $\text{cov}(\check{\boldsymbol{\varepsilon}}) = \check{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \mathbf{\Omega}^2$ , where  $\check{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda}$  is a  $\tilde{n} \times m$  matrix. For example, assuming only one single trait,  $m = 1$ , and  $n = 3$  it would be

$$\check{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_3 \\ \lambda_2 - \lambda_3 \end{pmatrix} \quad (1.14)$$

and assuming  $m = 3$  and  $n = 3$  it is

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \end{pmatrix} \quad (1.15)$$

and in both cases

$$\tilde{\Psi}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2 + \psi_3^2 + \omega_3^2 \end{pmatrix}. \quad (1.16)$$

Both models in (1.12) and (1.13) are equivalent, as (1.13) is simply a reparameterization, and therefore the same identification constraints apply. This model is also called the Thurstonian IRT model, as the reparameterization can be described as a normal ogive model

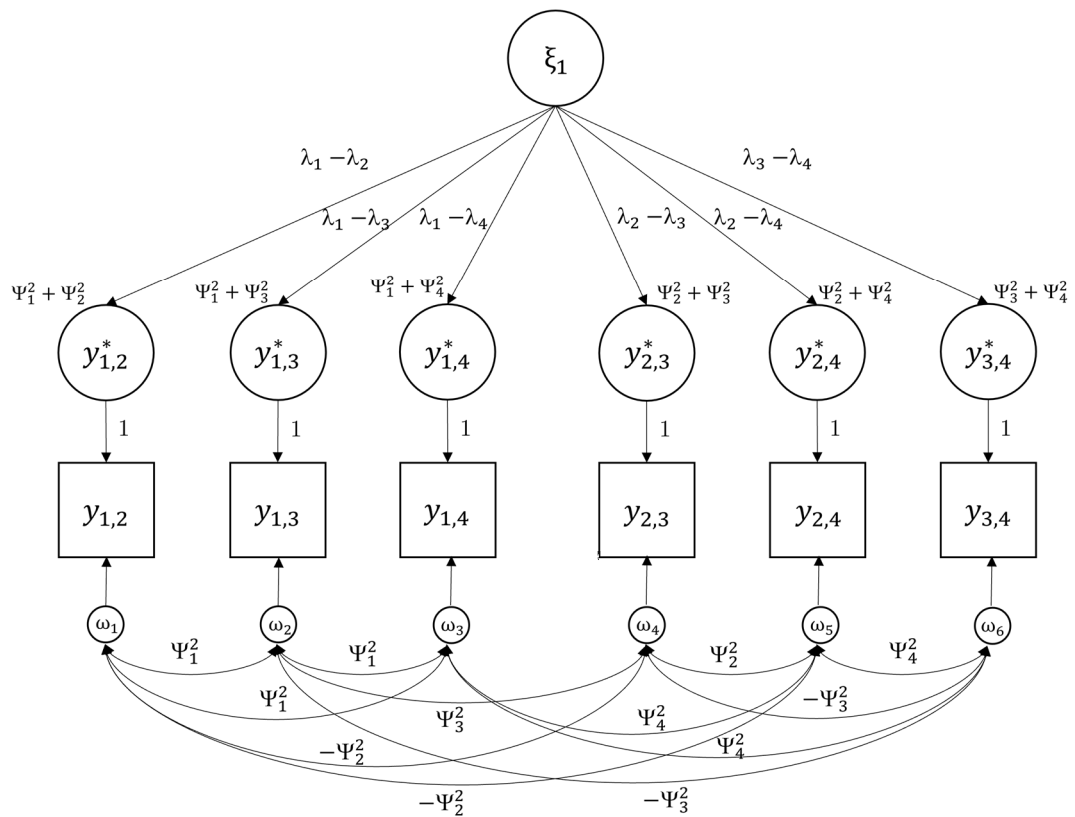
$$\Pr(y_l = 1 | \boldsymbol{\eta}) = \Phi \left( \frac{-\gamma_l + \tilde{\lambda}'_l \boldsymbol{\eta}}{\sqrt{\tilde{\psi}_l^2}} \right). \quad (1.17)$$

with structured  $\tilde{\lambda}'_l$  and  $\tilde{\psi}_l^2$ , and where the ICFs are not independent (for further details see Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). A graphical representation of the example with four items is given in Figure 1.3.

As for the IRT model, even in ranking tasks where  $\boldsymbol{\Omega}^2 = 0$ , the variances of the binary indicators are generally nonzero given that in (1.16) all  $\omega_l$  are zero, but  $\psi_i^2$  are generally not. Therefore, for the IRT model, latent trait scores can always be estimated (given fit and convergence of the model), which is done via maximum a posteriori (MAP) estimation (for further details see Brown & Maydeu-Olivares, 2011). For all Thurstonian models based on ranking data, there are redundancies in the thresholds and tetrachoric correlations. To get accurate model fit results, the degrees of freedom given by standard programs must be corrected by the number of redundancies.

**Figure 1.3**

*Covariance Structure of a Thurstonian IRT Model for  $n = 4$  and  $m = 1$ .*



*Note.* This Figure is identical to Figures 2.5, 3.3 and 4.3 from studies 1, 2 and 3.

### 1.3 Interim Summary and an Issue with Paired Comparison and Ranking Designs

The previous sections motivate the use of the FC method and present existing models for analyzing FC data. When collecting data of individuals, it is important to consider ways to reduce or, if possible, eliminate several response biases, which can distort data and interpretation of results. This is particularly important in high-stakes situations where respondents may respond in a deceptive way to achieve personal goals. The FC method, while resulting in ipsative data, also allows for the estimation of normative trait scores if a suitable Thurstonian model is applied. If transitivity is not assumed, then a paired comparison task allows for the estimation of latent trait scores with the Thurstonian factor model. However, most often the circumstances make a paired comparison task unsuitable and a ranking task is recommended, especially if the item set gets larger. For a ranking task, the

estimation of latent trait scores can be achieved with the Thurstonian IRT model. The main difference between the IRT and the factor model is that the focus shifts from item utilities (item parameter, item scaling) to the latent trait scores (person parameter). Both models are equivalent except for the constrained intercepts, which are needed to identify the means of the latent utilities.

In a world where respondents could and would respond to a near infinite number of items, where the computational resources are not limited, and a Thurstonian FC model could be estimated near instantly, the models of the preceding account would be perfectly suitable for any FC assessment. However, this is not the case, and there are three main limiting factors. First, the presentation of many paired comparisons is time-consuming, monotonous, and cognitively demanding for respondents (e. g. Sass et al., 2020). A typical psychological test for classic personality theories (e.g., big five or HEXACO) easily consist of 60 or more items, resulting in 1770 or more paired comparisons, of over and over repeating items. Second, ranking many items at the same time, while considerably faster, is still a very complex task for a respondent. Test motivation seems not to be affected by using four items to rank instead of three (Sass et al., 2020), however, this cannot easily be said about many larger blocks of  $k$  items. Also, the task of ranking many items at once is perceived as harder and more demanding (Sass et al., 2020). On the other hand, a previous study did ask respondents to rank 15 items at once, as well as presenting all 105 paired comparisons, and concluded that both options lead to comparable results (Jansen & Schulze, 2023a). That being said, a ranking task with a number of items that can be overlooked in a few saccades (for example 15 items) is, while cognitive demanding, quite possible. Third and notwithstanding the potential possibility to rank many items at once, there are the computational resources that are yet limited. For an item set of  $n$  items, and  $\tilde{n}$  derivable paired comparisons and thresholds, the dimension of the covariance matrix of the thresholds is  $\tilde{n} \times \tilde{n}$ . The estimation

process of a model with 25 items would (if even possible) take weeks or months with current PCs used in science. To counteract these problems, a second solution was proposed: the presentation of only a few item blocks (Brown & Maydeu-Olivares, 2011).

#### 1.4 The (Multidimensional) Block Design

Contrary to what the vast majority of research citations about the Thurstonian IRT model imply, the Thurstonian IRT model is not equivalent to the MFC format, which will be described in a few moments. The important reparameterization from the Thurstonian factor to an IRT model was proposed in 2010 (Maydeu-Olivares & Brown, 2010). However, the most-cited study considering Thurstonian models is from 2011 (Brown & Maydeu-Olivares, 2011). The main difference is that the study from 2011 proposes the IRT model in union with the idea of using only a few multidimensional item blocks, resulting in a manageable number of blocks and a more parsimonious amount of derived paired comparisons. Additionally, it is proposed that these blocks are multidimensional, as then the ICFs contain more information about the person's traits (Brown & Maydeu-Olivares, 2011). This can be seen from equation (1.17), as for the one-dimensional case it is

$$\Pr(y_l = 1 | \eta) = \Phi\left(\frac{-\gamma_l + \tilde{\lambda}_l \eta}{\sqrt{\tilde{\psi}_l^2}}\right) = \Phi\left(\frac{-\gamma_l + (\lambda_i - \lambda_j)\eta}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}\right) \quad (1.18)$$

while for multidimensional paired comparisons it follows

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_j \eta_b}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}\right). \quad (1.19)$$

If the difference in the item loadings is small, the trait scores of the respondents have virtually no influence on the ICF. For the multidimensional case, this is less severe, as long as the product of item loading and trait value are not identical for both items in a paired comparison.

For the block design, all previously presented equations still hold. The only difference is in the design matrix  $\mathbf{A}$ . Instead of one large block of  $n$  items, the idea is to present  $p$  blocks of  $k$  items, so that respondents are only asked to rank  $k$  items at a time. Also, no item is presented twice. A MFC test is constructed in a way that there are  $n = k \times p$  items. The design matrix does not hold all derivable paired comparisons, but only the set that can be derived from the corresponding blocks. For example, let  $k = 3$ , and  $p = 3$  then it is  $n = 9$  and the design matrix  $\mathbf{A}$  could be

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad (1.20)$$

In a full design (where all  $\tilde{n}$  paired comparisons are derivable), the nine items would yield  $\tilde{n} = 36$  paired comparisons. However, for the three blocks, only nine paired comparisons are considered. This reduces the complexity for respondents and the computations.

There already exist some MFC tests, examples being the Occupational Personality Questionnaire (Brown & Bartram, 2009–2011) or the Big Five triplets (Wetzel & Frick, 2020). These and similar tests are important as with them the effectiveness considering the reduction of faking can be studied. Also, normativity, validity, and reliability can be compared between response scales and FC measures.

Some studies examine if the scores estimated via the MFC format yield normative scores that are comparable between respondents. It was shown that if mixed-keyed blocks are used, trait scores are normative (Frick et al., 2021). For same-keyed blocks, ipsativity cannot



be resolved satisfactorily (Frick et al., 2021; Schulte, Holling et al., 2021). This is also the case if many traits are studied simultaneously (Schulte, Holling et al., 2021).

Considering the reduction of response biases, one study confirms that Halo effects are reduced by the use of the MFC design as compared to rating scales (Brown et al., 2017). As many common response biases are eliminated by design, there is not much research on the reduction of other biases. To study if the MFC format can reduce faking, a meta-analysis was conducted by Cao and Drasgow (2019). Included were 43 primary studies that compared the MFC assessment of the Big Five in a normal and in a simulated high-stakes situation. It was shown that mean score inflation (higher scores on desirable traits) was reduced considerably; however, the reduction depends on the traits of interest. While for rating scales, score inflation on neuroticism is high (around  $d = .45$ , Birkeland et al., 2006; Viswesvaran & Ones, 1999), it is eliminated for neuroticism and openness with the MFC format (Cao & Drasgow, 2019). However, for extraversion, agreeableness and conscientiousness, score inflation still exist with small effect sizes of  $d = .20$  (Cao & Drasgow, 2019). Furthermore, it is shown that desirability matching is important; otherwise, faking is not prevented. In contrast, there also are reports that, in spite of FC blocks matched by desirability, the FC method did not demonstrate faking resistance compared to response scales (Ng et al., 2021).

If results from response scales and the MFC format are compared in the ability to predict criteria and external traits, it was shown that similar or larger construct and criterion-related validities can be achieved (e.g., Frick et al., 2021; Salgado & Táuriz, 2014; Watrin et al., 2019; Wetzel et al., 2020). However, there also exists evidence that convergent and criterion-related validity of the factor scores is similar to the ipsative scores and that discriminant validity is low, even worse than for ipsative scores (e.g., Walton et al., 2020). Furthermore, research indicates that the MFC format and the IRT model can reduce bias and enhance validity. For example, Brown et al. (2017) showed that in organizational assessment,

a response scale approach yields a strong positive manifold of all traits, that is all traits are highly correlated, even though they are not expected to correlate this highly from a theoretical point of view. Using the FC method, this positive manifold can be greatly reduced, and correlations are more in line with theoretical expectations (Brown et al., 2017).

It has also been observed that correlations between traits can differ between the MFC and a response scale format, which is sometimes discussed as a potential problem (e.g., Guenole et al., 2018). There are two potential explanations for differing correlational results. First, in FC blocks, items are presented and compared to other items, which results in responses on items that are not independent from the other items (see also Lin & Brown, 2017). Second, if the FC format has any impact on reducing biases and faking, why should identical correlational results be expected, especially as changes in correlations due to faking were already reported (Moors, 2012)? It is more important that the observed pattern of correlation fits the theoretical expectations, which can also be observed (Guenole et al., 2018).

On the downside, the MFC format yields generally lower to insufficient reliabilities even under ideal circumstances, such as with mixed-keyed blocks (Frick et al., 2021; Schulte, Holling et al., 2021). This generally makes sense, as the information within ordinal binary responses is lower than in interval scales assumed when using response scales. Note also that there is a difference in how reliabilities are estimated and used between classical response scale analysis (often classical test theory) and IRT modeling. In classical test theory, reliability is estimated for the test, while in IRT setting, reliability is dependent on the trait scores. This makes the comparison and differentiation for reliability estimates more complex.

Other research is about the detection of differential item functioning (Lee et al., 2021) and measurement invariance (Lee & Smith, 2020). Overall, the results presented by previous research are mixed. While some studies show that the MFC format reduces fakeability,

biases, and enhances the validity of estimates, other studies focus on more problematic features such as reduced reliability and a paradoxical relationship between fakeability and estimation precision. What is more, the MFC format and the Thurstonian IRT model are used equivalently in the literature and virtually all findings based on simulations and empirical data are based on the MFC format and the corresponding Thurstonian IRT model (with exceptions being Maydeu-Olivares and Brown, 2010; Maydeu-Olivares and Böckenholt, 2005). Quite impressively, the initial models (simple and factor models; Maydeu-Olivares and Böckenholt, 2005) are barely considered at all, despite their many important use cases for item and desirability scaling in psychological science and application. One exception is a study on item scaling, which uses simple Thurstonian models to estimate desirability values for a set of 15 items (Jansen & Schulze, 2023a). The main goal of the study was to compare different methods of item scaling (paired comparisons, ranking, rating) and to test whether the items are scalable on an interval scale. In this study, a rating scale task, a ranking block of 15 items, and a full paired comparison design, with all 105 paired comparisons, were used. The main result is that item scaling works well with simple Thurstonian models. This study would fit well into the present work; however, it is based on a master's thesis of the author and is therefore not included in the present thesis. Importantly, the study of Jansen and Schulze (2023a) inspired the current work, as it was noticed that item scaling with the block design is not possible, as will be described subsequently.

### **1.5 Summary of Issues with the (Multidimensional) Block Design**

For the present work, two separate aspects of Thurstonian FC modeling are considered: the model type and the design. The model type is determined by the parameters of interest. If item scaling and the item utilities are of interest, simple Thurstonian or Thurstonian factor models are used. If the interest is in normative latent trait score estimation per respondent, then the IRT model is suitable if a ranking task is used. For intransitive data,

a factor model would also be suitable. The design, on the other hand, is defined by the design matrix  $A$  and determines which blocks of items are considered and which paired comparisons can be derived by the specific design.

The Thurstonian FC literature in recent years almost completely lacks a discussion about the item scaling perspective, which is surprising, and a severe problem. Coming back to a prominent goal of the FC method, the idea is to construct tests that are not prone to response biases and faking. A FC test is less fakable, if all items within a FC block are equally desirable (e.g., Bürkner, 2022; Bürkner et al., 2019; Edwards, 1953; Schulte, Holling et al., 2021; Schulte, Kaup et al., 2021); otherwise, respondents have an easy choice by choosing the more desirable option. While desirability values for items seem to be an important aspect of any test construction, especially FC test construction, the focus on desirability values is small. Often, desirability ratings are obtained by aggregating them from a response scale (e.g., Christiansen et al., 2005; Jackson et al., 2000; Wetzel & Frick, 2020) or expert scoring (e.g., psychology students; Dunette et al., 1962; Heggstad et al., 2006). The use of distance scaling or FC methods for desirability scaling is done seldomly (Edwards & Thurstone, 1952; Jansen & Schulze, 2023a), even though the statistical means to accomplish such scaling exist. This is even more impressive, as the MFC format and Thurstonian modeling is used. It seems natural that the same modeling framework would be used to get desirability values and to construct FC tests with matched blocks.

On a similar note, the current recommendations (see Brown & Maydeu-Olivares, 2011) for the FC test construction emphasize the importance to use mixed-keyed blocks. This means that a block contains at least one positively and one negatively keyed item. This can be explained by the already noted larger information in ICFs when the difference in item loadings is large, see equations (1.18) and (1.19). However, it is can be difficult, if not impossible, to construct mixed-keyed blocks that are matched by desirability. This is because

negatively keyed items of a trait often correspond to less desirable behaviors and properties of a person (see also Bürkner, 2022; Bürkner et al., 2019).

There are two more problems to consider that have not yet been mentioned in the literature. First, there seem to be some unspoken assumptions about the use of the block design. In general, it is assumed that each item of a trait is equivalent to any other item, so it does not matter which item of trait X is compared to another item of trait Y. While the general idea is understandable, the construction of FC tests is much more complex than the construction of response scales. Given a set of 15 items and constructing five triplets, there are more than 43 billion possible ways to build the test. When considering multidimensional blocks ( $m = 3$ ) the complexity is reduced, but there are still 14 400 combinations possible (more detailed analysis in Study 1). It is likely that the configuration of the items within blocks also affects the properties of the test, which has also been reported (Lin & Brown, 2017). More importantly, the question is about which model is considered in the first place. As noted, from one item set, many different tests can be constructed. However, the initial underlying model of interest should be the same for each of the tests. If it were not the same model, then the operationalization, that is the test, would determine the model to test. However, a given item set should already determine the unique model of interest, that is a full design with all  $\tilde{n}$  paired comparisons (more details in Study 1). So far this is not discussed within the Thurstonian IRT literature, on the contrary, simulation studies use data from the specific block designs, and test the fit to the specific model configuration, instead of the full design. Second, to identify the block models, the identification constraints need to be applied to every block. This is severe, as parameters (e.g., item loadings, utility means) are defined on different scales per block and thus, are not comparable. Taken together, the usefulness of Thurstonian modeling for desirability matching is reduced.

## 1.6 Overview of Manuscripts

The present work focuses on describing and discussing the afore mentioned Thurstonian models and proposes a (at least partial) solution to their shortcomings. The first study, presented in chapter 2, reviews the current literature on Thurstonian modeling and delves into the majority of the problems, discussed in the current chapter in more detail. The second study, presented in chapter 3, builds on this discussion by presenting a solution in adapting the design matrix  $\mathbf{A}$  by linking blocks. The linking procedure makes sure that the blocks and the parameter estimation are not independent anymore. Subsequent simulation studies show a more accurate model estimation compared to unlinked block designs. The overall parameter estimation is improved by only requiring one set of identification constraints for the entire item set, regardless of the specific set of blocks used. In chapter 3, the simulation is restricted to same-keyed blocks and to a large sample size for psychological science of  $N = 2000$ . The third study, presented in chapter 4, discusses the different sources of information with FC designs and the inclusion of mixed-keyed blocks, while also considering the importance of desirability matching. The integration of the Thurstonian linked block design with different information obtainable by FC data leads to accurate model estimation. This leads to updated recommendations for FC test construction, based on the advances of the present work. In chapter 5, an important software contribution is described, that made the simulation studies in chapters 3 and 4 possible. Finally, the overall results and implications are discussed in the final chapter.

## **2. Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues**

Jansen, M. T. & Schulze, R. (2023b). *Linear factor analytic Thurstonian forced-choice measurement: Current status and issues*. Manuscript submitted for publication in Educational and Psychological Measurement.

### **2.1 Summary**

#### **2.1.1 Review about Thurstonian Forced-Choice Models**

The first part of the manuscript reviews different FC models, with a focus on models based on Thurstone's law of comparative judgement (Thurstone, 1927). A brief differentiation from ideal-point models (Coombs, 1960) is also provided. Early approaches include full information maximum likelihood estimation of model parameters, which becomes unfeasible fast due to the need of numerical integration (Böckenholt, 1993; Maydeu-Olivares, 1999; Yao & Böckenholt, 1999). The manuscript then goes on to present and describe in detail the three already presented model types: the simple Thurstonian model (Maydeu-Olivares & Böckenholt, 2005), the Thurstonian factor model (Maydeu-Olivares & Böckenholt, 2005; Maydeu-Olivares & Brown, 2010) and the Thurstonian IRT model (Maydeu-Olivares & Brown, 2010). Additionally, the identification of these models and dealing with redundancies in a ranking task are discussed. Furthermore, the manuscript provides details on how to estimate MAP scores and about the reliability estimations for latent trait estimation. Finally, the multidimensional block design proposed by Brown & Maydeu-Olivares (2011) is described.

#### **2.1.2 Problems in Thurstonian Forced-Choice Modeling**

The second part of the manuscript addresses the problematic features of Thurstonian FC models, particularly the block design. These issues include the limiting factor of the large number of paired comparisons required for a full design, that is the complete information on

the  $\tilde{n}$  paired comparisons, and the paradoxical relationship between the need for mixed-keyed blocks for precise parameter estimation and the need to match items within a block by desirability. It appears that matching the desirability of items for test construction can only be achieved through using same-keyed blocks. The use of mixed-keyed blocks, on the other hand, would result in blocks that are not matched by desirability.

The manuscript then shifts to the estimation and simulation of parameters based on specific block designs (the operationalization), rather than the uniquely determinable full design. It is shown that the same item set with the same true parameters can lead to different parameter simulation, and thus estimation, based on the design matrix  $\mathbf{A}$ , that is by including another configuration of blocks. These claims are supported by reanalyzing empirical data from another study (Jansen & Schulze, 2023a) where data of all 105 paired comparisons of 15 items was available. The results show that the bias of model estimation can be substantial and is dependent on the specific design used. Additionally, it is highlighted that there are numerous identification constraints that need to be applied for each block, which makes item scaling challenging. This could only be solved if accurate assumptions on the order of all fixed parameters would be available.

### **2.1.3 Discussion and Conclusion**

The final conclusion of the first study is, that Thurstonian FC modeling is prone to be highly biased and published results should be used and interpreted with caution, due to the bias and fakeability of the blocks constructed thus far. Most importantly, a FC design should be chosen, so that parameter estimation is possible on the same scale for all item or person parameters.



## 2.2 Manuscript


### **Linear Factor Analytic Thurstonian Forced-Choice Models: Current Status and Issues**

Markus T. Jansen and Ralf Schulze

University of Wuppertal

#### **Author Note**

Markus T. Jansen  <https://orcid.org/0000-0002-5162-4409>

Ralf Schulze  <https://orcid.org/0000-0001-5780-8973>

The author(s) declared no conflicts of interest with respect to the authorship or the publication of this article. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence concerning this article should be addressed to Markus Thomas Jansen, Institute of Psychology, University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany. E-mail: [mjansen@uni-wuppertal.de](mailto:mjansen@uni-wuppertal.de)

### 2.2.1 Abstract

Thurstonian Forced-Choice modeling is considered to be a new powerful tool to estimate item and person parameters while simultaneously testing the model fit. This assessment approach is associated with the aim of reducing faking and other response tendencies that plague traditional self-report trait assessments. As a result of major recent methodological developments, the estimation of normative trait scores has become possible in addition to the computation of only ipsative scores. This opened up the important possibility of comparisons *between* individuals with Forced-Choice assessment procedures. A Multidimensional forced-choice (MFC) format has also been proposed to estimate individual scores. Customarily, items to assess different traits are presented in blocks, often triplets, in applications of the MFC, which is an efficient form of item presentation but also a simplification of the original models.

The present study gives a comprehensive review of the present status of Thurstonian Forced-Choice models and their variants. Critical features of the current models, especially the block models are identified and discussed. It is concluded that MFC modeling with item blocks is highly problematic and yields biased results. In particular, the often-recommended presentation of blocks with items that are keyed in different directions of a trait proves to be counterproductive considering the goal to reduce response tendencies. The consequences and implications of the highlighted problems are further discussed.

*Keywords:* Thurstonian modeling, multidimensional forced-choice format, item response theory, structural equation modeling

### **2.2.2 Linear Factor Analytic Thurstonian Forced-Choice Models: Current Status and Issues**

Many basic and applied empirical research efforts in psychology, as well as other social and behavioral sciences, include measurements of latent traits with self-reports. Pertinent examples can be found in psychiatric (Carey et al., 2004; Cochrane-Brink et al., 2000), educational and legal applications (e.g., Holden & Passey, 2010), or in the context of research on predicting vocational performance (Barrick & Mount, 1991; Tett et al., 1991). Given that many psychological constructs and behavioral tendencies (e.g., being emotionally stable) are often efficiently accessible via introspection at the individual level self-assessments appear to be a perfectly viable way of assessing the targeted constructs. In order to estimate scores of Emotional Stability, for example, an individual's self-reported degree of agreement on a rating scale (e.g., from "totally disagree" to "totally agree") to a statement (e.g., "In difficult situations I stay calm") may serve as an indicator. Of course, many such Likert-type items would generally be needed as indicators to arrive at a sufficiently precise estimate of the individuals' scores on the targeted construct and to be able to make interindividual comparisons.

A quite common problem with many forms of self-assessments, including the above-mentioned Likert-type items in particular, is their susceptibility to a myriad of response biases. Their proneness to such response distortions depends on the context as an abundance of empirical research shows and is particularly prevalent, but not limited to, high-stakes contexts (e.g., Ziegler et al., 2012). Typical response biases include the tendency to agree to a statement irrespective of its content (acquiescence bias), the tendency to very strongly agree or disagree (extreme responding), to give responses that are expected to be socially approved (social desirability bias, e.g., Cronbach, 1946; Jackson & Messick, 1958; Paulhus, 2002), and faking (Ziegler et al., 2012). In addition to response biases, other methodological issues also

prevail with Likert-type items. They include the question if interpretation and use of a given response scale is consistent both within and between individuals. Different approaches to mitigate the influences of response biases and faking have been proposed, but most of them are not without problems as they are associated with new challenges like complicated item-phrasing, restrictions in item sets (e.g., only “neutral” items), or the reduction of reliability and construct validity (Paulhus & Vazire, 2007).

The present paper focuses on one particular methodological approach to deal with the majority of these problems and challenges, namely the Forced-Choice design (FC). The main characteristic of the FC approach is that the respondents are required (“forced”) to make a choice between a set of stimuli (e.g., statements) with respect to a given criterion (e.g., the degree of fit of a statement as a descriptor of oneself). In contrast to Likert-type items, individual statements are not rated on a scale. Instead, comparisons between stimuli are required. Usually, two or more stimuli are presented in blocks (see Figure 2.1). The block may consist of only two stimuli as is the case in panel A of Figure 2.1. In this paired-comparison, the respondents compare the two statements as descriptors of themselves and choose the better one, even if none or both of the statements provides a very good description of the person. It is also possible and, in fact, more common to present more than two stimuli for a simultaneous comparison as is the case in panel B of Figure 2.1. Here, respondents are asked to rank the statements by choosing one statement that fits the *best*, and one statement that fits the *least* as a description of oneself. By implication, these two decisions result in a full ranking of all three stimuli. Note that individuals are not asked to respond to each of the items separately but to give a preference ranking over the combined block of statements.

One major drawback of FC designs is the limited information the resulting data provide about interindividual differences. Take the ranking of the three stimuli in the lower part of Figure 2.1 as an example. It indicates that a respondent A thinks that an Openness item

**Figure 2.1**

*Examples for the Forced-Choice Format.*

<b>A</b> Please select the option that describes you the best. <input type="radio"/> In difficult situations I stay calm. <input type="radio"/> I like partys.	
<b>B</b> Please rank the options according to how well they describe you.	
<input type="text" value="I like art."/>	<input type="text" value="1."/>
<input type="text" value="I seldom am in a good mood."/>	<input type="text" value="2."/>
<input type="text" value="I talk a lot."/>	<input type="text" value="3."/>

("I like art") is more descriptive of him- or herself as compared to an Extraversion item ("I talk a lot"). Hence, a within-person conclusion based on this information would be that the respondent may be more open to new experiences as compared to being extraverted. If a respondent B produced the exact same ranking, the same conclusion was valid for this person too. However, there is no information provided that could be used to make between-person conclusions like "respondent A is less extraverted than respondent B". In the same vein, traditional scoring methods for such data lead to so-called ipsative scale scores, which also cannot be compared to the scores of other respondents. This results from the fact that if the stimuli are assigned their ranks as scores, the same total number of points is distributed between stimuli within an FC block and for each block of the same size. Therefore, if such scores are added up within one block and across all blocks, the total score of a test is the same for every person. As a result, the usual score interpretations that refer to between-subjects comparisons and classic psychometric analysis are impossible (for further discussion see Baron, 1996). To overcome the limitations in interpretation with ipsative scoring two general model frameworks are considered here. They potentially allow for the estimation of so-called normative scores, that is, scores that allow for interindividual score interpretations.

The first model framework concentrates on the Ideal Point model (IP) whereas the second is based on Linear Factor Analysis models or the dominance response process (see Brown, 2016). The latter is more commonly used and it is assumed that each item is linearly related to only one factor. The higher a respondent's score on the factor, that is, the higher the utility is, the stronger the agreement to the item should be. To illustrate, take "I keep my desk orderly" as an item example. The higher a respondent's score on conscientiousness is, the more they should agree. Similarly, the same respondent should also tend to disagree with a statement such as "I sometimes forget where I put my things".

In IP models, it is considered that some items are not linear manifestations of a factor. For example, Thurstone (1928) argued that for the statement "Firearms should not belong in private hands" the psychological value (utility) is the highest for a person with the exact same level of the attitude towards Militarism. Statements that either represent relatively higher or lower attitudes toward the object should both have a lower utility. Similarly, the utility for the statement is lower for respondents with higher or lower attitude levels. The point of maximum utility between a person and the attribute continuum is called Ideal Point (Coombs, 1960). In IP models, the idea is that a person can be represented by a point on the attribute's continuum and that the utility of an item is larger, the lower the distance between the respondents and the items' location is.

In situations where IP models are valid, the use of dominance-type models as general factor analytic models would be inappropriate. Instead, other modeling approaches have been proposed. These include the Multi-Unidimensional Pairwise Preference Model (MUPP; Stark et al., 2005), the Generalized Graded Unfolding Model (GGUM; Roberts et al., 2000), and the McCloy-Heggstad-Reeve Unfolding model (McCloy et al., 2005). All of these IP models are based on the Bradley-Terry model (Bradley & Terry, 1952) and will not be further considered in the present paper. The Zinnes-Griggs IP model is yet another potential

framework to overcome the problem of ipsative scoring and is based on Thurstone's model.

However, within the scope of this study, we will only consider linear factor analytic models.

The original and typical use of Thurstone's Law of comparative judgment (Thurstone, 1927, 1931) was to scale stimuli according to specific criteria. Initially, the estimation of Thurstone models with latent variables was unfeasible for practical applications, especially when many (more than 10) items were to be compared (Maydeu-Olivares, 1999; Yao & Böckenholt, 1999). Though this is more of a problem with full information maximum likelihood estimation as the multivariate normal integral needs to be evaluated numerically (Böckenholt, 1993). Even today numerical integration still is considered computational heavy. With the development of better technical resources and limited information estimation methods, a practically useful confirmatory factor analytic (CFA) method for stimulus scales based on the Linear Factor Analysis model was proposed (Maydeu-Olivares & Böckenholt, 2005). These Thurstonian CFA models also made the differentiation and estimation of the different Thurstone cases possible. Moreover, by way of a reparameterization of the Thurstonian CFA models into IRT models a potential solution to the problem of ipsative scoring of respondent scoring was proposed (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). As result, both stimuli and respondents' scores can be estimated in this framework which makes it particularly attractive for research and applications.

Alas, several studies (e.g., Bürkner, 2022; Bürkner et al., 2019; Frick et al., 2021; Schulte et al., 2021) have already shown that Thurstonian modeling is not without issues and challenges, the review, identification, and specification of which will be the focus of the current study. This will be done in four sections. The first section sets the stage by describing how responses in FC settings are coded into binary outcome variables. In the second section, current Thurstonian models that all employ binary outcome variables will be specified and examined. In the third section, the status of current Thurstonian modeling and associated

prevailing issues as well as open questions are reviewed and summarized. Lastly, the results and consequences of the findings of the previous section are discussed.

### 2.2.3 Binary Coding of Forced-Choice Responses

The standard procedure of how to code responses in an FC design goes back to Thurstone's law of comparative judgment (Thurstone, 1927). The starting point are paired comparisons or rankings of a specific set of stimuli as described in the previous section. The coding as explained in this section is strongly oriented to Maydeu-Olivares and Böckenholt (2005), and Brown & Maydeu-Olivares (2011) for consistency.

It is assumed that at any given presentation each stimulus evokes a so-called discriminative process that results in a scale value  $t_i$  (for stimulus  $i$ ) on a latent utility continuum of a respondent. This scale value corresponds to the sensation that is elicited by the stimulus at presentation. A repeated presentation of a specific stimulus is not assumed to elicit the exact same sensation within or across respondents. Instead, the latent utilities follow some distribution (almost without exception assumed to be normal) with expected value  $\mu_{t_i}$  and standard deviation  $\sigma_{t_i}$ . The same is true for any other stimulus  $j$  that may be presented. Simultaneous presentation of both stimuli  $i$  and  $j$  results in a process depicted in Figure 2.2. The difference between the latent utilities determines if stimulus  $i$  or  $j$  is chosen (i.e., preferred, if preference is the criterion). The entire process is not observable, but the response  $y_i$  that results from this process is and it is coded as

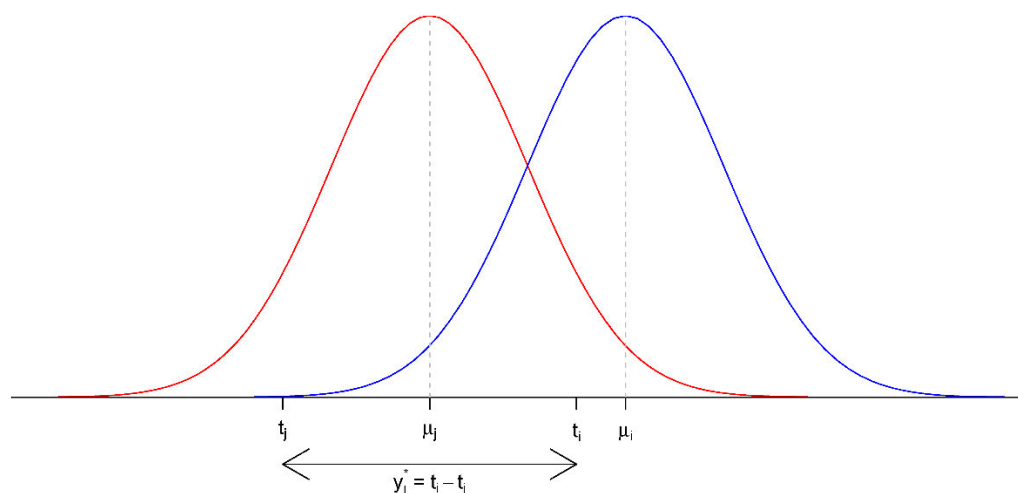
$$y_i = \begin{cases} 1 & \text{if } t_i \geq t_j \\ 0 & \text{if } t_i < t_j \end{cases} \quad (2.1)$$

According to Equation (2.1), the response is coded as 1 if stimulus  $i$  is preferred over stimulus  $j$  and 0 otherwise. As an example, consider  $n = 4$  items labeled as  $\{A, B, C, D\}$ . For these items  $\tilde{n} = n(n-1)/2 = 6$  nonredundant paired comparisons  $\{i, j\}$  can be constructed:



**Figure 2.2**

*Examples of the Discriminative Process of One Paired Comparison Between Stimulus  $i$  and Stimulus  $j$ .*



{A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}. If for {A, B} A is preferred over B then  $y_{\{A,B\}} = 1$ . This can be done for every paired comparison (for the moment we assume a full ranking on all items; Maydeu-Olivares & Böckenholt, 2005):

Ranking				Ordering			
A	B	C	D	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>
1	3	4	2	A	D	B	C

Binary outcomes

{A, B}	{A, C}	{A, D}	{B, C}	{B, D}	{C, D}
1	1	1	1	0	0

In cases where not the full, but only a partial ranking design is used, the full information cannot be retrieved. As an example, consider the respondent is only requested to provide the most and least preferred of the four items, not a ranking of all the items as before. Then the ranking would be (also see Brown & Maydeu-Olivares, 2011):

Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues

Ranking				Ordering			
A	B	C	D	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>
1	?	4	?	A	?	?	C

Binary outcomes					
{A, B}	{A, C}	{A, D}	{B, C}	{B, D}	{C, D}
1	1	1	1	.	0

As can be seen, the information about the comparison {B, D} is missing.

The difference between the processes  $y_i^* = t_i - t_j$  is not observable but the relationship between unobservable  $y_i^*$  and observed  $y_i$  is

$$y_i = \begin{cases} 1 & \text{if } y_i^* + e_i \geq 0 \\ 0 & \text{if } y_i^* + e_i < 0 \end{cases} \quad (2.2)$$

Note that there is now an error term  $e_i$  for each comparison. This is done in order to cover the issue of transitivity that may arise in the present context. Transitivity is given for three stimuli A, B, and C when A is preferred over B, B over C and also A over C. The latter choice is implied in rankings but may not be given in paired comparisons. Respondents are absolutely free in a paired comparison design to choose C over A, thereby violating transitivity. This is why the term  $e_i$  is included here. Again, a major consequence of a (full) ranking design is that all responses are always transitive. Thus, the error term  $e_i = 0$  for every comparison in a ranking design but not in paired comparison designs.

### 2.2.4 Thurstonian Models

There are several Thurstonian linear factor analytic models that differ in focus on estimation (item-centered vs. person-centered) and designs (full vs. block). In the following section, the different types of Thurstonian models will be specified and examined and their usability as well as differences will be pointed out.

## Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues

---

In its original form, Thurstone gave five cases for the estimation of latent utilities:

Case I is the least restrictive model and intended for one respondent with multiple responses on the same pair of items. It does not state any assumptions about the distributions of discriminative processes and involves correlations between discriminative processes (latent utilities). This leads to the model represented by

$$t_i - t_j = x_{ij} \sqrt{\sigma_{t_i}^2 + \sigma_{t_j}^2 - 2r\sigma_{t_i}\sigma_{t_j}} \quad (2.3)$$

where  $t_i$  is the latent utility of item  $i$ ,  $x_{ij}$  is the proportion of trials where  $i$  was chosen over  $j$ , and  $r$  is the correlation between the latent utilities of item  $i$  and item  $j$ . Case II is intended for multiple respondents and a normal distribution of each latent utility is assumed here. By additionally assuming that the correlations between latent utilities are zero, the following simplified model follows in Case III

$$t_i - t_j = x_{ij} \sqrt{\sigma_{t_i}^2 + \sigma_{t_j}^2}. \quad (2.4)$$

Under a Case IV model, the variance of each latent utility is assumed to be small. Finally further assuming equal variances of these utilities results in Case V. Equation (2.5) shows the Case V model where the scaling by the constant  $\sqrt{2}\sigma$  is ignored (Thurstone, 1927)

$$t_i - t_j = x_{ij}. \quad (2.5)$$

The matrix representation of the system of linear equations would need to be of full rank in order to be solved for the estimation of a unique vector of utilities. This means that the matrix of equations that need to be solved is transformable into a triangular matrix where all elements below the diagonal are zero. Simultaneously, the scale difference values have to eliminate each other in a specific pattern. This will be nearly never the case, leading to multiple possible results of utility vectors. To solve this problem, identification constraints need to be applied. Thurstone (1928) recommends building a  $n \times n$  matrix, ordering all items

by adding up the probabilities column-wise, and fixing the utility of the first item to zero to set the metric. Scale values are obtained by order of probability of endorsement. Alternative approaches to estimate the utilities are the simple Thurstonian model and Thurstone factor models, described in what follows.

#### 2.2.4.1 Simple Thurstonian Model

The simple Thurstonian model was introduced by Maydeu-Olivares and Böckenholt (2005). As it is based on Thurstone's Law of Comparative Judgment, its focus of estimation is on the item scale values. Writing all latent differences  $y_i^*$  between the latent utilities in vector-matrix form yields

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e} \quad (2.6)$$

for  $\mathbf{y}^*$ , which is a  $\tilde{n} \times 1$  vector. The  $n \times 1$  vector of the latent utilities is  $\mathbf{t}$ , the  $\tilde{n} \times n$  design matrix is  $\mathbf{A}$ , where the rows of  $\mathbf{A}$  correspond to the paired comparisons and the columns correspond to the choice alternatives. Finally,  $\mathbf{e}$  is a  $\tilde{n} \times 1$  vector of uncorrelated random errors terms. As a consequence, the covariance matrix  $\mathbf{\Omega}^2$  of the residuals is diagonal. For example, given  $n = 4$ , the design matrix is

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \quad (2.7)$$

Generally, it is assumed that in the population of respondents, the latent utilities follow a multivariate normal distribution (Maydeu-Olivares & Böckenholt, 2005; Thurstone, 1927), that is

$$\mathbf{t} \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \quad (2.8)$$

where  $\boldsymbol{\mu}_t$  is the mean vector and  $\boldsymbol{\Sigma}_t$  is the variance-covariance matrix of latent utilities.

Equation (2.9) includes the latent utilities in a structural equation model with binary indicators as follows

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t, \text{ and } \boldsymbol{\Sigma}_{y^*} = \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}' + \boldsymbol{\Omega}^2 \quad (2.9)$$

where  $\boldsymbol{\mu}_{y^*}$  is a vector with the means and  $\boldsymbol{\Sigma}_{y^*}$  is the variance-covariance matrix of the  $\tilde{n}$  latent differences. To estimate such models, the thresholds and tetrachoric correlations of the normal variables must be computed (Muthén, 1978). The elements of the covariance matrix  $\boldsymbol{\Omega}^2$  are not free to vary but constrained to be  $\boldsymbol{\Omega}^2 = \mathbf{I} - \text{diag}(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}')$ . This restriction is necessary to identify the variances in  $\boldsymbol{\Sigma}_{y^*}$ . However, this choice of restriction would imply  $\text{diag}(\boldsymbol{\Sigma}_t) = \mathbf{I}$ , which is not convenient for some Thurstone cases (II, III, and IV). Instead of  $\boldsymbol{\Omega}^2 = \mathbf{I} - \text{diag}(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}')$ , the latent differences are standardized to  $\mathbf{z}^* = \mathbf{D}(\mathbf{y}^* - \boldsymbol{\mu}_{y^*})$  with  $\mathbf{D} = \left[ \text{diag}(\boldsymbol{\Sigma}_{y^*}) \right]^{-1/2}$ . The standardized latent differences are still assumed to follow a multivariate normal distribution, now with  $\boldsymbol{\mu}_{z^*} = \mathbf{0}$ . To identify the model, the thresholds and the matrix of tetrachoric correlations  $\mathbf{P}_{z^*}$  must then be constrained instead of the covariance matrix  $\boldsymbol{\Omega}^2$ . The relationship between observed responses and standardized latent differences is one of the  $\tilde{n}$  thresholds  $\tau_l$

$$y_l = \begin{cases} 1 & \text{if } z_l^* \geq \tau_l \\ 0 & \text{if } z_l^* < \tau_l \end{cases} \quad (2.10)$$

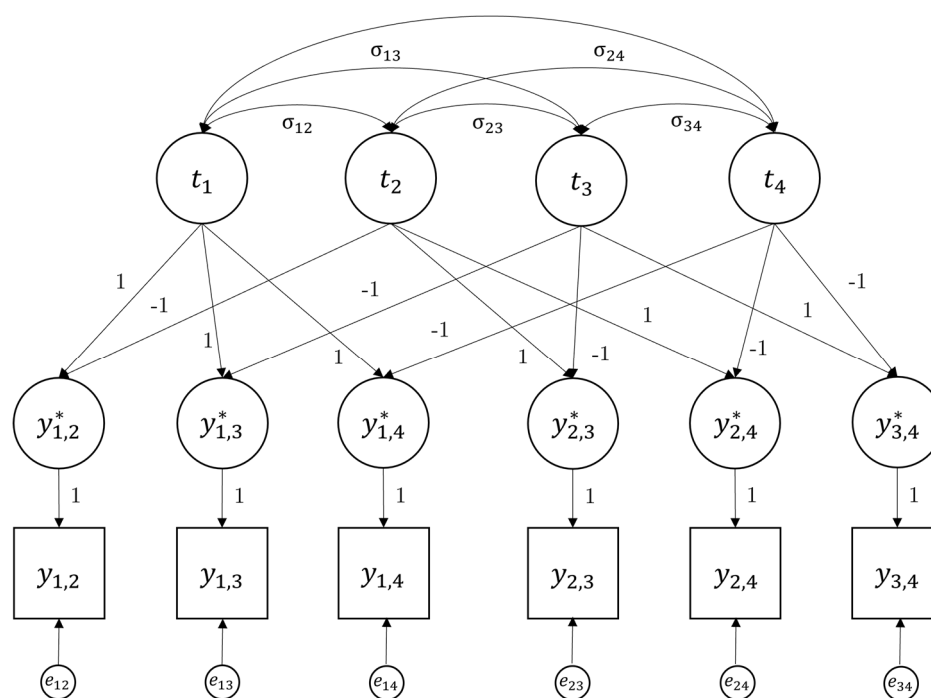
where the vector of thresholds is  $\boldsymbol{\tau} = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}_t$  and the tetrachoric correlation matrix is

$$\mathbf{P}_{z^*} = \mathbf{D}(\boldsymbol{\Sigma}_{y^*})\mathbf{D} = \mathbf{D}(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}' + \boldsymbol{\Omega}^2)\mathbf{D} \quad (2.11)$$

(Maydeu-Olivares & Böckenholt, 2005; Muthén, 1978). Figure 2.3 shows an example of a simple Thurstonian model for  $n = 4$ .

**Figure 2.3**

*Examples of a Covariance Structure of a Simple Thurstonian Model for  $n = 4$  Items.*



The simple Thurstonian model allows for the estimation of the mean latent utilities and the test of Thurstone’s cases. For example, for Thurstone’s Case V model it is assumed that all latent utilities  $\mathbf{t}$  are uncorrelated and that the variances of the latent utilities are equal. If such constraints were used for  $\Sigma_{\mathbf{t}}$ , then (comparative) tests of the restrictions implied by different Thurstone cases could be conducted and an assessment of the relative fit would be possible (for more details see Maydeu-Olivares & Böckenholt, 2005).

### 2.2.4.2 Thurstonian Factor Models

Instead of a correlated model as shown in Figure 2.3, there may be a known structure on the  $n$  items implied by a higher-order model with dichotomous indicators as shown in Figure 2.4, for example. Such a model is called the Thurstonian factor model (Maydeu-Olivares & Böckenholt, 2005). It is very common in psychological assessment that the assumption of such a structure can be assumed and justified for a set of stimuli. For example, in personality assessment, items presented as stimuli in an FC design may be uniquely

classified as indicators of traits such as conscientiousness, agreeableness, or any other known personality factors. Let  $m$  be the number of such common factors (also called latent traits).

The  $n$  latent utilities in  $\mathbf{t}$  can then be expressed as

$$\mathbf{t} = \boldsymbol{\mu}_t + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (2.12)$$

where  $\mathbf{\Lambda}$  is the  $n \times m$  matrix of factor loadings of the latent utilities on the latent traits,  $\boldsymbol{\eta}$  is a  $m \times 1$  vector of the latent traits, and  $\boldsymbol{\varepsilon}$  is a  $n \times 1$  vector of unique factors (error term). All common factors have a mean of zero and variances of unity but may be correlated as allowed for in an  $m \times m$  correlation matrix  $\boldsymbol{\Phi}$ . The unique factors are also assumed to have mean zero and to be uncorrelated. Thus, the variance-covariance matrix  $\boldsymbol{\Psi}^2$  of error terms is diagonal.

With Equation (2.6) and (2.12) the latent differences are

$$\mathbf{y}^* = \mathbf{A}(\boldsymbol{\mu}_t + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\mathbf{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} \quad (2.13)$$

which results in the mean and covariance structure

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t, \text{ and } \boldsymbol{\Sigma}_{y^*} = \mathbf{A}(\mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{A}' + \boldsymbol{\Omega}^2. \quad (2.14)$$

The implied tetrachoric correlations follow from Equation (2.11) and (2.14) to be

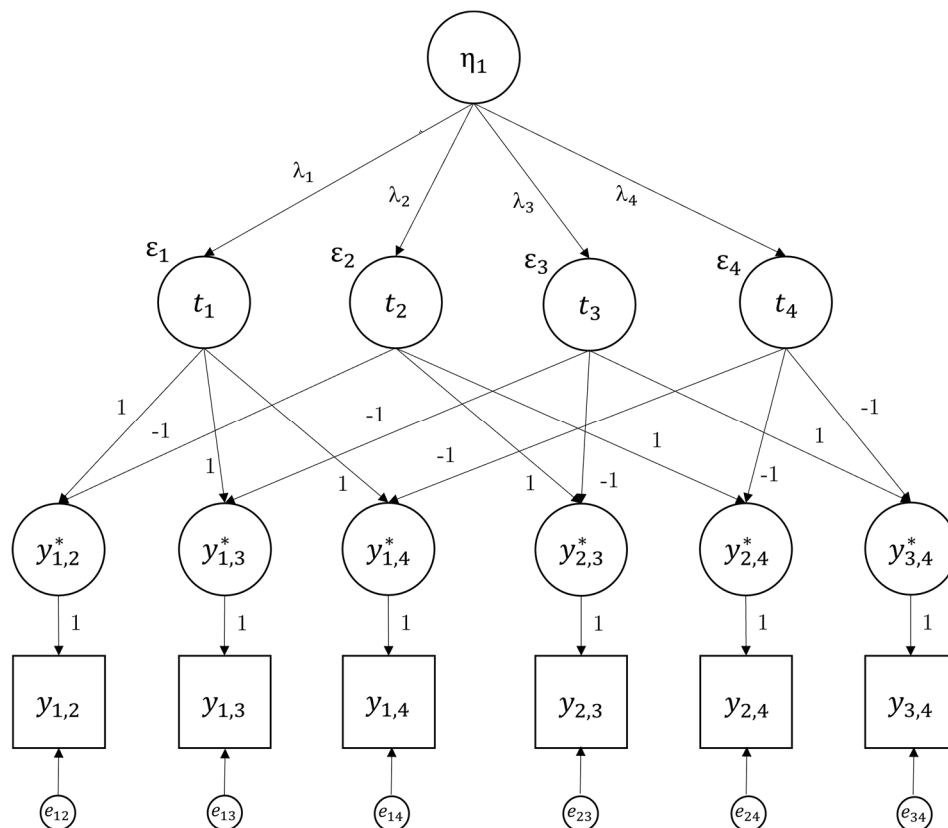
$$\mathbf{P}_z = \mathbf{D}(\boldsymbol{\Sigma}_{y^*})\mathbf{D} = \mathbf{D}(\mathbf{A}(\mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{A}' + \boldsymbol{\Omega}^2)\mathbf{D} \quad (2.15)$$

with the same relationship between standardized latent differences and observed responses as in Equation (2.10) and vector of thresholds  $\boldsymbol{\tau} = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}_t$ . Figure 2.4 gives an example for a Thurstonian factor model with  $n = 4$  and a single factor  $m = 1$ .

For full ranking data, intransitive responses are not possible. As a consequence, all model equations also hold in a ranking design, but with  $\mathbf{e} = 0$  and  $\boldsymbol{\Omega}^2 = 0$ . Additionally, rankings include only a subset of all possible paired comparisons, so the number of degrees of freedom must be adjusted, as there are

**Figure 2.4**

*Examples of a Covariance Structure of a Thurstonian Factor Model for  $n = 4$  and  $m = 1$*



$$r = \frac{n(n-1)(n-2)}{6} \quad (2.16)$$

redundancies for the tetrachoric correlations and thresholds (Maydeu-Olivares, 1999, Maydeu-Olivares & Böckenholt, 2005). Therefore,  $r$  must be subtracted from the number of degrees of freedom reported by structural equation modeling software, and fit indices must be adjusted accordingly.

### 2.2.4.3 Use and Identification of Simple Thurstonian Models and Thurstonian Factor

#### Models

One latent utility mean must be fixed to identify the model, like  $\mu_n = 0$ , for example.

All other latent utility means are estimated relative to the fixed value. Note that the



identification constraints are not unique so that the model parameters are always of relative nature. For example, using  $\mu_n = 5$  instead of  $\mu_n = 0$  as a constraint would result in different means for all other latent utility factors because estimation is relative to the constrained value. The results would be the same if 5 was added to all estimates that result from using  $\mu_n = 0$ . For simple Thurstone models at least one variance must additionally be fixed to  $\sigma_n^2 = 1$ , for example. If all parameters of the variance-covariance matrix  $\Sigma_t$  need to be estimated (Case II), two variances and all covariances of one of the items with fixed variances must be constrained. An example would be:  $\sigma_1^2 = \sigma_n^2 = 1$  and  $\sigma_{in}^2 = 0$  for all  $i$  from 1 to  $n$ . To summarize, for  $n = 4$  this results in

$$\boldsymbol{\mu}_t = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ 0^* \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_t = \begin{pmatrix} 1^* & \sigma_{21} & \sigma_{31} & 0^* \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & 0^* \\ \sigma_{31} & \sigma_{23} & \sigma_3^2 & 0^* \\ 0^* & 0^* & 0^* & 1^* \end{pmatrix}. \quad (2.17)$$

Identification constraints are marked with an Asterix \*. For Case III, all covariances are constrained to be zero, thus

$$\boldsymbol{\mu}_t = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ 0^* \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_t = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & 1^* \end{pmatrix}. \quad (2.18)$$

As all variances are constrained to be equal in Case V, they are all fixed to unity

$$\boldsymbol{\mu}_t = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ 0^* \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1^* \end{pmatrix}. \quad (2.19)$$

In order to estimate the means, all thresholds must also be fixed (e.g., to zero).

Furthermore, one factor loading must be fixed in Thurstonian factor models with only one latent trait (e.g.,  $\lambda_1 = 1$ ). This loading constraint would not be necessary if more than one latent trait is included in the model (Maydeu-Olivares & Brown, 2010). Finally, one variance of one latent utility and the variances of the latent traits need to be fixed to unity. These last constraints set the scales of the factor loadings and the unique variances.

Simple Thurstonian models and Thurstonian factor models are used whenever the means of the latent utilities or latent traits are of interest. Prototypical cases would be given by applications of item scaling, where item properties on a latent trait (Maydeu-Olivares & Böckenholt, 2005) or their social desirability (Jansen & Schulze, 2023a), for example, are of interest. In addition to item scaling, trait scaling may also be of interest. However, disregarding test construction studies, item and trait scaling are rather rare in psychological research. In most research efforts, the item properties are not of primary interest, but interest lies much more heavily on individual scores and individual differences. However, scaling methods are important for applications of Thurstonian IRT models which are reviewed and discussed in what follows (for an application and limited discussion, see Jansen & Schulze, 2023a).

#### ***2.2.4.4 Thurstonian Factor Models with Unrestricted Thresholds***

When latent utilities are not of interest, they are fixed to zero and the intercepts in the model are estimated instead. According to Equation (2.6) and (2.12) the latent differences are

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e}, \quad \mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (2.20)$$

The corresponding  $\tilde{n}$  intercepts are defined as  $-\boldsymbol{\gamma}$ . They were constrained for identification of the latent utility means as

$$\boldsymbol{\gamma} = -\boldsymbol{\Lambda}\boldsymbol{\mu}_t. \quad (2.21)$$

Accordingly, the model with unconstrained intercepts is defined by

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}\mathbf{t} + \mathbf{e}, \quad \mathbf{t} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (2.22)$$

The threshold structure is  $\boldsymbol{\tau} = \mathbf{D}\boldsymbol{\gamma}$  and unconstrained as it is a rescaling of  $\boldsymbol{\gamma}$  by  $\mathbf{D}$ .

Thurstonian factor models with unrestricted thresholds are hardly ever of particular interest.

However, this model type is important for the use of Thurstonian IRT models. The

Thurstonian factor models with unrestricted thresholds are considerably less constrained as compared to the Thurstonian factor models. With unconstrained thresholds (and restricted means), the consideration of individual differences would be possible. In a paired comparison design, factor scores can indeed be estimated. However, if only stimuli rankings are used, factor scores estimation is not possible anymore in the present framework. This is due to the non-positive residual variances of the categorical indicators with  $\mathbf{e} = \mathbf{0}$  and  $\boldsymbol{\Omega}^2 = \mathbf{0}$ .

#### 2.2.4.5 Thurstonian IRT Models

The main advantage of the unconstrained factor model is the possibility of a straightforward reparameterized into a first-order model. The reparameterized model can equivalently be expressed as an IRT model (Brown & Maydeu-Olivares, 2011). For instance, Equation (2.22) can be reparameterized to

$$\mathbf{y}^* = -\boldsymbol{\gamma} + \mathbf{A}(\mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = -\boldsymbol{\gamma} + \mathbf{A}\mathbf{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\boldsymbol{\gamma} + \check{\mathbf{\Lambda}}\boldsymbol{\eta} + \check{\boldsymbol{\varepsilon}} \quad (2.23)$$

with  $\check{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}$  and  $\text{cov}(\check{\boldsymbol{\varepsilon}}) = \check{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \boldsymbol{\Omega}^2$ , where  $\check{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda}$  is a  $\tilde{n} \times m$  matrix. Assuming a single latent trait, so that  $m = 1$  and  $n = 3$ , it would be

$$\check{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_3 \\ \lambda_2 - \lambda_3 \end{pmatrix}, \quad (2.24)$$

and with  $m = 3$  and  $n = 3$ , for example, it is

$$\check{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \end{pmatrix}. \quad (2.25)$$

In both cases

$$\check{\Psi}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2 + \psi_3^2 + \omega_3^2 \end{pmatrix}. \quad (2.26)$$

Both models in (2.22) and (2.23) are equivalent, as (2.23) is simply a reparameterization.

Therefore, both models have the same (but reparameterized) tetrachoric correlation matrix

$$\mathbf{P}_z^* = \mathbf{D}(\mathbf{A}\mathbf{A}\mathbf{\Phi}\mathbf{A}' + \mathbf{A}\check{\Psi}^2\mathbf{A}' + \mathbf{\Omega}^2)\mathbf{D} = \mathbf{D}(\check{\mathbf{A}}\check{\mathbf{A}}' + \check{\Psi}^2)\mathbf{D} \quad (2.27)$$

Let  $\Phi(x)$  be a standard normal distribution function at  $x$ ,  $\gamma_l$  the threshold for  $y_l$ ,  $\check{\lambda}_l$  the vector of factor loadings, and  $\check{\psi}_l^2$  the variance of the binary response. Then the item

characteristic function (ICF) for the binary response variable for items  $i$  and  $j$  is given by

$$\Pr(y_l = 1 | \boldsymbol{\eta}) = \Phi\left(\frac{-\gamma_l + \check{\lambda}_l \boldsymbol{\eta}}{\sqrt{\check{\psi}_l^2}}\right). \quad (2.28)$$

It is the ICF of a normal ogive model except that  $\check{\lambda}_l$  and  $\check{\psi}_l^2$  are structured and that the ICFs are not independent (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010).

With only one latent trait, that is  $m = 1$ , Equation (2.28) specifically is

$$\Pr(y_l = 1 | \eta) = \Phi\left(\frac{-\gamma_l + \check{\lambda}_l \eta}{\sqrt{\check{\psi}_l^2}}\right) = \Phi\left(\frac{-\gamma_l + (\lambda_i - \lambda_j)\eta}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}\right). \quad (2.29)$$

However, if  $m > 1$ , then for each comparison it is

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_j \eta_b}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}\right). \quad (2.30)$$

Expressed in intercept and slope notation it follows from

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}, \quad \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}, \quad \beta_j = \frac{\lambda_j}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}} \quad (2.31)$$

that, when  $m = 1$ , it is

$$\Pr(y_l = 1 | \eta) = \Phi(\alpha_l + (\beta_i - \beta_k)\eta). \quad (2.32)$$

whereas with  $m > 1$  it is for each comparison

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b). \quad (2.33)$$

Figure 2.5 gives an example for a Thurstonian IRT model with  $n = 4$  and a single factor  $m = 1$ .

#### 2.2.4.6 Latent Traits Estimation

Following the estimation of the IRT model parameters, latent traits scores for each individual can also be estimated using their pattern of binary outcome responses. Generally, the maximum a posteriori (MAP) estimation is used (Brown & Maydeu-Olivares, 2011), which maximizes the mode of the posterior distribution of the latent traits. When IRT scores are obtained with the MAP method, the posterior test information for each respondent at each point MAP estimate is evaluated. The empirical reliability for the MAP scores can be calculated by

$$\rho = \frac{\sigma^2 - \bar{\sigma}_{error}^2}{\sigma^2} \quad (2.34)$$

where  $\sigma^2$  is estimated using the variance of the MAP scores, and  $\bar{\sigma}_{error}^2$  is estimated by

$$\bar{\sigma}_{error}^2(\hat{\eta}) = \frac{1}{N} \sum_{j=1}^N SE_{estimate}^2, \quad (2.35)$$

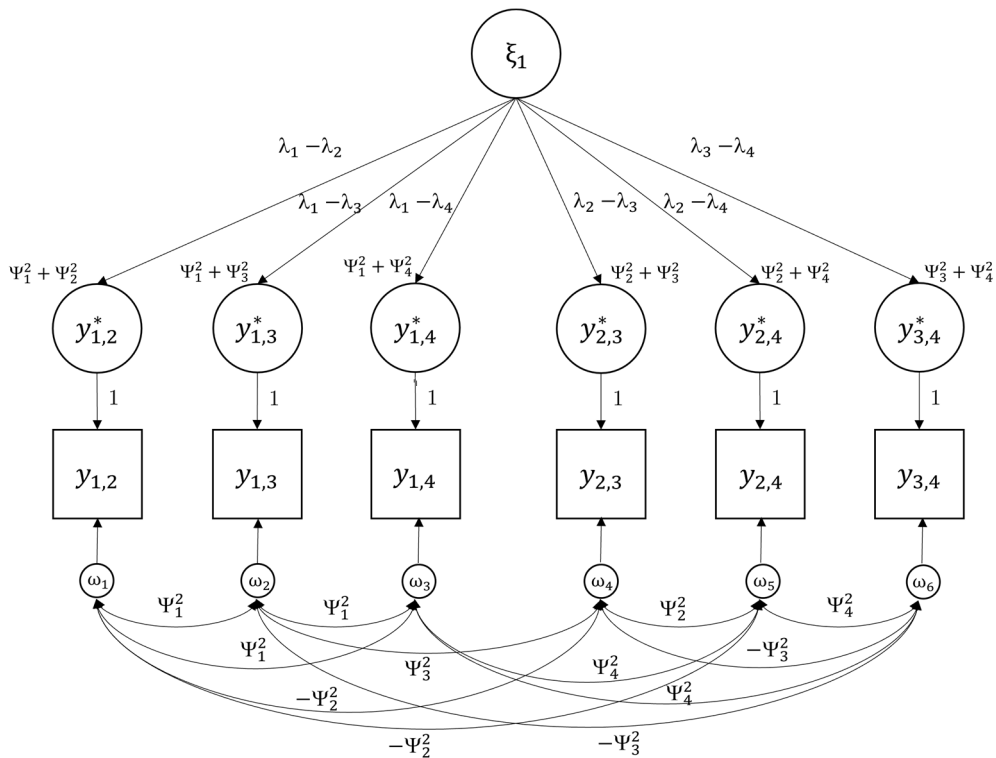
where  $SE_{estimate}^2$  are the squared standard errors of each estimate.

#### 2.2.4.7 Use and Identification of Thurstonian IRT Models

To identify the model, the variances of the latent traits and one error variance (uniqueness) need to be fixed to unity in the reparameterized model. The main advantage of the Thurstonian IRT model is the fact that the error variances are now non-zero and positive even for ranking designs. This allows for the estimation of the latent trait scores. From a

**Figure 2.5**

*Examples of a Covariance Structure of a Thurstonian IRT Model for  $n = 4$  and  $m = 1$ .*



practical viewpoint it is also noted that IRT model estimation is faster than the estimation of Thurstonian factor models via CFA methods.

So far, three types of Thurstonian models have been specified: Firstly, the simple Thurstonian model with correlated latent variables for the items. This model can be used to estimate item properties and test Thurstone's cases. Secondly, Thurstonian factor models that are higher-order models with latent traits (e.g., conscientiousness) as the second-order factors. These models can be used to estimate items or trait properties. Only with paired comparison designs, these models also allow for the estimation of the latent traits and person scores, because the uniquenesses are generally non-zero in this design type. Lastly, Thurstonian IRT models have been specified. These models can be used to estimate person scores in both paired comparison and ranking designs.

A major practical problem for applications of these models is that respondents have to perform many comparisons, even if the number of items is small. The number of necessary comparisons quickly becomes cumbersome and can escalate to an extent that may cause problems with data quality (respondent fatigue, noncompliance etc.). To illustrate, assume an investigation includes only  $n = 15$  items. This already results in  $\tilde{n} = 105$  paired comparisons for every respondent. A typical psychological questionnaire can easily exceed 30 items, which would require a prohibitively large number of not less than 435 paired comparisons. Hence, there is a strong need in practical applications to reduce the number of items, as requiring that many decisions from participants is mentally exhausting and not feasible. Fortunately, a potential solution to this problem was proposed by Brown and Maydeu-Olivares (2011) with the Multidimensional forced-choice format (MFC). The corresponding model will be called Thurstonian multidimensional block model (TMB) in the following.

#### ***2.2.4.8 Thurstonian Multidimensional Block Design***

To date, the MFC format and therefore the TMB design described in this section are by far the most often applied in practice and research of Thurstonian modeling (e.g., Brown et al., 2017; Guenole et al., 2018; Ng et al., 2021; Salgado & Táuriz, 2014). The main idea of the TMB is straightforward: instead of using all possible paired comparisons, only a selection of item blocks is presented. Blocks can contain a number of  $k$  items, that is, a triplet has  $k = 3$ , a quad  $k = 4$ , and so on. The total number of items  $n$  must be divisible by  $k$  so that  $p = n/k$  blocks are presented to a respondent. This could be accomplished with any  $k > 1$ . Moreover, the number of latent traits must be at least  $k$  for the construction of multidimensional blocks. The TMB design is considered to be a generalization of the Thurstonian factor (and IRT) model because the properties of the latter models hold within each block, only moving from one to multiple blocks. However, it will be shown later that it is not a generalization but a simplification.

To date, the TMB design is used and discussed only in the IRT setting as the reparameterized Thurstonian factor model with unrestricted thresholds. However, it is not restricted to the IRT setting and can be used for all of the aforementioned Thurstonian models. The model equations are also identical from a technical perspective but there is a move from the single block to the multiple block perspective.

As an IRT example, consider  $k = 3$  and  $p = 3$  so that  $n = 9$ . In this case, the structured design matrix  $\mathbf{A}$  would be

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad (2.36)$$

while the reparameterized matrix of loadings  $\tilde{\mathbf{\Lambda}}$  is

$$\tilde{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \end{pmatrix} \quad (2.37)$$

and the matrix of structured uniquenesses  $\tilde{\mathbf{\Psi}}^2 = \mathbf{A}\mathbf{\Psi}^2$  is given by





As has been mentioned in the introduction, the use of Thurstone models was mainly motivated by its potential to a) provide faking resistant measures of latent traits that have almost exclusively been measured with self-reports and b) eliminate any rating scale-specific response biases, such as extreme response styles or acquiescence. Beyond these potential benefits, rater biases like severity and leniency as well as halo effects are also expected to be eliminated or at least reduced (Wetzel et al., 2020). It is not entirely clear though, if Thurstone models have indeed lived up to all these promises.

In its current form, Thurstonian modeling cannot entirely eliminate faking (Schulte et al., 2021; Wetzel et al., 2020). Nevertheless, the effect of faking appears to be considerably reduced by using it. In a meta-analysis on faking in big five personality measures, the effect size of faking was found to be large (range 0.11–0.45; Birkeland et al., 2006). A recent meta-analysis of faking with the MFC format found remarkably smaller effect sizes of faking (range 0.00–0.23; Cao & Drasgow, 2019). This might be particularly the case if items presented in one block are successfully matched for their desirability, which is a challenge in itself. Fortunately, the simple Thurstonian and Thurstonian factor models can be applied to estimate items' desirability values (see Jansen & Schulze, 2023a), which in turn can be used for successful item desirability matching. Lastly, to estimate the Fakeability of item blocks, the faking mixture model was proposed (Frick, 2021). A combination of desirability matching and the use of the faking mixture model might be a good test construction strategy to eliminate or at least further reduce the effect of faking on the scores. In sum, there is preliminary evidence for a partial reduction of faking tendencies and some potential for additional developments to further reduce successful faking in Thurstonian modeling applications. However, it should also be noted that direct empirical evidence of a general reduction of response biases with applications of Thurstonian FC models is scarce at best. On the bright side, it also appears at least as good (i.e., valid) as traditional assessment

approaches. However, there also appears to be a “dark side” of the TMB that has not yet attracted sufficient attention to which we will now turn to.

### **2.2.5 Problems with Thurstonian Forced-Choice Models**

There have been occasional reports in the literature in recent years concerning issues with the use of the MFC format and the TMB design but most studies seem to add to the positive overall impression of the design’s potential. The more critical reports focus mainly on the issue of fakeability but also include the use of mixed keyed blocks (e.g., Bürkner, 2022; Bürkner et al., 2019; Schulte et al., 2021). In this section, we will add to a more critical perspective on the current state of the TMB design and its applications. This is done in service of the aim to identify and highlight current limitations and shortcomings on the one hand and point to routes of improvement on the other. Overall, the problems with Thurstonian FC models pointed out here will indicate its limited usefulness so far and lead to the conclusion that some of the results reported within the literature are biased.

A more general and the most limiting factor is the restricted number of comparisons that need to be performed if any of the non-block designs are used. Beyond the already mentioned practical limits imposed by the respondents’ abilities and motivation, it is also important to note that even with limited information estimators, the use of many items can easily exceed the computational resources available (software, memory, and computation power). This constrains the applicability of any of the Thurstonian non-block designs considerably.

As a consequence, the use of the TMB design has been widespread since the number of blocks that can be presented is far less limiting as compared to the number of paired comparisons. Because of its widespread use, the focus will henceforth be put on the TMB design and its problems. In what follows, a ranking model with  $\mathbf{e} = 0$  and  $\mathbf{\Omega}^2 = 0$  is assumed.

The main question in using the TMB design is about which selection of items should be presented in which constellation of blocks, or in other words, what is the design of the assessment? The most used block format is triplets. From a combinatoric point of view, the number of all possible constellations of triplets for a set of  $n$  items is easily found, assuming  $n$  is a multiple of three. Let three items be drawn at a time without replacement and without considering the order. Let  $k$  be the vector of the number of items within each block. The multinomial coefficient provides the number of combinations as

$$C_{n,k} = \binom{n}{k_1, k_2, \dots, k_p} = \frac{n!}{k_1! k_2! \dots k_p!}. \quad (2.39)$$

In a typical design with only triplets, that is all  $k_i = 3$ , the number of combinations is

$$C_{n,k} = \frac{n!}{3!3! \dots 3!} = \frac{n!}{p \times 6}. \quad (2.40)$$

For  $n = 15$  there are  $p = 5$  triplets to be constructed, resulting in more than 43 billion possible ways to construct a test. This number gets considerably lower if we assume that  $m > 2$  and that within each block no trait is represented twice. Assuming  $m = 3$  in the example, then for the first trait there is one combination to sort the five items into five (yet empty) blocks. Fix the order of the blocks for the second trait, as the order of the five items of the first trait is irrelevant. There are  $(n/3)!$  possible combinations to sort the  $n/3$  items of trait two and for each result  $(n/3)!$  combinations to sort the  $n/3$  items of trait three. Taken together, if within each block no trait is considered twice, then for  $n = 15$  and  $m = 3$  there are  $p = 5$  triplets, for which  $5! \times 5! = 120^2 = 14400$  combinations are possible. This illustrates that there are many possibilities to create a test in the framework of the TMB design. There are two main follow-up questions: Is there a best combination and does it even matter which combination is used?

### ***2.2.5.1 Problem 1: Precise Estimation vs. Fakeability***

The first question, namely if there is a single best way to combine a given set of items and, if so, how to identify it, is not easy to answer. From previous studies, we know that there are better and worse ways to combine items into a block design, especially for the latent trait estimation (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). These are rooted in the difference values within or between factors. From Equation (2.29) and Equation (2.30) it can be gathered that both thresholds and especially the loadings of the items have a considerable impact on the ICCs and, therefore, also on the latent trait estimation. For example, if the difference between loadings of two items in a comparison is very small in a one-factor situation (i.e.,  $m = 1$ ), the influence of the factor is small or even zero in the case of equal loadings. To a certain degree, this is also true for comparisons when  $m = 2$ . However, as two different traits are considered, the effect on the two-dimensional surface is small only if the difference between loadings and the difference between the latent traits is simultaneously small. In addition, estimation is also dependent on the size of the correlation between traits (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). All these observations lead to the general recommendation to construct every block with at least one item with a positive and one item with a negative loading. This procedure, as can be expected from the explanation above, results in better trait estimation (less bias and therefore higher reliability) as was shown in several simulation studies (see Brown & Maydeu-Olivares, 2011).

However, this recommendation to construct every block with at least one item with a positive and one item with a negative loading provokes a serious problem: Blocks are more susceptible to faking by design. The recommendation to construct every block with at least one item with positive and one item with negative loading offers respondents an easy choice if they are inclined to respond in socially favorable manner. In the dominance-response

model, one item is strongly more or strongly less desirable compared to the others if the direction of the loading is different. The item that is chosen to be most or least appealing is the one that is most or least desirable. To illustrate, assume a triplet with the following statements for conscientiousness, openness, and emotional stability:

1. I am orderly.
2. I have a rich vocabulary.
3. I get stressed out easily.

The loadings of the first two statements will be positive, the loading of the third statement will be negative. In a classical faking instruction for a job interview, the presentation of such a block almost certainly guarantees the third statement to be chosen as “least like me”. Two of the three paired comparisons that can be derived from the block are prone to faking and only one comparison yields useful information about the traits in question.

As a consequence, constructing MFC tests that have differently keyed items within each block is *not* advisable and may be counterproductive with respect to the validity of the assessment (see also Bürkner et al., 2019; Schulte et al., 2021). The construction principle conflicts with the initial goal to reduce the effect of response biases and particularly faking when these response distortions are relevant issues in the given situation.

#### ***2.2.5.2 Problem 2: Estimation of Specific (Block) Designs***

The second question addressed here is: Does it matter which block combination is used? For the Thurstonian IRT models, where the latent traits are of interest, it is generally assumed that it does *not* matter which combination is used, since specific items are indicators of their respective traits. In order to show and illustrate that this is not true, the assumptions on which this conclusion rests have to be specified first, so they will be explicated next.

Given are

1. A set of  $n$  items  $t_1, \dots, t_n$

2. A set of  $m$  latent traits  $\eta_1, \dots, \eta_n$
3. A mapping from  $t_1, \dots, t_n$  to  $\eta_1, \dots, \eta_n$ , that is, a matrix of loadings.

As a consequence of 2 and 3, the simple Thurstonian models are disregarded here, only Thurstonian factor and IRT models are considered.

When a specific MFC test is constructed and the focus is on the estimation of the specific block design, then a choice between a large set of (several thousand if not billions, see above) possible block designs that could be used for an assessment has to be made. The recommendation to only construct multidimensional block designs with at least one item with positive and one item with negative loadings indeed effectively reduces the set of designs to choose from, but was already dismissed in the previous subsection.

But what are the consequences of (randomly) choosing just any of the many admissible block designs? The most critical consequence of estimating a specific block design is that the specific MFC test (the operationalization of the assumed theoretical model) determines the test model that the data is applied to in a confirmatory sense (the specific block design), and not the other way around. To illustrate this intricate issue with a simplified example, the performance of a simulation study is considered. For illustrative purposes the following setting may be given:  $n = 9$ ,  $m = 3$  and  $p = 3$ . Data for one respondent will be generated from the Thurstonian factor model (Equation (2.13)), with

$$\Lambda = \begin{pmatrix} -.6 & 0 & 0 \\ .7 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & -.6 & 0 \\ 0 & .6 & 0 \\ 0 & .8 & 0 \\ 0 & 0 & .8 \\ 0 & 0 & -.6 \\ 0 & 0 & .4 \end{pmatrix} \quad (2.41)$$

and  $\boldsymbol{\mu}_t = (-.1, .3, .2, -.2, .3, -.1, .2, -.1, 0)$  as well as  $\boldsymbol{\eta} = (-.5, .1, .6)$ . The error terms  $\boldsymbol{\varepsilon}$  and  $\mathbf{e}$  are ignored here for simplicity. First, consider a multidimensional block design with exactly one item having a negative loading in each block as recommended in the literature (e.g., Brown & Maydeu-Olivares, 2011). The design matrix  $\mathbf{A}$  for such a design could be

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}. \quad (2.42)$$

This corresponds to blocks  $B_1 = (1,5,9)$ ,  $B_2 = (2,4,7)$ , and  $B_3 = (3,6,8)$  for which follows

$$\mathbf{y}^* = \mathbf{A}(\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta}) = \begin{pmatrix} -.16 \\ -.04 \\ .12 \\ .21 \\ -.73 \\ -.94 \\ -.03 \\ .41 \\ .44 \end{pmatrix} \Rightarrow \mathbf{y} = \begin{pmatrix} y_{15} \\ y_{19} \\ y_{59} \\ y_{24} \\ y_{27} \\ y_{47} \\ y_{36} \\ y_{38} \\ y_{68} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}. \quad (2.43)$$

Of course, the estimation of the model and calculation of MAP scores is not possible for one respondent only. However, we can focus on the ipsative ranking of the traits within a respondent. From Equation (2.43) it can be gathered that for the first block, item 1 (trait 1) is second in both paired comparisons, item 5 (trait 2) is chosen first in both paired comparisons, and item 9 (trait 3) is chosen over item 1 but not item 5. As a result, the ranks for the first block are  $(\eta_{11} = 3, \eta_{21} = 1, \eta_{31} = 2)$ . For the second block it is  $(\eta_{12} = 2, \eta_{22} = 3, \eta_{32} = 1)$ , and for the third it is  $(\eta_{13} = 2, \eta_{23} = 1, \eta_{33} = 3)$ , respectively. If we rank the traits over blocks, the



Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and  
Issues

mean ranks are 2.33 for the first trait, 1.67 for the second trait, and 2 for the third. Hence, the order of traits would be  $(\eta_2, \eta_3, \eta_1)$ . Now let's assume a different block design is given in the same situation that may have alternatively be chosen:

$$\mathbf{A}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix} \quad (2.44)$$

which corresponds to blocks  $B_1 = (1,6,9)$ ,  $B_2 = (2,4,7)$ , and  $B_3 = (3,5,8)$ . Items are only slightly shuffled as compared to the first block design. More specifically, only items 5 and 6 are switched. It follows

$$\mathbf{y}^* = \begin{pmatrix} .22 \\ -.04 \\ -.26 \\ .21 \\ -.73 \\ -.94 \\ -.41 \\ .41 \\ .82 \end{pmatrix} \Rightarrow \mathbf{y} = \begin{pmatrix} y_{16} \\ y_{19} \\ y_{69} \\ y_{24} \\ y_{27} \\ y_{47} \\ y_{35} \\ y_{38} \\ y_{58} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}. \quad (2.45)$$

Again, if we rank the traits over blocks, the mean ranks are 2 for the first trait, 2.33 for the second, and 1.67 for the third trait and the order would therefore be  $(\eta_3, \eta_1, \eta_2)$ . Let's assume yet another different block design in the same situation that may have alternatively be chosen:

Issues

$$\mathbf{A}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \end{pmatrix} \quad (2.46)$$

which corresponds to blocks  $B_1 = (1,5,7)$ ,  $B_2 = (2,6,8)$ , and  $B_3 = (3,4,9)$ . As can be seen, even more items have been shuffled across blocks but it is just another possible block design in the given situation. Now it follows

$$\mathbf{y}^* = \begin{pmatrix} -.16 \\ -.48 \\ -.32 \\ -.03 \\ .41 \\ .44 \\ .21 \\ -.29 \\ -.50 \end{pmatrix} \Rightarrow \mathbf{y} = \begin{pmatrix} y_{15} \\ y_{17} \\ y_{57} \\ y_{26} \\ y_{28} \\ y_{68} \\ y_{34} \\ y_{39} \\ y_{49} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \quad (2.47)$$

If we rank the traits within each block again, the mean ranks are 2.33 for the first, 2 for the second, and 1.67 for the third trait and the order would be  $(\eta_3, \eta_2, \eta_1)$ .

For a full design (all 36 paired comparisons) the ipsative order would be  $(\eta_3, \eta_2, \eta_1)$ , which corresponds to  $\boldsymbol{\eta} = (-.5, .1, .6)$ . What this elaborate example shows is that the simulated pattern of paired comparisons actually yields different results for the simulated preferences of traits for all three block designs. Note that this is true even without considering error terms. The only difference between the three designs is the block design as represented by matrix  $\mathbf{A}$ , all other fundamental parameters are identical. That is, *ceteris paribus*, the

second trait could be modeled as the most preferred, the least preferred, or in between just by choosing a specific design. Of course, this illustration is limited to a specific situation with  $n = 9$ ,  $m = 3$  and  $p = 3$ . Multidimensional blocks without the loading recommendation or unidimensional paired comparisons are not considered.

Nevertheless, the example shows that the choice of a specific block design can indeed matter and it can result in remarkable differences (not only) in simulated data. This implies that simulated data can already be biased. Given these circumstances, the result almost certainly is that the estimation of the model and the calculation of factor scores can also be biased. In order to illustrate this with an empirical example, data from Jansen and Schulze (2023a) is reanalyzed.

In this study, 15 positively keyed adjectives that are all intended to assess conscientiousness were used and the social desirability scores of these items were estimated. Participants were presented with the entire set of 105 possible paired comparisons and a block of all 15 items. This amounts to the fact that data for the full design is available. In the present reanalysis data is considered for ranking. If three different designs were chosen, the results differed not only between the three designs but also in comparison to the parameter estimation by the full design. Consider the arbitrarily chosen designs: Blocks are given by

1. items [1,2,3], [4,5,6], [7,8,9], [10,11,12], and [13,14,15]
2. items [1,5,6], [2,4,11], [7,12,15], [3,8,10], and [9,13,14]
3. items [1,9,14], [4,8,12], [2,5,7], [6,10,15], and [3,11,13]

The relative bias of the factor loadings on social desirability and item utilities of a Thurstonian factor model are given in Table 2.1. As can be seen, the estimated factor loadings and item utilities from the block designs deviate strongly from the estimates based on the full design. These relative biases appear to be beyond acceptable. Additionally, the correlation between factor scores (estimated with the corresponding Thurstonian IRT models)

Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues

of the block-designs and the full design are  $r_{Fd1} = -.42$ ,  $r_{Fd2} = .91$ , and  $r_{Fd3} = .87$ , respectively.

Interestingly, the block design with the lowest mean relative bias for loading and utility estimates is the design with the lowest correlation for estimated factor scores.

**Table 2.1**

*The Relative Bias of the Three Block Designs Compared to the Full Design.*

Item / Parameter	Loadings			Utilities		
	D1	D2	D3	D1	D2	D3
I01*	0.00	0.00	0.00	0.00	0.00	0.00
I02	-0.08	-0.69	-0.24	0.25	-2.68	-1.33
I03	-0.64	-0.70	-0.37	0.45	-0.55	-1.13
I04*	-0.48	-0.48	-0.48	-1.00	-1.00	-1.00
I05	-0.53	6.60	-0.17	-2.61	10.55	-1.58
I06	-0.76	9.35	-0.54	-2.04	10.56	-0.83
I07*	-0.16	-0.16	-0.16	-1.00	-1.00	-1.00
I08	1.12	-1.14	-1.04	-1.64	-1.28	-3.71
I09	-0.68	-0.57	-0.08	-2.11	-3.16	0.04
I10*	-0.58	-0.58	-0.58	-1.00	-1.00	-1.00
I11	-0.61	-0.45	-0.45	-0.49	-0.75	-0.71
I12	-0.61	-0.20	-0.37	-0.66	-0.76	-1.52
I13*	-0.52	-0.52	-0.52	-1.00	-1.00	-1.00
I14	0.03	-0.74	0.02	18.73	20.68	-0.52
I15	0.24	-0.38	-1.07	-1.63	-1.01	-0.80
Mean	-0.28	0.62	-0.40	0.28	1.84	-1.07

*Note.* D1, D2, and D3 denote designs 1, 2, and 3. Values with an Asterix \* are fixed values.

To mitigate or even remove the influence of the choice of a specific block design, the possibility of the many different designs that can be focused on needs to be eliminated. The solution to this appears to be almost trivial: The one (and only) design that should be in focus, especially for data generation and simulation studies, must be the design where all paired comparisons are considered (i.e., the full design). Given any item set and any set of traits, the full design is always unique. As a consequence, data should be generated with the full design in any simulation study with Thurstonian models. Previous simulation studies did generate data from the specific block designs, however. The block designs were tested (and compared) on the basis of the resulting data (e.g., Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2021). The same is true for any simulation study with the R package *thurstonianIRT* (Bürkner, 2019) because it also implements the simulation by each specific design. As we have illustrated, chances are high that these simulation studies provide biased results. An update on the results reported in previous simulation studies and further evidence that the simulations were indeed biased is given by Jansen and Schulze (2023b).

### ***2.2.5.3 Problem 3: Identification Constraints in Block Designs***

The reanalysis in the previous subsection also reveals another issue: Identification constraints need to be applied to every block, thereby setting different scales for the parameters. As stated before, the same identification constraints have to be used to identify the TMB design as for other Thurstonian models. However, the constraints need to be applied to every block. While the chosen identification constraints are statistically irrelevant (see Maydeu-Olivares & Böckenholt, 2005; Maydeu-Olivares & Brown, 2010), they are highly relevant for the specific estimates as they set the scale. This can be illustrated with a non-IRT TMB design. Assume that the focus is on estimating the means of the latent utilities of the items. Also assume  $n = 9$ ,  $m = 3$ ,  $p = 3$ , (again)  $\boldsymbol{\mu}_t = (-.1, .3, .2, -.1, .2, .1, .2, -.1, 0)$ , and  $\Lambda$  from equation (2.41). Then with blocks  $B_1 = (1,5,7)$ ,  $B_2 = (2,6,8)$ , and  $B_3 = (3,4,9)$  the

## Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and Issues

---

question is which of the mean latent utilities will be fixed to what value. A common approach would be to fix, for example, the last item per block to  $\mu_i = 0$ . Suppose parameter estimation would be perfect, then the result would be  $\boldsymbol{\mu}_{t_{est}} = (-.3, .4, .2, -.1, 0, .2, 0^*, 0^*, 0^*)$ , where the values with an Asterix \* are fixed values. In the original data, we had  $\mu_3 = .2$ ,  $\mu_5 = .2$  and  $\mu_6 = .1$  which corresponds to items 3 and 5 having the same and item 6 having a smaller mean utility. In contrast, the example estimates are  $\mu_3 = .2$ ,  $\mu_5 = 0$  and  $\mu_6 = .2$ , which corresponds to items 3 and 6 having the same and item 5 having a smaller mean utility.

The core problem is, that no reference point for any of the blocks exists. Without any further information, it is impossible to estimate all latent utility means simultaneously. Arbitrarily chosen identification constraints severely limit the use of the estimates. When identification constraints are arbitrarily chosen, they enforce an equivalence between two or more parameters that would only be valid, if further information was available to justify the equivalence. This is probably a rare rather than a common case in applications. Again, it is referred to Jansen and Schulze (2023a), where the goal was to estimate social desirability values for each item. With the estimated values from a full design, the matching of similar or equally desirable items per block was possible. This is the case because only one item utility needed to be fixed and the other utilities are estimated in relation to the fixed parameter. Assume a block design would be used. The arbitrary identification constraints within each block would identify the design. However, knowledge about the relationship between the latent utility means that are fixed for identification (differences and dispersion) is needed to justify the constraints. Fixing one latent utility mean per block to 0, amounts to the theoretical assumption that these fixed latent utility means are indeed equal. If the assumption is untrue, then using the constraints results in invalid estimates for between block relations. As a consequence, the estimated parameters have very limited usefulness. Furthermore, the

procedure can result in biased estimation for most model parameters. It can be gleaned from the analyses as presented in Table 2.1, for example, that the loadings and utilities are biased except for the parameters of the first item, which was used for identification constraints also in the full design.

#### **2.2.5.4 Problem 4: Estimation of Reliability and Recovery of Latent Traits**

In empirical studies the true latent trait scores are practically never known. The recovery and reliability of these scores can only be estimated using the respondents' pattern of binary outcome responses. As binary indicators are used and an IRT setting is given, with limited information estimators, factor scores are best estimated using the MAP estimation.

The *actual recovery* is defined as the correlation between the true trait scores and their estimates. The *empirical recovery* can be calculated as stated in Equations (2.34) and (2.35) by using the square of the reliability. Previous simulation studies (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010) indicate that the recovery of the latent traits is only reliable if positively and negatively keyed items are used within blocks. For designs with items keyed in only one direction the estimation of the empirical recovery is not recommended by Brown and Maydeu-Olivares (2011). To understand why, it is necessary to acknowledge that the IRT factor scores are dependent on the test information function  $\mathbf{I}^j(\boldsymbol{\eta})$  for trait  $j$ . The test information function is computed by using the item information functions  $\mathbf{I}_l^j(\boldsymbol{\eta})$ , where the  $l$  items are the binary outcomes. The information provided by one binary outcome for two traits  $a$  and  $b$  is (see Ackerman, 2005; Brown & Maydeu-Olivares, 2011)

$$I_l^a(\eta_a, \eta_b) = \frac{[\beta_i - \beta_k \text{corr}(\eta_a, \eta_b)]^2 [\phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b)]^2}{P_l(\eta_a, \eta_b) [1 - P_l(\eta_a, \eta_b)]} \quad (2.48)$$

and

$$I_l^b(\eta_a, \eta_b) = \frac{\left[-\beta_k + \beta_i \text{corr}(\eta_a, \eta_b)\right]^2 \left[\phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b)\right]^2}{P_l(\eta_a, \eta_b) \left[1 - P_l(\eta_a, \eta_b)\right]}, \quad (2.49)$$

where  $\alpha_i$  and  $\beta_i$  are the intercepts and slopes of the binary variables and the items, respectively, and  $P_l(\eta_a, \eta_b) = P(y_l = 1 | \eta_a, \eta_b)$ . The item information depends on the difference between slopes  $\beta_i$  which is small if the difference between loadings is small (see Equation (2.31)). The standard errors of the MAP estimates are estimated by the reciprocals of the square root of the posterior test information  $\mathbf{I}_p^j(\boldsymbol{\eta})$ :

$$SE(\hat{\eta}_j) = \frac{1}{\sqrt{\mathbf{I}_p^j(\boldsymbol{\eta})}}. \quad (2.50)$$

An additional problem is that the empirical recovery overestimates the actual recovery, especially in cases where the difference between loadings is small (see also Yousfi, 2019). However, Brown and Maydeu-Olivares (2011) provided evidence from a simulation study that the empirical recovery underestimates the actual recovery, or is close to it. We suspect that this discrepancy is also a result of the simulation from the block designs and not the full design (see previous subsection).

To illustrate, a small simulation study for full designs was conducted. We simulated 100 data sets with 1000 respondents from full Thurstonian factor models (Equation (2.13)) with  $m = 3$  uncorrelated factors and six items per factor ( $n = 18$ ). The first design included only items with positive item loadings, which were drawn from a uniform distribution at the interval  $[0.3, 0.9]$ . The intercepts were drawn from the interval  $[-1, 1]$ . The second design had the same specifications, only half of the loadings were negative. The results on the MAP scores and the correlation with the true scores are given in Table 2.2.

As can be seen in Table 2.2, the actual recoveries are relatively low for designs where items load only in one direction and reliabilities are not acceptable. In designs including



Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and  
Issues

**Table 2.2**

*Results for the Latent Trait Recoveries and Reliabilities of the Simulation Study.*

	Actual			Empirical		
	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
<b>Recovery</b>						
+	.722	.707	.670	.818	.817	.803
+/-	.876	.860	.838	.889	.886	.881
<b>Reliability</b>						
+	.521	.499	.449	.669	.668	.644
+/-	.768	.739	.702	.790	.785	.775

*Note.* The actual recovery is computed as the correlation between true scores and MAP score estimates. The empirical recovery is calculated using the square root of Equation (2.34). + = positive loadings only, +/- positive and negative loadings.

items loading in both directions, the information provided by each paired comparison is larger and latent trait recovery is better. Additionally, we see that the empirical recovery overestimates the actual recovery, especially if only positively keyed items are used.

In sum, it was illustrated that latent trait recovery is problematic and generally seems to be overestimated in Thurstonian FC modeling. Additionally, in model designs where only blocks are used, this problem is aggravated. Past simulation studies already indicated lower reliability in Thurstonian modeling compared to rating scale reliabilities (e.g., Wetzel & Frick, 2020). Lastly, in IRT settings, the reliability of trait scores depends also on the respondents' location on the traits.

### **2.2.6 Discussion**

The present paper provided an overview of the status and issues with current linear factor analytic Thurstonian FC models and their variants. The main reasons why these models have been introduced is because they are connected to concerns about the validity of responses in the assessment of constructs from the realm of typical psychological behavior. Threats to the validity of responses can arise from unintentional response distortions and biases as well as intentional response tampering such as faking. The use of this class of models is associated with the expectation that they provide solutions to these longstanding problems in psychological assessment. Recent developments of Thurstonian FC models also go beyond an adaptation of Thurstone's original models used in psychophysics and offer modern model testing and parameter estimation. Generally, these models are suitable for flexible modeling binary responses from FC questionnaires with items from a dominance-response perspective. Thurstonian FC modeling can focus on the estimation of latent utilities of items or the estimation of latent factor scores of respondents, thereby providing a framework that includes both item and person scaling. The models encompass both CFA and IRT procedures for these purposes. With software templates and packages available, the estimation can be straightforwardly implemented and, with limited information estimation, is fast, especially compared to full information maximum likelihood estimation. All these highly attractive features of Thurstonian FC models indeed justify the rise in publications and applications with this method in recent years.

As was elaborated in this paper, though, it is not entirely clear yet whether the use of Thurstonian FC models as they are currently used actually lead to enhanced validity of the assessments. The potential benefits are also accompanied by a number of theoretical and practical issues and problems that threaten or at least hamper the realization of the benefits. One of the more obvious difficulties for the use of the Thurstonian FC models is that the use

of the full designs, which is where all possible paired comparisons are needed, is infeasible.

This is true for both the estimation of the models as well as the respondents that need to perform all paired comparisons. Hence, the design of the assessment is one of the major topics of research on Thurstonian FC models. Some proposals to make the use of Thurstonian FC models feasible with triplets, for example, have already been made and used in applications. In the present paper, four problem areas that are given with the proposed solution were addressed and scrutinized. Some of these problems have implications mainly for the practical use of the Thurstonian FC models, others also pertain to the design of simulation studies and the interpretation of their results. For example, it was shown that the MFC format and its constraint to multidimensional blocks, and the estimation with the TMB designs, can lead to biased results. Also, the recommendation to use both positively and negatively keyed items within each block results in potentially highly fokable blocks and does, therefore, contradict the main advantage to potentially reduce faking and response biases (see also Bürkner, 2022; Bürkner et al., 2019).

Given these results, a change in recommendations for the assessment design with Thurstonian FC model is derived, namely *not* to use differently keyed items in blocks when the MFC format or any other block format is used. It should be noted, though, that following this recommendation is expected to lead to responses less prone to faking but at the cost of less reliable estimations (e.g., Brown & Maydeu-Olivares, 2011; Bürkner, 2022; Maydeu-Olivares & Brown, 2010; Schulte et al., 2021).

In addition, this also draws attention to the fact that fakeability and/ or desirability of items and item blocks should explicitly be accounted for. It can be assumed that blocks with items that have different scores on a desirability scale are easily fokable, even when the items in a block are keyed in the same direction. If faking is of concern when using the Thurstonian FC models, as a first step it is therefore recommended to carefully estimate the desirability

scores of items or item blocks before assembling items into blocks. This can be done with the simple Thurstonian or Thurstonian factor model (e.g., Jansen & Schulze, 2023a) or using a rating scale approach (e.g., Wetzel & Frick, 2020). A second step could be to use the recently introduced faking mixture model, to estimate the Fakeability of item blocks (Frick, 2022) after these blocks were matched for desirability.

With respect to simulation studies with Thurstonian FC models, the issues identified here lead the recommendation that simulated data should always be derived from the full design. If any Thurstonian block designs are used, paired comparisons that are simulated but not of interest can just be discarded. However, it was shown that the use of any TMB is problematic and leads to biased results. This is also and especially true since for each block arbitrary identification constraints must be applied. To overcome this problem, it is recommended to apply the identification constraints in relation to one another, or to establish the relation or linkage between the blocks by other justifiable means (such a linkage is proposed by Jansen & Schulze, 2023b).

Overall, the Thurstonian FC models are a very important recent psychometric development for the assessment of typical psychological behavior. The fact that some issues and problem areas as explicated in this paper still prevail for the models in their current status does not imply at all that the use of the models would be not advisable. Research on the Thurstonian FC models is still a vibrant field and it appears to be quite reasonable to expect solutions, or at least significant improvements, in the problem areas as outlined in this paper. Until then, careful interpretation of results from the use of Thurstonian FC models taking the problems areas into account appears to be advisable.

### 2.2.7 References

- Ackerman, T. A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. Mc Ardle (Eds.), *Contemporary psychometrics* (pp. 3–26). Erlbaum.
- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56.  
<https://doi.org/10.1111/j.2044-8325.1996.tb00599.x>
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26.  
<https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335.  
<https://doi.org/10.1111/j.1468-2389.2006.00354.x>
- Böckenholt, U. (1993). Applications of Thurstonian models to ranking data. In *Probability models and statistical analyses for ranking data* (pp. 157–172). Springer.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.
- Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. <https://doi.org/10.1007/s11336-014-9434-9>
- Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148.  
<https://doi.org/10.1177/1094428116668036>

- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502.  
<https://doi.org/10.1177/0013164410375112>
- Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147.  
<https://doi.org/10.3758/s13428-012-0217-x>
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52.  
<https://doi.org/10.1037/a0030641>
- Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software, 4*(42), 1662–1663.
- Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854. <https://doi.org/10.1177/0013164419832063>
- Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. <https://doi.org/10.1037/apl0000414>
- Carey, K. B., Neal, D. J., & Collins, S. E. (2004). A psychometric analysis of the self-regulation questionnaire. *Addictive Behaviors, 29*(2), 253–260.  
<https://doi.org/10.1016/j.addbeh.2003.08.001>
- Cochrane-Brink, K. A., Lofchy, J. S., & Sakinofsky, I. (2000). Clinical rating scales in suicide risk assessment. *General Hospital Psychiatry, 22*(6), 445–451.  
[https://doi.org/10.1016/S0163-8343\(00\)00106-7](https://doi.org/10.1016/S0163-8343(00)00106-7)
- Coombs, C. H. (1960). A theory of data. *Psychological Review, 67*(3), 143–159.  
<https://doi.org/10.1037/h0047773>

- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475–494.
- Frick, S. (2022). Modeling Faking in the Multidimensional Forced-Choice Format: The Faking Mixture Model. *Psychometrika, 1*–22. <https://doi.org/10.1007/s11336-021-09818-6>
- Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research, 1*–29. <https://doi.org/10.1080/00273171.2021.1938960>
- Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment, 25*(4), 513–526. <https://doi.org/10.1177/1073191116641181>
- Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49*(5), 446–450. <https://doi.org/10.1016/j.paid.2010.04.015>
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252. <https://doi.org/10.1037/h0045996>
- Jansen, M. T., & Schulze, R. (2023a). *Item scaling of social desirability using conjoint measurement: A comparison of ratings, paired comparisons, and rankings*. Manuscript in preparation.
- Jansen, M. T., & Schulze, R. (2023b). *The Thurstonian linked block design: Improving Thurstonian modeling for paired comparison and ranking data*. Manuscript submitted.

- Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences, 123*, 229–235. <https://doi.org/10.1016/j.paid.2017.11.031>
- Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika, 64*(3), 325–340.  
<https://doi.org/10.1007/BF02294299>
- Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10*(3), 285–304.  
<https://doi.org/10.1037/1082-989X.10.3.285>
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioural Research, 45*(6), 935–974.  
<https://doi.org/10.1080/00273171.2010.531231>
- 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*(2), 222–248. <https://doi.org/10.1177/1094428105275374>
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551–560. <https://doi.org/10.1007/BF02293813>
- Muthén, L. K., & Muthén, B. O. (1998-2022). Mplus User's Guide. Eighth Edition. Muthén & Muthén.
- Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment, 103*(2), 224–237.  
<https://doi.org/10.1080/00223891.2020.1739056>



Study 1: Linear Factor Analytic Thurstonian Forced-Choice Measurement: Current Status and  
Issues

---

- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Routledge.
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford Press.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3–32. <https://doi.org/10.1177/01466216000241001>
- Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. <https://doi.org/10.1080/1359432X.2012.716198>
- Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats?. *Educational and Psychological Measurement, 81*(2), 262–289. <https://doi.org/10.1177/0013164420934861>
- 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*(3), 184–203. <https://doi.org/10.1177/0146621604273988>
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703–742. <https://doi.org/10.1111/j.1744-6570.1991.tb00696.x>

- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *3*(4), 529–554.
- Thurstone, L. L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology*, *14*(3), 187–201. <https://doi.org/10.1037/h0070025>
- Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. <https://doi.org/10.1037/pas0000781>
- Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales. *European Journal of Psychological Assessment*, *36*(4), 511–515. <https://doi.org/10.1027/1015-5759/a000609>
- Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*(1), 79–92. <https://doi.org/10.1348/000711099158973>
- Yousfi, S. (2019). Person parameter estimation for IRT Models of forced-choice data: Merits and perils of pseudo-likelihood Approaches. In *The Annual Meeting of the Psychometric Society* (pp. 31–43). Springer.
- Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569–590. <https://doi.org/10.1177/1094428119836486>
- Ziegler, M., MacCann, C., & Roberts, R. (2012). *New perspectives on faking in personality assessment*. Oxford University Press.

### **3. Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data**

Jansen, M. T., & Schulze, R. (2023c). *The Thurstonian linked block design: Improving Thurstonian modeling for paired comparison and ranking data*. Manuscript submitted for publication in Psychometrika.

#### **3.1 Summary**

##### **3.1.1 Thurstonian Models and Non-Comparability Between Independent Blocks**

The preceding manuscript focused on problematic features but also hinted at a potential solution for most of the shortcomings described. The following manuscript picks up where the last one ended. The relevant Thurstonian models (factor and IRT model) and the block design are presented again. It is described how most of the issues arise from the independence of the blocks. The emphasis is on the rank of  $\mathbf{A}$ , which in a full design is  $n - 1$ . For the block designs proposed by Brown and Maydeu-Olivares (2011), the rank of  $\mathbf{A}$  is  $(k - 1)p$ , which corresponds to the need for one set of identification constraints per block. If the block design could be altered so that  $\mathbf{A}$  has a rank of  $n - 1$ , then this problem could be solved and estimates would be comparable on the same scale.

##### **3.1.2 The Thurstonian Linked Block Design and Simulation Study**

This is done by the Thurstonian linked block design. Each initial block is linked to every other block by adding linking blocks that contain items from different initial blocks. A full linking is done if the rank of  $\mathbf{A}$  is  $n - 1$ .

A simulation study is conducted to compare the unlinked block with the linked block design. Additionally, as there are more blocks in the linked block design by definition, which should contribute to a larger amount of information, some partially linked block designs are also used. This way, more blocks yield more information, but the incomplete linking does not

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

solve the problems stated before. Simulation study parameters include the number of traits (one, three, and five), the number of items per trait (6, 12, and 18), and the correlation between traits (uncorrelated or correlated). For all combinations, a data set of 2000 respondents is simulated with 1000 repetitions. All items were positively keyed, hence only same-keyed blocks were used. For data sets with a maximum of 18 items, the full design is also estimated. For all data sets five different block designs are analyzed. For all data sets and all block designs, the Thurstonian factor and IRT models are both estimated to obtain results on item and person parameters.

### 3.1.3 Results

Results show that in general, the more items a test has, the more accurate the results are. Furthermore, partially linked block designs cannot overcome the mentioned problems, as expected, due to the lack of full linking. Overall, unlinked block designs yield lower parameter recovery, as well as lower convergence rates, and larger than expected empirical rejection rates compared to linked block designs. Multidimensional blocks perform best, while unidimensional blocks and blocks matched by loadings perform worst. However, for linked block models, results are mostly satisfactory, linked block designs outperform unlinked designs by far. The results and some limitations are discussed. The most important limitations are the restriction to use only positively keyed items and using large sample sizes of 2000 respondents, which limits the transfer to real empirical data.

In the appendix, it is proven and discussed how the number of redundancies among the thresholds and tetrachoric correlations is determined. This is not trivial, as the result for a full design (Maydeu-Olivares, 1999) is generalized to any FC design, whether linked or unlinked, and for any size of item set or blocks. The result is that there is no closed form solution, but the number of redundancies is still easily determinable.

### 3.2 Manuscript


#### **The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data**

Markus T. Jansen and Ralf Schulze

University of Wuppertal

#### **Author Note**

Markus T. Jansen  <https://orcid.org/0000-0002-5162-4409>

Ralf Schulze  <https://orcid.org/0000-0001-5780-8973>

Supplement material available at

[https://osf.io/jd4uc/?view\\_only=2b8edd3a07a74a07b49f30ad1c36a3ce](https://osf.io/jd4uc/?view_only=2b8edd3a07a74a07b49f30ad1c36a3ce)

The author(s) declared no conflicts of interest with respect to the authorship or the publication of this article. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence concerning this article should be addressed to Markus Thomas Jansen, Institute of Psychology, University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany. E-mail: [mjansen@uni-wuppertal.de](mailto:mjansen@uni-wuppertal.de)

### 3.2.1 Abstract

Thurstonian forced-choice modeling is a powerful tool for the estimation of item and person parameters. It has most often been used in psychological research to assess constructs from the realm of typical behavior (e.g., personality constructs) in situations where response biases or faking may be an issue. This is due to the expectation that it effectively helps to mitigate or even eliminate the detrimental effects of such response distortions on the validity of psychological measures. However, critical issues with Thurstonian modeling in general and the multidimensional forced-choice (MFC) format in particular have recently surfaced that call its utility into question. As a remedy, an adaptation and generalization of the MFC format is proposed: the Thurstonian linked block (TLB) design. The TLB design is flexible and can be applied in both person- or item-centered situations. Comprehensive simulation studies are conducted to investigate the bias in parameter estimation and latent trait recovery of the new linked and traditional unlinked block designs in both the factor analytic and IRT settings. The results show that unlinked block designs produce biased results. In contrast, TLB designs yield unbiased parameter estimates, somewhat better latent trait score recovery and almost accurate model rejection rates. Therefore, it is advisable to replace traditional unlinked block designs with TLB designs in research and applications of Thurstonian forced-choice modeling.

*Keywords:* Thurstonian modeling, forced-choice format, item response theory, structural equation modeling, block linking

### **3.2.2 The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data**

In almost any area of psychology constructs are of interest whose assessment require introspection from respondents in order for them to provide information on their (latent) construct score. The result of the introspection is typically just reported by the respondent, though this self-report may take on different forms. A very typical and highly popular form of self-report are Likert-type items or similar forms of written responses, even if the assumption of a perfect and bias-free process of introspection was true, the self-report can be subject to many response distortions, both intentional or unintentional. Accordingly, common problems with self-reports via rating scales, for example, include their susceptibility to a) response biases (e.g., the tendency to very strongly agree- or disagree irrespective of item content [extreme responding]), b) social desirability bias (e.g., Cronbach, 1946; Jackson & Messick, 1958; Paulhus, 2002), and faking (Ziegler et al., 2012). In sum, there is a high demand for valid assessment procedures for constructs typically assessed with self-reports on the one hand and strong concerns about their validity because of potential response distortions on the other. Forced-choice (FC) assessment procedures may offer a solution for this conundrum by reducing or eliminating potential response distortions in self-reports. This is the main reason why these procedures are highly attractive for almost any field of psychology and there has recently been a surge of new developments and applications of this type of assessment. How are FC assessment procedures expected to realize these benefits?

The central feature of the FC method is the requirement for the respondents to make a choice between two or more stimuli (e.g., statements about an individual's behavior) according to a predefined criterion (e.g., best description of oneself). An example with a block of two options to choose from can be seen in panel A – Paired Comparison in Figure 3.1. Obviously, statements are not rated on a scale individually, as would the case with

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Likert-type items. Instead, the stimuli are compared and ranked by the respondent according to the criterion. Whereas it is possible when using ratings to maximally agree with both statements, this is impossible in a FC setting. Thereby, some response biases are eliminated with the FC format (e.g., acquiescence bias).

One of the most popular (psychophysical) theories for the FC format is Thurstone's Law of Comparative Judgment (LCJ; Thurstone, 1927, 1931). Somewhat unusual from the perspective of common psychometric models is its typical use to scale stimuli and not individuals according to specific criteria. While the estimation of Thurstone models was initially unfeasible for practical applications (Maydeu-Olivares, 1999; Yao & Böckenholt, 1999), recent technical developments made the estimation of Thurstone models possible. Unfortunately, responses gathered with a FC format are generally ipsative. That is, ipsative scores are not directly comparable interindividually. However, by using limited information estimation, a confirmatory factor analytic (CFA) method for the estimation of stimulus scales was proposed in the seminal work by Maydeu-Olivares and Böckenholt (2005). These new Thurstonian CFA models made the differentiation and estimation of the different classical Thurstone cases easily and computationally efficiently available. Moreover, in some cases, factor scores can be estimated with these models that allow for a comparison between respondents (i.e., they can be used as so-called normative scores).

### Figure 3.1

*Examples for the Forced-Choice Format.*

<p><b>A – Paired Comparison</b> Please select the option that describes you the best.</p> <p><input type="radio"/> I talk a lot when I am at parties. <input type="radio"/> I keep my desk always orderly.</p>	<p><b>B - Triplet</b> Please rank the options according to how well they describe you.</p> <ol style="list-style-type: none"><li>1. Sometimes I just want to cry.</li><li>3. I am always prepared.</li><li>2. I like to go to the museum.</li></ol>
--	---



## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Mainly due to the fact that using an FC format as shown in the left panel in Figure 3.1 leads to assessments that are too burdensome for the respondents, more efficient ways of gathering data for Thurstonian models have been sought. The typical solution is shown in the right panel of Figure 3.1 where three stimuli are presented at once (a block of a so-called triplet) and the respondents' task is to establish a ranking of these stimuli. Hence, instead of paired comparisons, such a ranking design is often used in applied settings.

The ranking design has two main advantages: first, the workload of a respondent is considerably lower in ranking designs as compared to paired comparison designs. Given three stimuli, one triplet is used in rankings while three paired comparisons would be needed to be performed. Second, the ranking design eliminates the possibility of intransitive responding. To illustrate, consider  $n = 3$  stimuli that are labeled as  $\{A, B, C\}$ . For these stimuli  $\tilde{n} = n(n-1)/2 = 3$  nonredundant paired comparisons  $\{i, j\}$  can be constructed:  $\{A, B\}$ ,  $\{A, C\}$ ,  $\{B, C\}$ . Suppose a respondent prefers A over B and B over C. It would then seem logical because of transitivity that the respondent also prefers A over C. However, it would be possible and it actually happens quite regularly in practice that a respondent prefers C over A instead. This would be intransitive responding and considered as an observed error in a paired comparison design. As a consequence of ranking and transitive responses, the factor scores for the respondents cannot be estimated anymore via the Thurstonian CFA models (Jansen & Schulze, 2023a; Maydeu-Olivares & Brown, 2010). Hence, the Thurstonian CFA models are not useful in such cases when interest shifts from item scaling to person scaling. Fortunately, Maydeu-Olivares and Brown (2010; see also Brown & Maydeu-Olivares, 2011) proposed a solution to this problem by reparametrizing the Thurstonian CFA model into an IRT model. This very important step in the given field of research led to the widespread use of the Thurstonian IRT model by constructing multidimensional forced-choice (MFC) questionnaires.

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Alas, some problematic features of the MFC design became apparent in recent years that limit the usefulness of Thurstonian models. For example, Brown and Maydeu-Olivares (2011) reported that parameter estimation and latent trait score recovery are only satisfactory if blocks with both positively and negatively keyed items are constructed. On the other hand, blocks constructed according to this recommendation are more susceptible to faking which runs counter the initial goal why FC designs were introduced to begin with. It seems very implausible that blocks with mixed keyed items can be constructed while simultaneously controlling for the desirability of each statement in a block (see Bürkner et al., 2019; Schulte et al., 2021).

Another problematic feature is that the use of a model incorporating only MFC blocks results in biased estimation of parameters and biased simulation of data (Jansen & Schulze, 2023a). Moreover, identification constraints need to be applied for every block to identify a model which results in incomparable scales between blocks. Especially in applications where the means of the items (i.e., expected values of the latent utilities) are of interest, this results in arbitrary constraints that are not comparable between blocks. This is the case because one latent utility mean needs to be fixed per block for estimation purposes and this results in some means being arbitrarily fixed to be equal, given the same value is used for identification across blocks. Matching by the items' latent utility is possible, however, it would most likely be invalid (Jansen & Schulze, 2023a). Lastly, the fact that model rejection rates are inflated for block designs (Brown & Maydeu-Olivares, 2011) is yet another problematic feature of the MFC design.

The current article provides a modeling framework for FC designs, incorporating any number of traits assessed with paired comparisons or rankings, that at least substantially reduces the impact of the aforementioned problems. There are four sections. The first section provides a necessary description of the response coding from paired comparisons and

rankings into binary outcome variables (see also Maydeu-Olivares & Böckenholt, 2005). The subsequent section describes the Thurstonian factor and IRT models, including the MFC format where only multidimensional blocks are used (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Böckenholt, 2005). In the third section, the Thurstonian linked block design is presented. It is a modification of the Thurstonian block designs as introduced by Brown and Maydeu-Olivares (2011). In section four, the results of comprehensive simulation studies are reported to compare the block with linked block designs and illustrate parameter, latent trait recovery as well as model rejection rates.

### 3.2.3 Binary Coding of Responses of Paired Comparisons and Rankings

The standard procedure to code responses goes back to Thurstone's LCJ (Thurstone, 1927, 1931) and has also been described elsewhere (e.g., Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Jansen & Schulze, 2023a; Maydeu-Olivares & Böckenholt, 2005).

In the context of the LCJ it is assumed that at any given presentation a stimulus evokes a so-called discriminative process. This process results in a scale value  $t_i$  (latent utility) for stimulus  $i$  on a certain psychological continuum of a respondent. Since the repeated presentation of a stimulus does not necessarily elicit the same psychological sensation, latent utilities follow some distribution. The distribution has expected value  $\mu_{t_i}$  and standard deviation  $\sigma_{t_i}$  and is almost without exception assumed to be normal. The same is true for any other stimulus  $j$  that may be presented. When two or more stimuli are presented at the same time and a comparison is required, the latent utilities of the stimuli are discriminated. If two stimuli  $i$  and  $j$  are compared, then the difference between the latent utilities determines whether stimulus  $i$  or  $j$  is chosen. The entire process is not directly observable, but the response that results from this process is observed. The observed response  $y_i$  is coded as

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

$$y_l = \begin{cases} 1 & \text{if } t_i + e_l \geq t_j \\ 0 & \text{if } t_i + e_l < t_j \end{cases} \quad (3.1)$$

The choice between  $i$  and  $j$  is coded as 1 if stimulus  $i$  is chosen over stimulus  $j$  and 0 otherwise. Note that, there is an error term  $e_l$  that represents the error for intransitive responses in a paired comparison design. The transitivity enforced in ranking designs implies that the error term is  $e_l = 0$  for every comparison in these designs.

In a ranking task, all choice alternatives are presented at once, but the coding scheme of Equation (3.1) may nevertheless be used for each possible comparison. For example, consider again  $n = 3$  stimuli labeled as  $\{A, B, C\}$ . If for  $\{A, B\}$  A is chosen over B then  $y_{\{A,B\}} = 1$  and 0 otherwise. This can be done for every paired comparison and ranking task (see Maydeu-Olivares & Böckenholt, 2005):

Ranking			Ordering		
A	B	C	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
1	3	2	A	C	B
Binary outcomes					
	{A, B}	{A, C}	{B, C}		
	1	1	0		

The latent difference between the utilities is denoted as  $y_l^* = t_i - t_j$  and not directly observable. However, we can specify the relationship between unobservable  $y_l^*$  and observed  $y_l$  as

$$y_l = \begin{cases} 1 & \text{if } y_l^* + e_l \geq 0 \\ 0 & \text{if } y_l^* + e_l < 0 \end{cases} \quad (3.2)$$

### 3.2.4 Thurstonian Models

In the following, we consider only Thurstonian models that are applicable when at least one common factor (latent trait) is assumed. So-called simple Thurstonian models (Jansen & Schulze, 2022a; introduced by Maydeu-Olivares & Böckenholt, 2005), where the stimuli are only correlated and Thurstone's cases can be tested, are not considered here. In other words, we focus on the Thurstonian factor and the Thurstonian IRT models.

#### 3.2.4.1 Thurstonian Factor Model

The Thurstonian factor model was introduced in the seminal work by Maydeu-Olivares and Böckenholt (2005). It is useful, for example, whenever the means of the latent utilities or latent traits are of interest. An appropriate application is therefore item scaling where item properties are of interest, for example, on a specific latent trait (Maydeu-Olivares & Böckenholt, 2005) or even social desirability (Jansen & Schulze, 2022b).

In general, writing all latent differences  $y_i^*$  between the latent utilities in vector-matrix form and including the error term,  $\mathbf{y}^*$  yields

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e}. \quad (3.3)$$

Here,  $\mathbf{y}^*$  is a  $\tilde{n} \times 1$  vector,  $\mathbf{t}$  is a  $n \times 1$  vector of the latent utilities, and  $\mathbf{A}$  is a  $\tilde{n} \times n$  design matrix. The rows of  $\mathbf{A}$  correspond to the paired comparisons and the columns correspond to the choice alternatives. Finally,  $\mathbf{e}$  is a  $\tilde{n} \times 1$  vector of uncorrelated random error terms. Thus, the covariance matrix  $\mathbf{\Omega}^2$  of the residuals is diagonal. An example for a design matrix with  $n = 4$  is given by

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \quad (3.4)$$

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Generally, it is assumed that the latent utilities follow a multivariate normal distribution (Maydeu-Olivares & Böckenholt, 2005; Thurstone, 1927), that is

$$\mathbf{t} \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \quad (3.5)$$

where  $\boldsymbol{\mu}_t$  is the mean vector and  $\boldsymbol{\Sigma}_t$  is the variance-covariance matrix of the latent utilities.

With  $m$  as the number of latent traits, the  $n$  latent utilities in  $\mathbf{t}$  can be expressed as

$$\mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (3.6)$$

where  $\boldsymbol{\Lambda}$  is the  $n \times m$  matrix of factor loadings of the latent utilities on the latent traits,  $\boldsymbol{\eta}$  is a  $m \times 1$  vector of the latent traits, and  $\boldsymbol{\varepsilon}$  is a  $n \times 1$  vector of unique factors (residual term). The means and variances of the common factors are zero and one, respectively. The common factors may be correlated as specified in a  $m \times m$  correlation matrix  $\boldsymbol{\Phi}$ . The unique factors (uniquenesses) are also assumed to have a mean of zero and to be uncorrelated. Thus, the variance-covariance matrix  $\boldsymbol{\Psi}^2$  of residual terms is diagonal. With Equation (3.3) and (3.6) the latent differences are

$$\mathbf{y}^* = \mathbf{A}(\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} \quad (3.7)$$

which results in the mean and covariance structure

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t, \text{ and } \boldsymbol{\Sigma}_{y^*} = \mathbf{A}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{A}' + \boldsymbol{\Omega}^2. \quad (3.8)$$

To estimate Thurstonian factor models, the thresholds and tetrachoric correlations of the normal variables must be estimated since all observed data is categorical (Muthén, 1978). To obtain thresholds, the latent differences have to be standardized as follows:  $\mathbf{z}^* = \mathbf{D}(\mathbf{y}^* - \boldsymbol{\mu}_{y^*})$

with  $\mathbf{D} = \left[ \text{diag}(\boldsymbol{\Sigma}_{y^*}) \right]^{-1/2}$  (Maydeu-Olivares & Böckenholt, 2005). The standardized latent differences follow a multivariate normal distribution with  $\boldsymbol{\mu}_{z^*} = \mathbf{0}$  and the tetrachoric correlation matrix given by

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

$$\mathbf{P}_{z^*} = \mathbf{D}(\boldsymbol{\Sigma}_{y^*})\mathbf{D} = \mathbf{D}(\mathbf{A}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{A}' + \boldsymbol{\Omega}^2)\mathbf{D} \quad (3.9)$$

The relationship between the standardized latent differences and the observed responses is one of the  $\tilde{n}$  thresholds  $\tau_l$

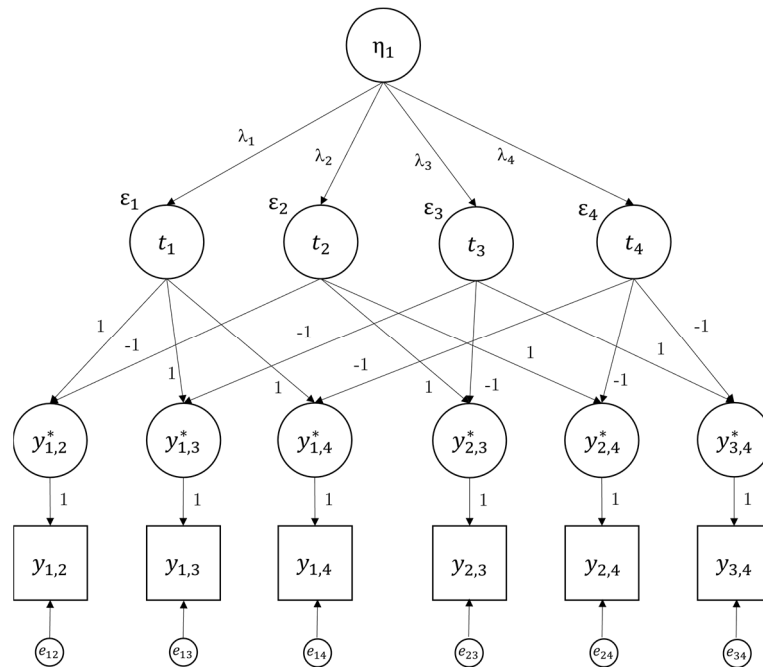
$$y_l = \begin{cases} 1 & \text{if } z_l^* \geq \tau_l \\ 0 & \text{if } z_l^* < \tau_l \end{cases} \quad (3.10)$$

where the vector of thresholds is  $\boldsymbol{\tau} = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}_t$  (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Böckenholt, 2005; Muthén, 1978). Figure 3.2 gives an example for a Thurstonian factor model with  $n = 4$  and  $m = 1$ .

All fundamental model equations also hold in a ranking design, but with  $\mathbf{e} = 0$  and  $\boldsymbol{\Omega}^2 = 0$ , as intransitive responses are not possible.

**Figure 3.2**

*Example of a Covariance Structure of a Thurstonian Factor Model for  $n = 4$  and  $m = 1$ .*



### 3.2.4.2 Identification of Thurstonian Factor Models

For item scaling, the means of the latent utilities are of interest. Originally, the matrix representation of the system of linear equations that would have to be solved, would need to

be of full rank for the estimation of a unique vector of utilities. However, since the design matrix  $\mathbf{A}$  has rank  $n - 1$  the design has a location indeterminacy.

To estimate the latent utility means all thresholds must be fixed. Additionally, for identification purposes and to solve the location indeterminacy, one latent utility mean has to be fixed (e.g.,  $\mu_n = 0$ ). All other latent utility means are estimated relative to the constrained value. In addition, one factor loading must also be fixed (e.g.,  $\lambda_1 = 1$ ) if only one latent trait is included in the model. If more than one latent trait is incorporated, constraints on loadings are not necessary (Maydeu-Olivares & Brown, 2010). Furthermore, one variance of one latent utility and, lastly, the variances of all latent traits need to be fixed (e.g., to unity). These constraints set the scales of the factor loadings as well as the unique variances and, in a sense, substitute the one missing rank of  $\mathbf{A}$ .

### 3.2.4.3 Thurstonian IRT Models

For the derivation of the Thurstonian IRT models (Maydeu-Olivares & Brown, 2010), a Thurstonian factor model with unconstrained thresholds must be reparameterized. To do so, all latent utility means are fixed to zero. This establishes identification and enables estimation of the intercepts. The  $\tilde{n}$  intercepts are defined as  $-\gamma$  and were constrained for identification of the latent utility means as

$$\gamma = -\mathbf{A}\boldsymbol{\mu}_t. \quad (3.11)$$

From Equation (3.7) it follows that the model with unconstrained intercepts is defined by

$$\mathbf{y}^* = -\gamma + \mathbf{A}\mathbf{t} + \mathbf{e}, \quad \mathbf{t} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (3.12)$$

and the threshold structure is  $\boldsymbol{\tau} = \mathbf{D}\boldsymbol{\gamma}$ , which is also unconstrained as it is a rescaling of  $\boldsymbol{\gamma}$  by  $\mathbf{D}$ . Given unconstrained thresholds and restricted means, factor score estimation and the consideration of individual differences would be possible in a paired comparison design.



Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

However, if a ranking design is given, the estimation of factor scores is not possible due to the non-positive residual variances of the categorical indicators with  $\mathbf{e} = 0$  and  $\mathbf{\Omega}^2 = 0$ .

By a reparameterization of the higher-order model in Equation (3.12) into a first-order model, the Thurstonian IRT model is defined (Maydeu-Olivares & Brown, 2010). The reparameterization is done by

$$\mathbf{y}^* = -\gamma + \mathbf{A}(\mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = -\gamma + \mathbf{A}\mathbf{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\gamma + \tilde{\mathbf{\Lambda}}\boldsymbol{\eta} + \tilde{\boldsymbol{\varepsilon}} \quad (3.13)$$

with  $\tilde{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}$  and  $\text{cov}(\tilde{\boldsymbol{\varepsilon}}) = \tilde{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \mathbf{\Omega}^2$ , where  $\tilde{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda}$  is a  $\tilde{n} \times m$  matrix. To illustrate, with  $m = 1$  and  $n = 3$  it would be

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_3 \\ \lambda_2 - \lambda_3 \end{pmatrix} \quad (3.14)$$

and setting  $m = 3$  and  $n = 3$  the result is

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \end{pmatrix}. \quad (3.15)$$

In both cases the covariance matrix of unique errors is

$$\tilde{\boldsymbol{\Psi}}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2 + \psi_3^2 + \omega_3^2 \end{pmatrix}. \quad (3.16)$$

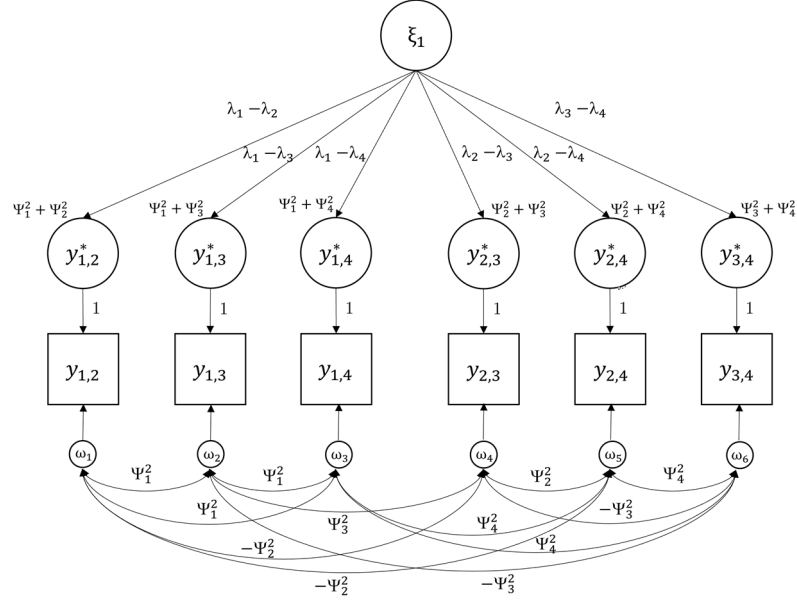
Both models in (3.12) and (3.13) are equivalent because (3.13) is simply a reparameterization of (3.12) and, therefore, both models have the same (but reparameterized) tetrachoric correlation matrix

$$\mathbf{P}_z^* = \mathbf{D}(\mathbf{A}\mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}'\mathbf{A}' + \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \mathbf{\Omega}^2)\mathbf{D} = \mathbf{D}(\tilde{\mathbf{\Lambda}}\boldsymbol{\Phi}\tilde{\mathbf{\Lambda}}' + \tilde{\boldsymbol{\Psi}}^2)\mathbf{D} \quad (3.17)$$

Figure 3.3 gives an example for a Thurstonian IRT model with  $n = 4$  and  $m = 1$ .

**Figure 3.3**

Example of a Covariance Structure of a Thurstonian IRT Model for  $n = 4$  and  $m = 1$ .



The reparameterized model can equivalently be viewed as a normal ogive model with some special features. Let  $\Phi(x)$  be a standard normal distribution function at  $x$ ,  $\gamma_i$  the threshold for  $y_i$ ,  $\tilde{\lambda}_i$  the vector of factor loadings, and  $\tilde{\psi}_i^2$  the variance of the binary response. Then, for each binary response variable for items  $i$  and  $j$ , the item characteristic function (ICF) is given by

$$\Pr(y_i = 1 | \boldsymbol{\eta}) = \Phi \left( \frac{-\gamma_i + \tilde{\lambda}_i \boldsymbol{\eta}}{\sqrt{\tilde{\psi}_i^2}} \right). \quad (3.18)$$

Equation (3.18) provides the ICF of a normal ogive model but with structured  $\tilde{\lambda}_i$  and  $\tilde{\psi}_i^2$ .

Also, the ICFs are not independent (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). When  $m = 1$  the ICF is

$$\Pr(y_i = 1 | \eta) = \Phi \left( \frac{-\gamma_i + \tilde{\lambda}_i \eta}{\sqrt{\tilde{\psi}_i^2}} \right) = \Phi \left( \frac{-\gamma_i + (\lambda_i - \lambda_j) \eta}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_i^2}} \right) \quad (3.19)$$

and if  $m > 1$ , for each comparison it is

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi \left( \frac{-\gamma_l + \lambda_i \eta_a - \lambda_j \eta_b}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}} \right). \quad (3.20)$$

Expressed in intercept and slope notation, it follows from

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}, \quad \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}}, \quad \beta_j = \frac{\lambda_j}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_l^2}} \quad (3.21)$$

that when  $m = 1$ , the ICF is

$$\Pr(y_l = 1 | \eta) = \Phi(\alpha_l + (\beta_i - \beta_k)\eta). \quad (3.22)$$

Given  $m > 1$ , then for each comparison it becomes

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi(\alpha_l + \beta_i \eta_a - \beta_k \eta_b). \quad (3.23)$$

To identify Thurstonian IRT models, the same identification constraints need to be applied as is the case for Thurstonian factor models, except for the constraints on the means. That is, one of the error variances (uniquenesses) of the latent utilities needs to be fixed and the variances of the latent traits need to be fixed as well (e.g., both to one). Also, one factor loading must be fixed (e.g.,  $\lambda_1 = 1$ ) if only one latent trait is included. If more than one latent trait is considered, the loading constraint is not necessary (Maydeu-Olivares & Brown, 2010). Again, these constraints set the scales of the factor loadings, the unique variances, and substitute the one missing rank of  $\mathbf{A}$  as was the case for the Thurstonian factor models.

#### **3.2.4.4 Latent Trait Estimation**

The main advantage of the Thurstonian IRT model is that the error variances of the binary indicators are structured (Equation (3.16)). This means that they are non-zero and positive for both a paired comparison and also a ranking design. Customarily, the maximum a posteriori (MAP) estimation is used for the estimation of factor scores (Brown & Maydeu-Olivares, 2011). When IRT scores are obtained with the MAP method, the posterior test

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

information for each respondent at each MAP point estimate is evaluated. The empirical reliability for the MAP scores can be calculated by

$$\rho = \frac{\sigma^2 - \bar{\sigma}_{error}^2}{\sigma^2} \quad (3.24)$$

where  $\sigma^2$  is estimated using the variance of the MAP scores, and  $\bar{\sigma}_{error}^2$  is estimated by

$$\bar{\sigma}_{error}^2(\hat{\eta}) = \frac{1}{N} \sum_{j=1}^N SE_{estimate}^2, \quad (3.25)$$

where  $SE_{estimate}^2$  are the squared standard errors of each estimate (see also Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010).

#### 3.2.4.5 Estimation of Thurstonian Models

Thurstonian models can be estimated with limited information estimation methods (Maydeu-Olivares & Böckenholt, 2005). The necessary steps involve the estimation of the tetrachoric correlations and thresholds. Based on these, model parameters are estimated using diagonally weighted least squares (DWLS) or unweighted least squares (ULS) estimators (Maydeu-Olivares & Böckenholt, 2005; Maydeu-Olivares & Brown, 2010). It was shown, that the differences in the results between estimation methods are negligible (Forero et al., 2009).

In general, the number of degrees of freedom can be calculated by using the number of thresholds and tetrachoric correlations of the dichotomous responses. Let  $\tilde{n}$  be the number of dichotomous response variables, then  $\tilde{n}$  thresholds and  $\tilde{n}(\tilde{n}-1)/2$  tetrachoric correlations are given, resulting in  $\tilde{n}(\tilde{n}+1)/2$  known parameters. In paired comparison designs, the number of degrees of freedom is  $df = \tilde{n}(\tilde{n}+1)/2 - q$ , where  $q$  is the number of parameters to estimate. As rankings include only a subset of all possible paired comparisons, the degrees of freedom have to be adjusted since there are

$$r = \frac{n(n-1)(n-2)}{6} \quad (3.26)$$

redundancies among the tetrachoric correlations and thresholds (Maydeu-Olivares, 1999; Maydeu-Olivares & Böckenholt, 2005). As a consequence,  $r$  must be subtracted from the degrees of freedom as reported by common structural equation modeling software. The correct number is  $df_{corr} = \tilde{n}(\tilde{n}+1)/2 - q - r$  and fit indices must also be adjusted accordingly.

#### 3.2.4.6 Thurstonian Block Designs

A major practical problem with the use of Thurstonian models is that respondents need to respond to a great number of comparisons, even if the number of stimuli is small. For example, considering only  $n = 20$  stimuli would already result in 190 paired comparisons. These comparisons would need to be presented one by one, or coupled in a ranking design. While a ranking of many stimuli at once is considerably faster than the sequential presentation of many paired comparisons, ranking a lot of stimuli can be a considerably complex task, depending on the type of stimuli. There is evidence (Sass et al., 2020) that from the perspective of a respondent, the presentation of rankings with more than four stimuli already seems to inflict a substantial workload, while test motivation appears to be unaffected.

The MFC format was proposed by Brown and Maydeu-Olivares (2011) as a solution to this problem. In this format, stimuli or items are presented in blocks of  $k$  and all blocks include items assigned to at least  $k$  different traits. The number of items  $n$  must be divisible by  $k$  so that there is a total of  $p = n/k$  blocks presented to a respondent in this format. The corresponding design will be called the Thurstonian block design (TB design).

The TB design can be applied to all Thurstonian models (Jansen & Schulze, 2023a). To date, it has been used and discussed only within the IRT context. From a technical perspective, the TB design equations are identical to ones for the models specified in the



Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

As can be seen, the equations that are true for a full design, naturally also hold for multiple blocks of smaller size. Note, that for the TB design, the rank of each submatrix (block) is always  $k - 1$ . For a design with  $n = pk$  items,  $\mathbf{A}$  is of rank  $(k - 1)p$ . To identify the TB design, the same identification constraints have to be used as described before. However, the identification constraints need to be applied for every block to substitute the missing rank within each block. For rankings, the number of degrees of freedom must again be adjusted, as there are

$$r = \frac{k(k-1)(k-2)}{6} \quad (3.30)$$

redundancies among the tetrachoric correlations and thresholds for each block (Brown & Maydeu-Olivares, 2011). Hence, as before,  $r$  must be subtracted from the degrees of freedom as reported in the output of common structural equation modeling software. The correct number is  $df_{adj} = \tilde{n}(\tilde{n} + 1) / 2 - q - r$  and fit indices must also be adjusted accordingly. Special care has to be taken for designs of block size two (Brown & Maydeu-Olivares, 2011, Appendix A). In this specific case, uniquenesses are not identifiable so they are all fixed for identification of the model by fixing the variance of the latent difference responses to 1.

### 3.2.5 The Thurstonian Linked Block Design

The independent estimation of the parameters per block can be regarded as the most central critical issue of the TB design. The reason is that parameter estimates remain mostly ipsative as a consequence of the independent estimation. To illustrate, assume the design as given in Equations (3.27) to (3.29). The blocks are  $B_1 = (1,2,3)$ ,  $B_2 = (4,5,6)$ , and  $B_3 = (7,8,9)$ , respectively. The corresponding matrix of loadings is

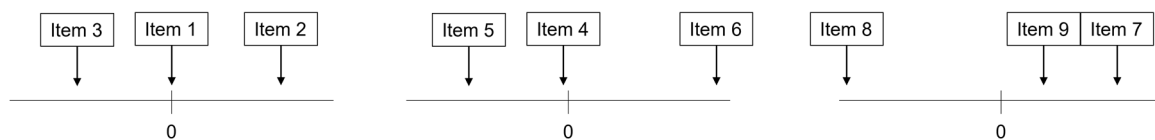
Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ \lambda_4 & 0 & 0 \\ 0 & \lambda_5 & 0 \\ 0 & 0 & \lambda_6 \\ \lambda_7 & 0 & 0 \\ 0 & \lambda_8 & 0 \\ 0 & 0 & \lambda_9 \end{pmatrix}. \quad (3.31)$$

Now, a respondent provides a ranking for each block but comparisons are made only for all the items *within* a block. None of the items is compared to any of the others *between* blocks. Therefore, the scale must be redefined for each block, to identify the model. The consequence is that items in different blocks are not on the same scale and, therefore, not comparable. For an illustration see Figure 3.4. In this case, items 1 and 4 have similar values on the utility scale, with item 4 having a slightly lower value than item 1. Since no direct or indirect comparison was done between the two items, it is unclear if these two items are truly very similar in utility. Comparisons between item 4 and the items in the first block could yield either of the following results: item 4 is (a) less attractive than item three, (b) more attractive than item 3, but less than item 1, (c) more attractive than item 1, but less than item 2, or (d) more attractive than item 2. Unless such comparisons are done, any comparison of items between blocks is not meaningful because the items of each block are located on separate scales.

**Figure 3.4**

*Example of Responses for Each Block Compared to the Utility Scale.*





## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

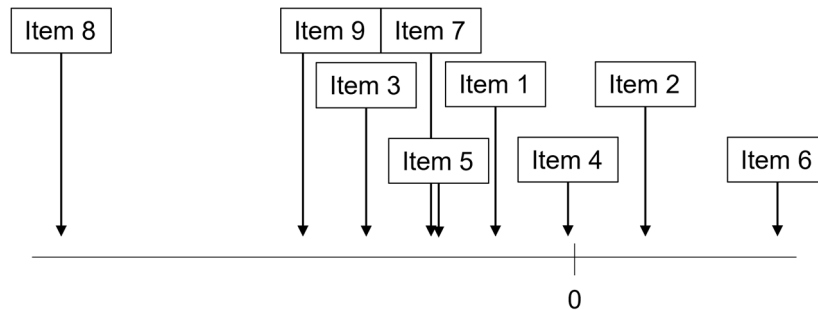
The Thurstonian linked block (TLB) design is designed to overcome or at least reduce this important problem. The main goal associated with the TLB is to estimate all model parameters in a design including multiple blocks with the same set of identification constraints as would be needed if only one block (i.e., a full design) was present. The implications of this type of design will be discussed after a more elaborate presentation of the TLB approach.

Naturally, the fundamental model equations of the Thurstonian factor and IRT models also hold for the TLB design. The focal difference between the designs lies in the design matrix  $\mathbf{A}$ . To estimate every parameter in relation to the parameter fixed for identification, all items need to be linked to one another. In technical terms, this is related to the fact that matrix  $\mathbf{A}$  is of rank  $n - 1$ . For each rank of matrix  $\mathbf{A}$  smaller than  $n$ , one set of identifications constraints must be applied. Creating a design of rank  $n - 1$  can be achieved by two strategies. One way is to assume a full design with all possible paired comparisons available. From this design, paired comparisons are expunged under the condition that each item is still connected to every other item through retained item comparisons, so that the rank is  $n - 1$ . A second more easily specified way is to construct an initial TB design as proposed by Brown and Maydeu-Olivares (2011) and add blocks that include the item comparisons needed to have all items linked to one another through a chain of comparisons.

Going back to the example of Equations (3.27) to (3.29) and Figure 3.4, shows that a simple way to achieve a TLB design would be to add one triplet with, for example, items 1, 4, and 7 (see also Figure 3.5). If the order of attractiveness would be item 7, item 1, and item 4, then naturally, we have no full ranking and there are still “missing” binary indicators. However, the parameters can now be estimated in relation to the same identification constraint and therefore on the same scale. In our example,  $\mathbf{A}$  is of rank  $3(3 - 1) = 6$  for the

**Figure 3.5**

*Example of Responses for Each Block Compared to the Utility Scale in a Thurstonian Linked Block Design. In the Linking Block, the Order of Attractiveness is Item 7, Item 1, then Item 4.*



TB design. By adding a block with items of three yet unconnected blocks,  $\mathbf{A}$  is now of rank  $9 - 1 = 8$ .

It is noteworthy that in the case of rankings some paired comparisons can be deduced via the implied transitivity relation. For example, we know for the specific respondent in Figure 3.5 that item 2 is more attractive than item 1 and both items are included in the first block.

Similarly, it is known that items 8 and 9 are less attractive than item 7. As a result of including the linking block, it is also established that item 1 is more attractive than item 7.

Transitivity now implies that items 7, 8, and 9 are less attractive than item 2 and item 1.

To illustrate the structure imposed by a TLB design on the matrices, the example is translated into the model equations. Again, matrix  $\mathbf{A}$  of Equation (3.27) is of rank 6. By adding the block  $B_4 = (1,4,7)$ ,  $\mathbf{A}$  is changed to

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired  
Comparison and Ranking Data

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}, \quad (3.32)$$

that is, the linking block is added. Note that, the number of columns corresponds to the number of items and has not changed. This matrix is of rank  $n - 1 = 8$ . Hence, a TLB design is given. An equivalent design is given by

$$\mathbf{A}^* = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad (3.33)$$

as only the order of the binary indicators is changed. Given Equation (3.32), the matrix of loadings of the reparameterized model is

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

$$\tilde{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \\ \lambda_1 - \lambda_4 & 0 & 0 \\ \lambda_1 - \lambda_7 & 0 & 0 \\ \lambda_4 - \lambda_7 & 0 & 0 \end{pmatrix} \quad (3.34)$$

as all items of  $B_4 = (1,4,7)$  measure trait one. Of course, we also could define  $B_5 = (1,5,9)$  as a multidimensional linking block. In this case  $\tilde{\mathbf{\Lambda}}$  would be

$$\tilde{\mathbf{\Lambda}} = \mathbf{A}\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \\ \lambda_1 & -\lambda_5 & 0 \\ \lambda_1 & 0 & -\lambda_9 \\ 0 & \lambda_5 & -\lambda_9 \end{pmatrix} \quad (3.35)$$

and matrix  $\tilde{\Psi}^2$  would be patterned accordingly. The TLB design is identified by the same constraints that are used for the full (one block) design. The constraints per block are not necessary anymore. Again, special care has to be taken for designs of block size two. The uniquenesses are again fixed for identification by fixing the variance of the latent difference responses to 1 (if no three comparisons are about the same items). In contrast to unlinked

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

block designs, some loadings and uniquenesses appear more than once in the matrices. For the loadings, corresponding equality constraints must be applied.

In light of the critique about the fakeability of a block and its implications (Bürkner et al., 2019; Jansen & Schulze, 2023a; Schulte et al., 2021) it is advisable to only include items with the same key direction within each initial block. By linking blocks, items with negative and positive loadings, or at least larger differences in loadings have to be compared nonetheless. Such blocks should then contribute to the maximal absolute difference, and therefore, enhance parameter estimation and latent trait recovery (as discussed by Brown & Maydeu-Olivares, 2011).

It is emphasized that no other restrictions are needed to use the design. In theory, one block can be unidimensional, multidimensional, or include both uni- and multidimensional paired comparisons. Also, blocks can be of different sizes. For example, a design can include both triplets and quads. The number of items is not restricted in any way and it does not have to be constrained by  $n = pk$  in particular. However, in comparison to the TB design, more blocks are necessary to achieve a design matrix of rank  $n - 1$ . If only same-sized blocks of size  $k$  are used,

$$\left\lceil \frac{p-1}{k-1} \right\rceil \quad (3.36)$$

additional blocks are needed, where  $\lceil \cdot \rceil$  is the ceiling function and corresponds to rounding up the result. If the number of blocks in a design is critical and there is a practical need to minimize it, for example, it is noteworthy that not all blocks have to be linked. However, this would come at the cost of having to include additional identification constraints, as the rank of  $\mathbf{A}$  is not  $n - 1$ .

Lastly, it should be kept in mind that in a ranking design there are redundancies among the thresholds and the tetrachoric correlations. Previous results (Brown & Maydeu-

Olivares, 2011; Maydeu-Olivares, 1999) can be generalized to a TLB design but only if no comparison is presented twice. Then, for each block the number of redundancies can be calculated by Equation (3.30) (also see Appendix A). For cases where two comparisons are presented at least twice, but transitivity is still assumed, there is no closed form to calculate the redundancies. This is discussed in more detail in Appendix A.

### 3.2.6 Simulation Studies

In this section, simulation study results on the Thurstonian models are presented. The aims of the simulation studies presented in this section are to empirically demonstrate issues and shortcomings of the TB design that have been detailed elsewhere (see Jansen & Schulze, 2023a) and – most importantly -- compare the TB design to the TLB design to show the latter's superiority. At first glance it may appear trivial that the TLB design leads to better parameter estimations because of the higher number of blocks alone. However, the simulation results will show that an increase in the number of blocks is not sufficient to enhance quality.

A major first question to design the simulation studies is: From which model should the data be generated? In the case of Thurstonian modeling, each specific design (i.e., matrix **A**) yields different results in a simulation (Jansen & Schulze, 2023a). As a specific design should not determine the test model, the one and only model design that data should be generated from is a full design with all possible paired comparisons. From the resulting data, only the paired comparisons are extracted that fit the specific design. Furthermore, to reduce complexity, data is generated from a Thurstonian factor model with restricted thresholds (Equation (3.7)), assuming a ranking task. As a consequence,  $\mathbf{e} = 0$ . All models are applied to the same data sets. For all data sets, 2000 respondents were simulated with 1000 repetitions per condition.

The Thurstonian factor as well as the Thurstonian IRT model are used for the same data sets in order to study differences in parameter and trait recovery for both the item-

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

centered and the person-centered approaches. Only triplets are used as blocks. For each condition, there are three block model design types: unlinked block designs (U designs; estimated with the TB design;  $\mathbf{A}$  is of rank  $(3 - 1)p = 2p$ ), partially linked block designs (P designs; estimated with the TLB design;  $\mathbf{A}$  is of rank larger than  $2p$ , but lower than  $n - 1$ ), and a complete linked block design (C designs; estimated by the TLB design;  $\mathbf{A}$  is of rank  $n - 1$ ) are used. For a P design,  $\left\lceil \frac{p-1}{3-1} \right\rceil$  additional blocks were chosen, so that the connectedness of blocks was as small as possible (i.e., the rank of  $\mathbf{A}$  was as low as possible). Thus, P and C designs had the same number of blocks but differed with respect to the connectedness of blocks. The initial blocks were the same for all three block design types. Additional blocks were chosen at random, consistent with the block design definition, of course.

With respect to keying of the items within blocks, Brown and Maydeu-Olivares (2011) recommend that each block should contain items that are keyed in different directions of their respective traits. As pointed out before and elsewhere, this recommendation is problematic for different reasons (Bürkner et al., 2019; Jansen & Schulze, 2023a) which is why only consistently positively keyed items are used in the blocks in the simulation studies. Different block compositions will be compared, though, based on a different number of traits and based on utilities as well as loadings.

### ***3.2.6.1 Simulation Study 1: A Forced-Choice Questionnaire Measuring One Trait***

In the first simulation study only one latent trait is given in all designs. The number of items was varied (12 or 18). Block compositions were as follows:

1. Item assignment to blocks according to their utilities. Items with similar utilities (e.g., items with the lowest three utilities) are combined into blocks.
2. Item assignment to blocks according to their loadings. As in 1, the block assignment groups similar items but with respect to their loadings.

3. Random assignment of items to blocks.

It can be expected that block composition (2) is the least reliable, as the influence of the trait is small (see Equation (3.19)). In addition to the three block compositions, a full one-block design with all possible paired comparisons is also considered. The true item parameters were drawn from a uniform distribution between .30 and .90 for the factor loadings, and between -1 and 1 for the utility means. True uniquenesses were specified to be  $\psi_i^2 = 1 - \lambda_i^2$ , similar to the procedure by Brown and Maydeu-Olivares (2011). If 18 items were drawn, these include the same twelve items and six additional items.

To estimate the IRT models the procedure was as follows for the different block compositions: a) U designs: one factor loading and one uniqueness were fixed to unity for each block. b) P designs: one factor loading and one uniqueness were fixed for each block that has no relation with other blocks. For example, for initial U design blocks  $B_1 = (1,2,3)$ ,  $B_2 = (4,5,6)$ ,  $B_3 = (7,8,9)$ , and  $B_4 = (10,11,12)$ , the two additional blocks could be  $B_5 = (1,2,4)$  and  $B_6 = (1,2,5)$ . As blocks  $B_1$  and  $B_2$  are connected, setting one identification constraint in  $B_1$  identifies the part of the design including  $B_1$ ,  $B_2$ ,  $B_5$ , and  $B_6$ . Additional identification constraints must only be applied for  $B_3$  and  $B_4$ . c) C design and the full one-block design: one factor loading and one uniqueness were fixed. To identify the factor models, one utility mean was additionally fixed to zero, either per (unrelated) block, or one for the whole design.

**3.2.6.1.1 Item Parameter Recovery**

For all conditions investigated, Table 3.1 and Table 3.2 provide the number of converged repetitions (IRT/factor), the average relative bias across estimated loadings, utility means, and thresholds as well as the average bias of their standard errors. The loading estimates of the Thurstonian factor and IRT model are not directly comparable. Hence, only results of the IRT models are reported because convergence rates were lower for the



## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Thurstonian factor model. Naturally, the bias for loadings in the factor model is smaller, as data was simulated from factor models. For each of the separate unrelated blocks within U and P designs, the estimations (loadings/ latent utility means) were regressed on the true parameters and then rescaled so that the scale and the origin were equal between estimations and true parameters. For the full one-block design and the C design, the same was done in relation to the constrained item parameters. Note that this procedure conceals the fact that the scale is redefined for every block and, therefore, is not comparable whatsoever. This means that the determined bias is smaller than the real bias. On the other hand, if all estimations were rescaled only to one of the items that the constraints were put on, the bias would potentially be bloated and artificially much larger than the real bias. For the purposes of the current simulation, results from both methods are presented for the loadings. Later we will rescale the estimates for each block, but report correlations between estimates and the true parameter values. A rescaling is not needed for the thresholds.

Overall, the results from the full design can be seen as a benchmark for the different block and linkage designs. As can be seen in both Table 3.1 and 3.2 the number of converged repetitions depends on the model type (Thurstonian factor or IRT model), the number of items (12 vs. 18), and linking status (unlinked, partially linked, completely linked). Generally, convergence rates are considerably higher for IRT models as compared to factor models (cf. Table 3.1 and 3.2) This is particularly true for non-C designs. For Thurstonian factor models (see Table 3.2), U and P designs do not provide satisfactory convergence rates, whereas with C designs convergence rates are substantially higher.

In general, thresholds and standard errors of thresholds are estimated accurately (relative bias of about 1%) for all designs. If loadings are rescaled per block, the relative bias is small for all designs, with only a few exceptions. However, if the rescaling is done with respect to only one item (identification constraints), the U and P designs perform worse than

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

the C design, in many cases at an unacceptable level (using a cutoff of 10% for relative bias).

This is also reflected by the correlation between estimates and true parameters that is generally higher for the C design. With respect to test length, the results seem to become more stable the more items are used. As predicted, an exceptional case is block composition (2). Per construction, the underlying trait has only a small influence on the results (see Equation (3.19)), so the convergence rates, the relative bias, and the correlation between true and estimated parameter values, are poor.

### ***3.2.6.1.2 Latent Trait Score Recovery***

For all conditions, including the full design, latent trait score recovery was unsatisfactory. For the full design, the actual recovery (i.e., correlation between true scores and MAP scores) was .54 for  $n = 12$  and .63 for  $n = 18$ , respectively. For the block designs, recovery was even smaller. Given  $n = 12$ , mean recoveries were .28, .30, and .37 for U, P, and C designs, respectively. With  $n = 18$ , mean recoveries were .33, .35, and .40. The empirical recovery (square of Equation(3.24)) for the full design was .81 for  $n = 12$  and .93 for  $n = 18$ , respectively. It could not be computed for most unlinked block designs. The variance of the MAP estimates was smaller than the mean of the standard errors. For  $n = 12$  mean empirical recoveries were .40 and .31 for P and C designs, and for  $n = 18$  mean empirical recoveries were .35 and .34 for P and C designs.

### ***3.2.6.1.3 Discussion***

Since data sets are generated from a Thurstonian factor model with restricted thresholds (Equation (3.7)), a comparison of rejection rates should be done for the latent utility factor model. However, as convergence rates are deemed too low, model rejection rates are considered to be meaningless in the given case. Rejection rates will be reported in simulation study 2.

**Table 3.1**
*Thurstonian IRT Models: Results from Simulation Study 1.*

Blocks	<i>n</i>	Convergence			Loadings ARB									Loadings recovery			
					Estimates (per block)			Estimates (once)			<i>SE</i>			<i>U</i>	<i>P</i>	<i>C</i>	
		<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>				
Full	12	1000			.02			.02									.97
Block 1	12	962	894	995	.02	-.28	.05	.15	.02	.05	-.11	-.15	-.12	.90	.41	.91	
Block 2	12	546	584	704	-.07	-.09	-.17	-.82	-.41	-.19	2.32	.55	-.19	.26	.23	.68	
Block 3	12	727	830	949	-.01	.21	.07	.43	.27	.05	.30	-.29	-.25	.67	.77	.84	
Full	18	1000			.01			.01									.98
Block 1	18	999	971	1000	.01	.02	.03	.13	.13	.03	-.37	-.35	-.04	.91	.09	.95	
Block 2	18	528	556	718	-.04	-.01	-.15	.54	9.56	-.14	.69	-.02	-.22	.27	.31	.57	
Block 3	18	952	915	997	-.02	.14	.04	-.42	-.05	.02	-.37	-.44	-.20	.52	.08	.91	

*Note.* *n* = number of items, *U* = unlinked block designs, *P* = partially linked block designs, *C* = completely linked block designs, ARB = average relative bias.

**Table 3.2**

*Thurstonian Factor Models: Results from Simulation Study 1.*

Blocks	<i>n</i>	Convergence			Utility Means ARB						Utility Means recovery		
					Estimates (per block)			<i>SE</i>			Estimates		
		<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>
Full	12	1000			-.07			-.02			1.00		
Block 1	12	0	8	576	-	-.03	-.07	-	.17	-.03	-	.45	1.00
Block 2	12	0	310	484	-	-.0	.04	-	-.02	-.02	-	.71	.98
Block 3	12	1	107	954	-.01	-.01	-.01	-	-.11	-.04	.92	.96	.99
Full	18	1000			-.06			.01			1.00		
Block 1	18	723	829	463	-.03	-.03	-.07	.11	.09	-.02	-.18	-.32	1.00
Block 2	18	19	385	499	-.03	-.01	-.04	.18	.18	-.13	.46	.58	.96
Block 3	18	29	731	989	.00	.00	-.08	-.03	-.03	-.01	.72	.70	.99

*Note.* *n* = number of items, *U* = unlinked block designs, *P* = partially linked block designs, *C* = completely linked block designs, ARB = average relative bias.

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

The first simulation study shows that, overall, one-dimensional Thurstonian FC-designs with 12 items yield unreliable results. U designs have higher relative biases (partly at unacceptable levels) and lower correlations between true and estimated scores. The result for the P designs make it clear that just adding blocks to a U design does not necessarily improve the estimation accuracy. Instead, adding blocks according to C designs yields the best outcomes, though constellations with unsatisfactory results remain. It is worth noting that in comparison to the analyses of Brown and Maydeu-Olivares (2011) blocks in the current designs were restricted to include only positively keyed items.

In sum, there are some recommendations that can already be given based on the first simulation study. First, if Thurstonian models are used for one trait, at least 18 items should be used. Item assignment to blocks according to their loadings as is done in block design 2 should definitely be avoided. A C-design yields the most reliable results overall and is therefore recommend in the situations under investigation in simulation study 1.

### ***3.2.6.2 Simulation Study 2: A Forced-Choice Questionnaire Measuring Five Uncorrelated Traits***

Five uncorrelated traits were used in the second simulation study. Both number of items per trait (6 vs. 12) and the block compositions were varied. Block compositions were as follows:

1. Item assignment to blocks according to their utilities. Items with similar utilities (e.g., items with the lowest three utilities) are combined into blocks.
2. Item assignment to blocks according to their loadings. As in 1, the block assignment groups similar items but with respect to their loadings.
3. Random assignment of items to blocks (including uni- and multidimensional blocks).

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

4. Item assignment to initial blocks according to their factor assignment resulting in purely unidimensional blocks.
5. Item assignment to blocks according to their factor assignment resulting in purely multidimensional blocks.

With respect to block composition (4) it should be noted that even if all initial blocks are unidimensional, linking blocks must include multidimensional blocks. Again, block composition (2) should be the least reliable as the influence of the traits is small (see Equation (3.20)). However, the problem should be less pronounced as compared to simulation study 1 since multiple traits are involved now. The same strategy as in simulation study 1 was used with respect to the sampling of the true parameters and identification constraints. Relative biases for parameters are reported as rescaled per block and the correlation between estimated and true parameters. All results are presented in Tables 3.3 and 3.4.

The convergence rates shown in Table 3.3 are very problematic at least for the block designs with similar factor loadings (Block 2) and unidimensional blocks (Block 4) unless a C design is used. Overall, C designs have very high convergence rates for both factor and IRT models. Only the block designs with similar factor loadings (Block 2) seem to provoke nonconvergence to a substantial degree, but in comparison to U and P designs, they are less serious by far. Again, this illustrates that just the addition of more blocks does not improve convergence rates.

### ***3.2.6.2.1 Item Parameter Recovery***

Considering the loadings, estimation of parameters is biased immensely whenever U or P designs are used, especially, but not limited to, the block designs with similar loadings and unidimensional blocks (Table 3.3). Although this does not seem to be serious for the estimates rescaled per block, the correlations between estimates and true parameters imply

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

heavily biased results. The C design recovers the loadings satisfactorily if a cut-off of 10% is used (mean bias is 0%). Again, this is also evidenced by the correlations between true and estimated parameters, which is about .90 for C designs in all cases. The standard errors of the loadings are accurately estimated except for the block designs with similar loadings and unidimensional blocks (Table 3.3). Only C designs yield accurate estimations for the standard errors of the loadings, with one exception for block design with similar loadings.

As shown in Table 3.4, all designs yield sufficiently precise estimates of the latent utilities if they are rescaled within each block, except for the first block design (blocks of items with similar utilities). For block composition 1 only C designs yield precise estimates. As before, correlations between true and estimated utilities indicate that only C designs yield unbiased results because the correlations are mostly small for U and P designs. The difference between the C design on the one hand and both U and P designs on the other is immense. Whereas unbiased estimates are almost guaranteed in all scenarios with the C design, biased results can safely be expected with U and P designs. The standard errors of the estimated utilities are accurately estimated for all designs, except when only unidimensional blocks are used (Block 4). In the latter case only C designs yield unbiased results.

The result configuration is given for the correlations of the factors in Table 3.4. While all designs can recover correlation estimates sufficiently accurate, C designs yield accurate correlation estimates (mean bias: .02) for all blocks, including block designs (2) and (4). Standard deviations between correlation estimates are lowest for C designs. Overall, thresholds are accurately estimated by all designs and designs except block composition (1). The results do not differ substantially when twelve items per factor are used.

### ***3.2.6.2.2 Latent Trait Score Recovery***

Results for the actual and empirical latent trait recovery for U and C designs are presented in columns labelled “Zero” in Table 3.5. Overall, the empirical recovery

underestimates the actual recovery, as was also shown by Brown and Maydeu-Olivares (2011). For blocks with similar latent utilities (1), random blocks (3), and multidimensional blocks (5), the advantage of the C design is small. However, for block designs (2) and (4), a C design performs considerably better than a U design. It can be seen by comparing the upper part with the lower part of Table 3.5 that the higher the number of items per factor is, the better are the latent trait recoveries.

### ***3.2.6.2.3 Goodness-of-Fit Tests***

Empirical rejection rates per block design are depicted in Figure 3.6. Empirical rejection rates vary by the different block designs. For designs with similar utilities, a random assignment of items and multidimensional designs, U and P design rejection rates are in general larger than the target rejection rates. For designs with similar loadings, U and P designs reveal much smaller rejection rates than the target values and for unidimensional blocks U designs yield smaller and P design in tendency larger rejection rates, than the corresponding target values.

In contrast, rejection rates are closer on target for C designs, but they are still slightly higher than their target values, especially with a lower number of items (30 instead of 60). However, in comparison to the rates reported by Brown and Maydeu-Olivares (2011), the rejection rates in the present study are only slightly off target.

### ***3.2.6.3 Discussion***

The second simulation study shows that, overall, the traditional unlinked block design yields precise results only for multidimensional blocks if the parameters are rescaled per block. Again, the results show that U designs have higher (unacceptable) relative biases and lower correlations between true and estimated scores than C designs. As was the case in study 1, just presenting more blocks in P designs does not yield more precise estimates when multiple factors are given. C designs yield overall satisfactory results for all block



**Table 3.3**

*Results for Valid Iterations, Loading Estimates, and Standard Errors in the Simulation Study with Five Uncorrelated Traits.*

Blocks	<i>n</i>	Convergence						Loadings ARB						Loadings recovery		
		IRT			factor			Estimates			<i>SE</i>			<i>U</i>	<i>P</i>	<i>C</i>
		<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>			
Block 1	30	1000	557	1000	995	1000	1000	.00	.03	.00	-.03	-.03	-.02	.69	.06	.94
Block 2	30	237	241	928	249	301	924	-.06	.32	.00	-.29	-.29	-.09	.24	.55	.89
Block 3	30	972	746	1000	884	901	997	-.01	.09	.00	-.08	-.08	-.03	.60	.11	.94
Block 4	30	15	69	999	2	18	973	-.02	-.07	.00	2.17	1.66	-.08	.13	.15	.90
Block 5	30	967	688	1000	962	971	1000	.01	.26	.00	-.01	-.01	-.04	.57	.23	.93
Block 1	60	1000	904	1000	1000	999	1000	.00	.08	.00	-.01	-.01	.01	.66	.25	.94
Block 2	60	197	207	924	220	270	925	-.08	.16	.00	-.42	-.34	-.14	.13	.36	.87
Block 3	60	965	979	1000	958	971	1000	.02	.02	.00	-.02	-.02	-.02	.42	.08	.93
Block 4	60	196	206	1000	12	129	998	.01	-.08	.00	.18	.23	-.01	.14	.05	.93
Block 5	60	988	686	1000	986	986	1000	.01	.24	.00	-.02	-.02	-.03	.54	.13	.91

*Note.* *n* = number of items, *U* = unlinked block designs, *P* = partially linked block designs, *C* = completely linked block designs, ARB = average relative bias.

**Table 3.4**

*Results for Utility, Factor Correlation Estimates and Standard Errors in the Simulation Study with Five Uncorrelated Traits.*

Blocks	<i>n</i>	Utility Means ARB						Utility Means recovery			Factor correlations					
		Estimates			<i>SE</i>			Estimates			Mean			<i>SD</i>		
		<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>
Block 1	30	-.16	-.15	-.01	.01	.01	-.01	.10	-.32	1.00	-.01	-.01	-.01	.08	.08	.08
Block 2	30	.00	.00	-.02	.01	-.01	-.01	.59	.53	.99	.50	.53	-.04	.41	.47	.24
Block 3	30	-.01	-.01	-.01	-.01	.00	-.01	.86	.87	1.00	-.02	-.02	-.01	.13	.14	.09
Block 4	30	.01	.00	-.03	2.36	1.37	-.01	.38	.51	.99	-.10	.01	-.01	.19	.32	.11
Block 5	30	.00	.00	-.01	.02	.03	.01	.54	.52	1.00	-.02	-.02	-.02	.12	.12	.11
Block 1	60	-.20	-.18	-.02	.01	.01	.02	-.21	-.24	1.00	-.01	-.01	-.01	.06	.06	.05
Block 2	60	.00	.00	-.02	.01	.01	-.02	.57	.54	.99	.44	.50	-.09	.44	.45	.24
Block 3	60	.00	.00	-.01	-.01	-.01	-.01	.53	.55	.99	-.02	-.02	-.01	.09	.08	.07
Block 4	60	-.05	-.02	-.04	.51	.20	.01	.69	.71	.99	-.01	-.01	-.01	.26	.23	.06
Block 5	60	.01	.00	.00	.00	.00	-.02	.52	.51	.99	-.02	-.02	-.01	.07	.06	.06

*Note.* *n* = number of items, *U* = unlinked block designs, *P* = partially linked block designs, *C* = completely linked block designs, ARB = average relative bias.

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

**Table 3.5**

*Results for the Actual and Empirical Latent Trait Recovery for Five Uncorrelated (Simulation Study 2) and Five Correlated Traits (Simulation Study 3).*

	<i>n</i>	Actual Recovery				Empirical Recovery			
		Zero		Non-Zero		Zero		Non-Zero	
		<i>U</i>	<i>C</i>	<i>U</i>	<i>C</i>	<i>U</i>	<i>C</i>	<i>U</i>	<i>C</i>
Block 1	30	.72	.74	.73	.76	.52	.67	.55	.74
Block 2	30	.47	.71	.46	.72	.45	.63	.42	.59
Block 3	30	.66	.71	.72	.73	.29	.63	.50	.68
Block 4	30	.20	.66	.28	.62	-	.30	-	.18
Block 5	30	.73	.75	.73	.75	.54	.72	.56	.74
Block 1	60	.81	.83	.82	.84	.75	.84	.78	.86
Block 2	60	.53	.78	.53	.81	.62	.79	.58	.81
Block 3	60	.80	.81	.80	.83	.74	.82	.76	.86
Block 4	60	.33	.74	.46	.75	-	.63	-	.68
Block 5	60	.81	.82	.82	.84	.76	.84	.78	.86

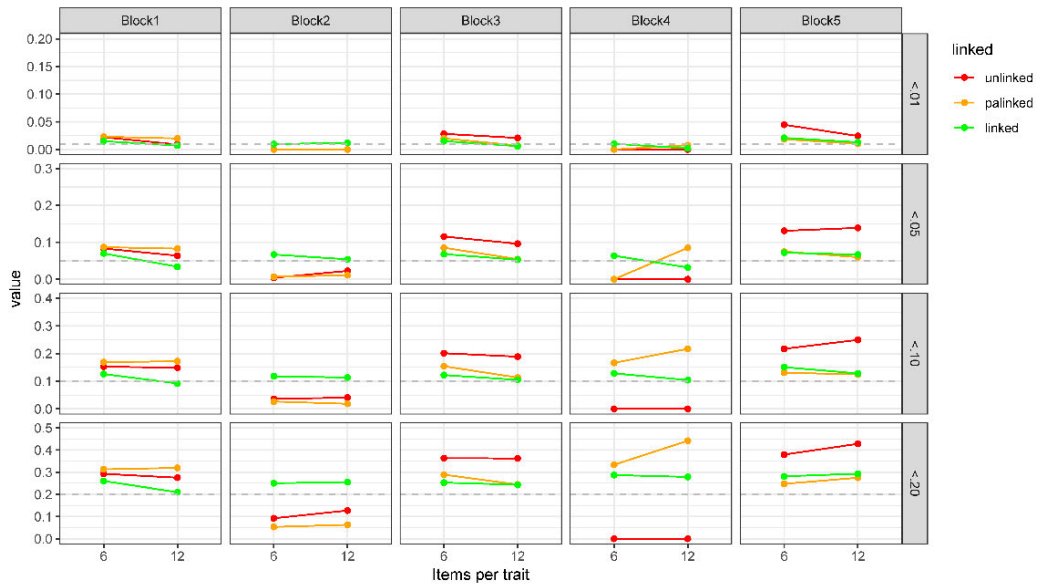
*Note.* The table shows results both from simulation study 2 and 3. *n* = number of items, Zero = results from simulation study 2 where all 5 factors have zero correlation, Non-Zero = results from simulation study 3 where all 5 factors have non-zero correlations, *U* = U design, *C* = C design.

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

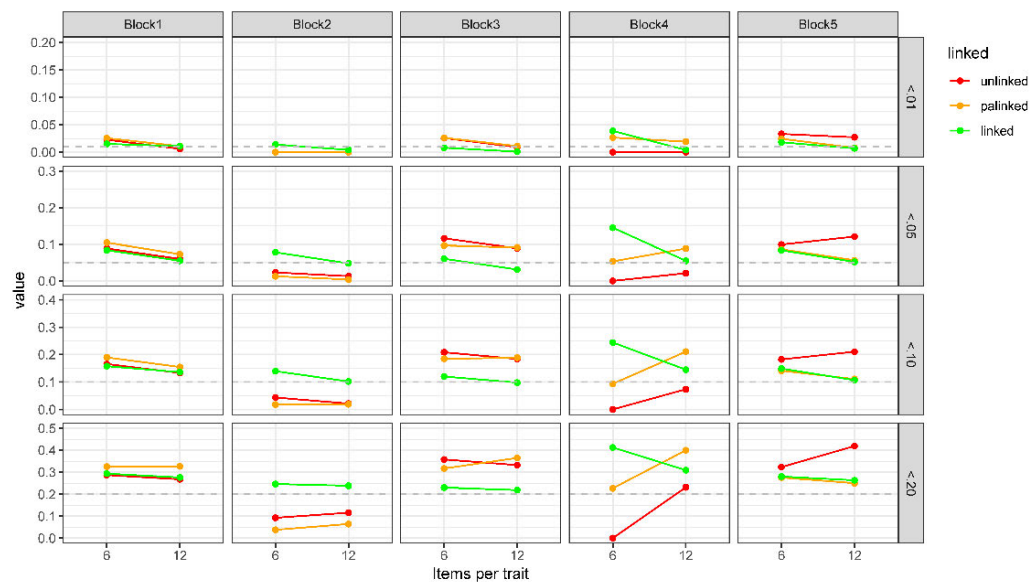
**Figure 3.6**

*Empirical Rejection Rates for the Simulation Studies with Five Factors.*

**A**



**B**



*Note.* The figure show results both from simulation study 2 (A) and 3 (B). Degrees of freedom are adjusted by the number of redundancies.  $n$  = number of items, unlinked = U design, palinked = P design, linked = C design.

compositions. The latent trait recovery (and therefore reliability) is small, however. It should be kept in mind, though, that compared to the analyses of Brown and Maydeu-Olivares (2011), the block designs are restricted to only positively keyed items in this study.

In sum, the C design performs at least as good, and in most cases much better, than both the P and U design. This is true for all the quality criteria under investigation. As a result, it is not only the design of choice in a single latent trait situation, but also with multiple (uncorrelated) factors.

### ***3.2.6.3 Simulation Study 3: A Forced-Choice Questionnaire Measuring Five Correlated Traits***

In the third simulation study, five correlated traits were used. All other parameters of the simulation study were identical to simulation study 2. For comparability to Brown and Maydeu-Olivares (2011) we used the same values for the Big Five factors as true correlations. These are: -.21, 0, -.25, -.53, .40, 0, .27, 0, 0, .24.

#### ***3.2.6.3.1 Item Parameter Recovery***

The results are very similar to the results of simulation study 2. Thus, they are not repeated here in detail but are available as online supplement. One main difference in the results is that the number of convergences for unidimensional blocks is considerably higher for correlated models as compared to uncorrelated models (around 900). However, this does not coincide with smaller biases in parameter estimation. As the data was simulated with correlated traits, we will focus on the results for the correlations (Table 3.6). In general, bias and SEs for the correlations are small only with C designs (mean bias: .01). For U and P designs, block designs (2) and (4) result in highly biased and therefore unacceptable estimates. Empirical rejection rates are again depicted in Figure 3.6. Overall, the rates are unacceptable for P designs and slightly higher than expected for U and C designs. Rates for C

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

**Table 3.6**

*Results for the Trait Correlations in Simulation Study 3.*

Blocks	<i>n</i>	Factor correlations					
		Bias Estimate			Bias SE		
		<i>U</i>	<i>P</i>	<i>C</i>	<i>U</i>	<i>P</i>	<i>C</i>
Block 1	30	-.01	-.01	.00	.14	.13	.11
Block 2	30	-1.42	-1.82	.01	.54	.46	.21
Block 3	30	.04	.06	.06	.13	.12	.11
Block 4	30	-1.03	-1.02	.01	.29	.25	.11
Block 5	30	-.01	-.02	.03	.12	.11	.10
Block 1	60	.04	.03	.05	.06	.06	.05
Block 2	60	-1.87	-2.31	.06	.41	.38	.14
Block 3	60	-.02	-.03	-.01	.07	.07	.06
Block 4	60	-1.12	-1.02	-.03	.24	.25	.06
Block 5	60	.01	-.02	.01	.06	.06	.05

*Note.* *n* = number of items, *U* = U design, *P* = P design, *C* = C design.

designs are good except for a level of .20. Again, rejection rates are much better than reported by Brown and Maydeu-Olivares (2011).

**3.2.6.3.2 Latent Trait Score Recovery**

Results for the latent trait recovery can be found in Table 3.5. Overall, the empirical recovery underestimates the actual recovery, as was also shown by Brown and Maydeu-Olivares (2011). For multidimensional blocks, random blocks, and blocks with similar latent utilities, the advantage of the TLB design is small. However, for block designs 2 and 4, the TLB design performs considerably better than the TB design. For the latent trait recovery, the higher the number of items per factor is, the higher are the recoveries.

### **3.2.6.3.3 Discussion**

The overall results in the third simulation study are not fundamentally different from the second simulation study. Nevertheless, since correlated factor models are very common in practice, it seemed necessary to include a simulation study with such a factor model in addition to simulation study 2 to confirm that results are indeed similar.

For correlated models, the relative bias is in tendency smaller than for models with uncorrelated traits. However, this benefit appears to be small and not systematic. There also are some parameters that are estimated with a larger relative bias. U designs seem to recover the correlation between traits only in some block compositions, whereas C designs yield precise estimates independent of the block composition. In sum, conclusions and recommendations do not differ to those of simulation study 2.

### **3.2.7 Overall Discussion**

In the current article, the Thurstonian linked block (TLB) design was introduced. It is an adaptation of the Thurstonian block (TB) design and its special case, the multidimensional forced-choice (MFC) format proposed by Brown and Maydeu-Olivares (2011). While, so far, the TB design was only discussed within an IRT context and has rather limited use in a factor analytic setting (Jansen & Schulze, 2023a), the TLB can be used in both settings. With respect to the number of traits, the use of uni- and multidimensional blocks, and the number of items per block the TLB is highly flexible and has almost no restrictions. Unfortunately, in comparison to hitherto used designs, the use of the TLB comes at the cost of a necessity for more blocks. In order to reap the benefits of the TLB design somewhat longer assessments are needed because the stimuli in a design have to be linked. A formula was given in the present paper that can be used to calculate the number of additional blocks needed (if only same-sized blocks are used). The additional cost of the TLB design can therefore easily be quantified. It is argued, though, that the design is generally worth this price.

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

Via simulation studies, it was demonstrated that traditional U designs with unlinked blocks, estimated by a TB design, yield biased results, even when multidimensional blocks are used. It was also shown that simply adding any blocks of already used items does not substantially improve the recovery. Hence, the benefits of the TLB are not trivially due to higher test length. An important result of the simulation studies was that with linked blocks (i.e., C designs) true parameters can accurately be recovered. Considering the recovery of the latent traits, the TLB design with linked blocks generally outperforms the traditional TB designs, when disregarding some special cases (the difference between loadings within a block is small) where the benefit is small if only positively keyed items are used.

With respect to the absolute goodness-of-fit test, all designs and designs yield at least slightly higher rejection rates than permissible, especially for quite large significance levels. However, again the TLB designs show the best performance and yield rejection rates closest to the target values. In general, all designs show less aggravated rates as compared to the results reported by Brown and Maydeu-Olivares (2011). This last result is somewhat surprising: The results reported here show that with respect to fit estimation, the TB designs perform much better if all possible response variables are simulated, in contrast to the case where the specific TB design's response variables are simulated.

At present, the most commonly used block format still is triplets. To reduce the number of blocks a respondent needs to work on, using quads or pentads could be considered. While the number of items seems not to inflict a reduction in test motivation (Sass et al., 2020), the cognitive workload to respond to one block naturally gets higher. At the same time, even the use of a block with  $k = 15$  seem to be a valid possibility and, yield similar results compared to a paired comparison task (see Jansen & Schulze, 2023b). It can be suspected that it may not only simply be the number of stimuli in a block that increases the workload but the type of stimuli and the complexity of the criterion that respondents have to



## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

use to establish a ranking. In any case, the use of any block size and designs with varying linked block sizes is straightforward with the TLB design, too.

Some additional remarks on how to link blocks in the TLB design seem to be in order. As mentioned before, there are two possibilities to link multiple blocks so that the design matrix  $\mathbf{A}$  has a rank of  $n - 1$ . For example, assume blocks  $B_1 = (1,2,3)$ ,  $B_2 = (4,5,6)$ ,  $B_3 = (7,8,9)$ ,  $B_4 = (10,11,12)$ ,  $B_5 = (13,14,15)$ , and  $B_6 = (16,17,18)$  are given. Assume further, that a)  $B_7 = (1,4,7)$  connects  $B_1$ ,  $B_2$ , and  $B_3$ , and, b)  $B_8 = (10,13,16)$ , connects  $B_4$ ,  $B_5$ , and  $B_6$ . Now,  $B_7$  and  $B_8$  have to be linked. One way is direct linking by using two items that are included in the linking blocks and for the third slot in a triplet any other item could be used, like  $B_9 = (4,13,18)$ , for example. Considering test motivation of the respondents, it appears not to be far-fetched to assume that the repeated presentation of the same items, even if they occur in different blocks, may be exhausting and result in detrimental effects on test motivation. One advantage of the TLB is that such a construction is not necessary. On the contrary, to connect  $B_1$ ,  $B_2$ ,  $B_3$ , with,  $B_4$ ,  $B_5$ , and  $B_6$ , it is sufficient to use an indirect linking. Any item that is in one of the blocks, for example,  $B_9 = (3,12,15)$  would do. For reasons of test motivation, it seems advisable to consider such indirect TLB designs in test construction and design.

### ***3.2.7.1 Software Implementation***

With regard to the implementation of Thurstone models in SEM software, another advantage of the TB design is that no dependencies between stimuli in different blocks exist. The code to implement the models is therefore highly structured and less error-prone. While the general strategy of implementation is the same, the additional loadings, uniquenesses, and correlations between uniquenesses, plus the equality constraints needed, result in the need for an automated code generation system. There already exist useful Excel syntaxes (Brown & Maydeu-Olivares, 2012) and R packages (Bürkner, 2019). However, they do not include the

TLB design. To accomplish automated syntax creation the R package *Thurmod* is a viable candidate (Jansen, 2023a). This package is suitable to simulate data, write codes for *Mplus* and *lavaan*, read *Mplus* and *lavaan* outputs, determine the number of redundancies, and has even more functionalities. With this tool, the analysis of any Thurstonian model as described here is made easily available.

### 3.2.7.2 *Direction of Items*

As mentioned repeatedly, only positively keyed items were used in the present study to account for the critique put forth against mixed keyed item blocks. The main argument is that it is highly implausible to be able to construct blocks with items of equal or at least similar desirability values and simultaneously include positively and negatively keyed items (Bürkner et al., 2019; Jansen & Schulze, 2023a). At the same time, the present results and the results of previous studies too (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010) show that using only unidirectionally keyed items per block has the consequence of lower recovery rates for the latent trait scores. The TLB design offers a potential solution to this problem. Note that in a TB design, if only unidirectionally keyed items are used within a block, the information per block is rather small. With the TLB design, though, linking blocks does automatically lead to the inclusion of blocks with both negatively and positively keyed items, as otherwise, it would not be possible to create a design with rank  $n - 1$ . A viable strategy for test construction with the TLB design would be to have initial blocks being matched for desirability and fakeability. Some of the linking blocks can then serve two purposes: First, they yield information about the maximum relative difference and should improve the latent trait recovery this way. Second, they can function as attention control checks in high stakes situations. A detailed analysis of the use and benefit of such linkage blocks is given by Jansen and Schulze (2023c).

### ***3.2.7.3 Incomplete Ranking***

Similar to the TB design, incomplete rankings can be modelled in the TLB design. Incomplete rankings may occur, if blocks of size  $k = 4$  are used, for example, but respondents are only asked to mark the alternatives that are most and least fitting to themselves. In such cases, no information on the comparisons in the middle portion of the ranking would be available. This information is missing by design. Another application would be a full design where all possible response variables are of interest but only a subset of paired comparisons or rankings is presented. These cases could be present, for example, if sorting algorithms are used to determine the preferences of each respondent by assuming the transitivity of rankings, but after  $x$  rankings participation is aborted. One sorting algorithm that neatly fits into the use of FC designs in general is Binary or N-ary Path Sort (Jansen, 2023b; Jansen & Doeblér, 2023). The algorithm has two steps: first, it builds a  $k$ -ary tree, where  $k$  is the number of items per block, and second, it sorts all elements of the tree by presenting only blocks that are not determinable via transitivity. With a TLB design, the first step is already accounted for, as the linking of blocks allows for building the individual  $k$ -ary tree. With incomplete rankings, that is missings by design, one should still be able to estimate Thurstonian models appropriately (Jansen, 2023b; Jansen & Doeblér, 2023).

### ***3.2.7.4 Conclusions***

The TLB design introduced in the current study is an adaptation of the Thurstonian block designs. Therefore, it is also of interest for any situation where forced choice assessments constructed with the MFC format are used. The TLB designs can very flexibly be used and have a number of desirable properties outlined in the present paper. Using the TLB design and considering a matching of stimuli by fakeability may help the forced choice method to live up to its promise of a reduction or even elimination of response biases and faking. It was shown here that the TLB design not only has desirable features and potential

but also that it is superior to its predecessors, albeit at a somewhat higher expenditure associated with its use. Overall, the results reported in this paper are in line with the more theoretically derived features by Jansen and Schulze (2023a). The main conclusion is that the use of unlinked block designs cannot be recommended anymore and they should be replaced by linked block designs, even at the cost of somewhat longer assessments. The TLB is therefore generally recommended for use in research and applications with forced choice assessments.

### **3.2.7.5 Limitations**

Of course, the conclusions in this article that are based on the results from the simulation studies have to be qualified due to some limitations. No matter how lavish the design of a simulation study, there are always design factors or levels missing that might have delivered useful insights. This is indeed the case in the current work. All simulation studies were run with a sample sizes of 2000 respondents. For the purposes of comparisons between the models that are focused here, this sample is deemed appropriate. At the same time, it is conceded that varying the sample size would have provided interesting information for convergence rates of the models and the precision of parameter estimates, for example. In addition, a sample size of 2000 or larger is unfortunately not very common in psychological research, so the results reported here may not generalize to studies with samples more typical in psychological research. Further limitations pertain to other design factors, such as the number of traits and their correlational structure as well as the number of items, in particular. Hence, investigating the (relative) performance of the TLB with smaller sample sizes and more varied assessment designs would certainly be a worthwhile endeavor. However, it is not expected that the conclusions drawn here would change with smaller sample sizes or a different number of items.

### 3.2.8 References

- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502.  
<https://doi.org/10.1177/0013164410375112>
- Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147.  
<https://doi.org/10.3758/s13428-012-0217-x>
- Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software, 4*(42), 1662–1663.
- Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854. <https://doi.org/10.1177/0013164419832063>
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475–494.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625–641.  
<https://doi.org/10.1080/10705510903203573>
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252. <https://doi.org/10.1037/h0045996>
- Jansen, M. T. (2023a). *Thurmod: An R package for Thurstonian modeling*. Package submitted.
- Jansen, M. T. (2023b). *N-ary Path Sort: Adaptively sorting stimuli via forced-choice blocks of size N*. Manuscript in preparation.

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

- Jansen, M. T., & Doebler, P. (2023). *Binary Path Sort: An adaptive sorting algorithm for forced-choice designs with focus on a single respondent*. Manuscript in preparation.
- Jansen, M. T., & Schulze, R. (2023a). *Linear factor analytic Thurstonian forced-choice models: Current status and issues*. Manuscript submitted.
- Jansen, M. T., & Schulze, R. (2023b). *Item scaling of social desirability using conjoint measurement: A comparison of ratings, paired comparisons, and rankings*. Manuscript in preparation.
- Jansen, M. T., & Schulze, R. (2023c). *Forced-choice with Thurstonian models: Assessment design and the role of item keying*. Manuscript submitted.
- Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64*(3), 325–340.  
<https://doi.org/10.1007/BF02294299>
- Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*(3), 285–304.  
<https://doi.org/10.1037/1082-989X.10.3.285>
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioural Research*, *45*(6), 935–974.  
<https://doi.org/10.1080/00273171.2010.531231>
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. <https://doi.org/10.1007/BF02293813>
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Routledge.
- Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker’s perspective: Response process and test motivation in multidimensional forced-choice versus rating

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

scale instruments. *Assessment*, 27(3), 572–584.

<https://doi.org/10.1177/1073191118762049>

Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats?. *Educational and Psychological Measurement*, 81(2), 262–289.

<https://doi.org/10.1177/0013164420934861>

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

Thurstone, L. L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology*, 14(3), 187–201. <https://doi.org/10.1037/h0070025>

Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1), 79–92. <https://doi.org/10.1348/000711099158973>

Ziegler, M., MacCann, C., & Roberts, R. (2012). *New perspectives on faking in personality assessment*. Oxford University Press.

### 3.2.9 Appendix A

Number of redundancies in a ranking task for Thurstonian forced-choice designs

In the following, further detail is provided on how to determine the number of redundancies of any forced-choice ranking design, that is, if transitivity is assumed. To this end, the derivation for full designs as provided by Maydeu-Olivares (1999; Appendix C) is revisited. The final arguments of the proofs in this important work used  $\tilde{n}$  as the argument of the number of dichotomous random variables. This is discussed here for any number of response variables  $c$ .

As has been done in the main text, a forced-choice design matrix  $\mathbf{A}$  is defined as a  $c \times n$  design matrix where the rows of  $\mathbf{A}$  correspond to the paired comparisons and the columns of the matrix correspond to the items. Only in a full design, it is  $c = \tilde{n} = n(n-1)/2$ . In a full design, the number of redundancies was determined to be as given in Equation (2.16), that is

$$r = \frac{n(n-1)(n-2)}{6}.$$

The distribution of the first stage estimator is asymptotically normal (Maydeu-Olivares, 1999; Appendix C).

Now, consider vector  $\mathbf{y}$  to consist of  $c$  dichotomous random variables  $y_l = \{0,1\}$  for all  $l$  in  $1 \dots c$ . A  $2^c$  contingency table can now be constructed and from that  $\mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix}$  is constructed, where  $\mathbf{p}_1$  consist of the first order marginals and  $\mathbf{p}_2$  of the nonredundant second order marginals of the contingency table from a random sample. Further,  $\boldsymbol{\pi}$  represents the true marginals.

Assume that the true probability for a response variable comes from a dichotomization of a normal density



Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

$$\Pr(y_b) = \int_{\dots} \int_R \phi(\mathbf{z}^* : \boldsymbol{\mu}_{z^*}, \mathbf{P}_{z^*}) d\mathbf{z}^*, \quad b = 1, \dots, 2^c \quad (3.37)$$

where  $R$  is the area of integration defined by (3.37). Define  $\hat{\mathbf{P}}_{\boldsymbol{\mu}_{z^*}}$  to be a vector that stacks all lower diagonal elements of  $\mathbf{P}_{\boldsymbol{\mu}_{z^*}}$  on a column vector. Assuming  $c = \tilde{n}$  and given alternative identification restrictions, it was shown that with the transformation  $\hat{\boldsymbol{\kappa}}$  of  $\mathbf{p}$

$$\hat{\boldsymbol{\kappa}} = \begin{pmatrix} \boldsymbol{\mu}_{z^*} \\ \hat{\mathbf{P}}_{\boldsymbol{\mu}_{z^*}} \end{pmatrix} \quad (3.38)$$

it is

$$\sqrt{N}(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma} = \boldsymbol{\Lambda} \mathbf{H} \boldsymbol{\Xi} \mathbf{H}' \boldsymbol{\Lambda}'), \quad (3.39)$$

where  $\boldsymbol{\Lambda} = \begin{pmatrix} \partial \boldsymbol{\pi} \\ \partial \boldsymbol{\kappa}' \end{pmatrix}_{\boldsymbol{\kappa}=\boldsymbol{\kappa}_0}^{-1}$  and  $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\pi}_r) - \boldsymbol{\pi}'_r \boldsymbol{\pi}_r$  (Maydeu-Olivares, 1999; Muthén, 1978). This

result is important, as the first-stage estimator is asymptotically normal and the degrees of freedom of the design can directly be determined by the rank of  $\boldsymbol{\Gamma}$ , which is equal to the rank of  $\mathbf{H}$  (Maydeu-Olivares, 1999). For the construction of  $\mathbf{H}$  a  $n! \times \tilde{n}$  matrix  $\mathbf{Y}$  is considered that contains all ranking patterns in binary form, which are possible with the design  $\mathbf{A}$ . For example, for  $n = 3$  it would be

$$\begin{matrix} 1 & 2 & 3 \\ 1 & 3 & 2 \\ 2 & 1 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \\ 3 & 2 & 1 \end{matrix} \Rightarrow \mathbf{Y} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad (3.40)$$

Now,  $\mathbf{H}$  is defined by

$$\begin{aligned} \mathbf{h}'_l &= \mathbf{y}'_l, \forall \mathbf{h}'_l \in \mathbf{H}_1 & l &= 1, \dots, c \\ \mathbf{h}'_{l,l'} &= \mathbf{y}'_l \odot \mathbf{y}'_{l'}, \forall \mathbf{h}'_{l,l'} \in \mathbf{H}_2 & l &= 2, \dots, c; l' = 1, \dots, c-1; \end{aligned} \quad (3.41)$$

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

where  $\odot$  is the Hadamard product. That is, we need  $\mathbf{Y}$  and the Hadamard product of every possible column combination of the  $c$  variables. For our example it is

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}' \quad (3.42)$$

The rank of  $\mathbf{H}$  in (43) is 5 and we have  $1 = 3 \times 2 \times 1 / 6$  redundancies. Obviously, each redundancy for  $n$  items is also a redundancy in any  $v$  items with  $v > n$ . The assumption  $c = \tilde{n}$  in Maydeu-Olivares (1999) is only necessary, if  $\mathbf{H}$  is represented by matrix  $\mathbf{H} = \mathbf{GT}$ , where  $\mathbf{G}$  is constructed similarly to (3.41) but based on a  $2^c \times c$  matrix  $\mathbf{X}$  containing all possible ranking patterns in binary form, including those that are not possible by design. To compensate,  $\mathbf{T}$  is a  $2^c \times n!$  augmentation matrix with zeros in appropriate positions, so that all non-possible patterns are eliminated.

For dichotomous response variables it is easy to obtain the redundancies by computing a basis for the nullspace of  $\mathbf{H}$ . However, there is a more direct connection between the redundancies and the transitivities in ranking. By definition, the redundancies are the dependencies in the linear equations of the mapping defined by  $\mathbf{H}$ . These are caused by the implied paired comparisons due to transitivity. For each implied transitivity (unique up to order) there is exactly one redundancy among the linear equations defined by the thresholds and tetrachoric correlations in  $\mathbf{H}$ . In turn each redundancy is a result of the two missing columns in  $\mathbf{H}$  that are inversed (equal up to order) and that correspond to intransitive responding. For the example of  $n = 3$ , Equation (3.40) is considered, where rows three and five define the same transitivity, only in reversed order.

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

With the correspondence of redundancies and transitivities, it is sufficient only to count the number of all combination of each three items without considering the order, which simply is

$$\binom{n}{3} = \frac{n!}{3!(n-3)!} = \frac{n(n-1)(n-2)}{6}. \quad (3.43)$$

More importantly, we know exactly which combination of response variables form the redundancies, as it is every combination of three items that imply one transitivity. For  $n = 3$ , there is only one transitivity (unique up to order) which is implied by  $1 > 2$  and  $2 > 3$ . For  $n = 4$ , there are four possible transitivities as given by

$$\begin{aligned} &1 > 2, 2 > 3 \\ &1 > 2, 2 > 4 \\ &1 > 3, 3 > 4 \\ &2 > 3, 3 > 4 \end{aligned} \quad (3.44)$$

Now, assume  $c < \tilde{n}$ . Most importantly, Equations (3.39) and (3.41) still hold. It will now be shown that for a Thurstonian block design, it is sufficient to calculate the redundancies per block. If a redundancy is present in a block where only a part (the blocks) of the  $c$  response variables is considered, then it must also be present when all  $c$  response variables are considered (each block can be mapped on a subspace with zeros at appropriate positions). We need to show that the opposite is also true. As mentioned before, the redundancies are implied by the transitivities. First, the set of all transitivities that is implied by  $\tilde{n}$  is constructed. However, there are combinations of dichotomous variables that are not present in the specific set of  $c$  response variables. As there is no linkage between the blocks, all dichotomous variables including combinations of items between blocks must be discarded. This implies, that there are only valid transitivities (and therefore redundancies) for the items within one block, hence the assertion. For linked blocks this is also true as long as each dichotomous variable is present in

Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

no more than one block. Otherwise, the paired comparisons that are present in multiple blocks, imply transivities across blocks.

To illustrate, items 1 to 5 are taken for Example 1 as well as the two blocks  $B_1 = (1,2,3)$  and  $B_2 = (3,4,5)$ . All possible transivities assuming all paired comparisons are given in the first column of Table 3.7.

**Table 3.7**

*Illustration of Redundancies for Different Scenarios.*

<i>Transivities</i>	<i>Example 1</i>	<i>Example 2</i>	<i>Example 3</i>	<i>Example 4</i>	<i>Example 5</i>
1>2,2>3	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>no</i>
1>2,2>4	<i>no</i>	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>no</i>
1>2,2>5	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>no</i>
1>3,3>4	<i>no</i>	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
1>3,3>5	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>no</i>	<i>yes</i>
1>4,4>5	<i>no</i>	<i>no</i>	<i>no</i>	<i>no</i>	<i>no</i>
2>3,3>4	<i>no</i>	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
2>3,3>5	<i>no</i>	<i>yes</i>	<i>yes</i>	<i>no</i>	<i>yes</i>
2>4,4>5	<i>no</i>	<i>no</i>	<i>no</i>	<i>no</i>	<i>no</i>
3>4,4>5	<i>yes</i>	<i>yes</i>	<i>no</i>	<i>no</i>	<i>no</i>

All transivities with a “no” are not possible. For example 2, we take blocks  $B_1 = (1,2,3)$ , and  $B_2 = (2,3,5)$ . There are more than two transivities possible, even though item 4 is not included in the design (Example 2; Table 3.7). No transitivity including item 4 is possible, while all others are. This directly implies that in cases where some dichotomous variables are present in more than one block, there is no closed form solution to the calculation of the number of redundancies. This can be shown by a direct counterexample. First assume we

## Study 2: The Thurstonian Linked Block Design: Improving Thurstonian Modeling for Paired Comparison and Ranking Data

---

have a full design. Then remove the comparison  $y_{5,4}$ , so no redundancy including this comparison is possible (Example 3; Table 3.7) resulting in seven redundancies. If now,  $y_{5,3}$  is excluded, five redundancies are given (Example 4; Table 3.7). If instead  $y_{2,1}$  is excluded, four redundancies are given (Example 5; Table 3.7). Therefore, there is no direct closed form, as it is dependent on the specific design. However, the number of redundancies is easily determinable.

## **4. Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying**

Jansen, M. T., & Schulze, R. (2023d). *Forced-choice with Thurstonian models: Assessment design and the role of item keying*. Manuscript submitted for publication in *Multivariate Behavioral Research*.

### **4.1 Summary**

#### **4.1.1 The Information Obtainable from Forced-Choice Responses**

The preceding manuscript discussed the presentation of the Thurstonian linked block design, but also had some limitations in its simulation study that limit its transfer to real-world empirical data. The next manuscript continues where the last one left off and focuses on the information obtainable from FC responses. Broadly, there are three main perspectives on this information. First, the Fisher information matrix, that includes information for reliable latent trait score recovery. Second, an optimal item design through A- and/or D-optimality, and third, information on the items' position on the true ranking per participant. To eliminate fakeability, FC blocks should be constructed by matching items by desirability. It is also noted that the direction of traits is arbitrary and dependent only on the interpretation given by psychological theories (e.g., neuroticism can be interpreted as reversed score emotional stability). However, this interpretation of a trait's direction does not affect the information obtained by a corresponding paired comparison.

If all sources of information are considered, it seems necessary to include both same- and mixed-keyed blocks. The Thurstonian linked block design allows for this by linking blocks with large differences in item loadings, except if loadings of all items are very similar. For a test to be faking resistant, it is suggested to use a small number of mixed-keyed blocks.

#### **4.1.2 Simulation Study**

A simulation study was conducted to compare the unlinked with the linked block design. Simulation study parameters are the same as in study 2, that is the number of traits (one, three, and five), the number of items per trait (6, 12, and 18), and the correlation between traits (uncorrelated or correlated). Additionally, the sample size (200, 500, 1000), and number of negatively keyed items per trait (1 or 2) is varied. More negatively keyed items result in more mixed-keyed blocks. The simulation was repeated 1000 times for all conditions. Three different block designs were constructed. Both, the Thurstonian factor and IRT models are considered to obtain results on both item and person parameters.

#### **4.1.3 Results**

The results of the simulation study show that in general, the more items a test has, the more accurate the results are. Overall, the results of study 2 are supported, the linked block designs outperform the unlinked designs considering recovery and convergence rates, as well as bias and recoveries. Results for linked block designs are excellent for larger sample sizes, but not for unlinked designs. When only 200 respondents were considered, results are not acceptable for any of the designs or model types. Empirical rejection rates are acceptable for linked blocks and at least 500 respondents. The inclusion of a second negatively keyed item per trait does not significantly enhance model and parameter estimation.

#### **4.1.4 Recommendations for FC Test Construction**


To reduce fakeability, it seems important to define all traits in the same direction of desirability. Desirability values for each item should be obtained by desirability scaling and these values should be used to match initial blocks by desirability. Additionally, all initial blocks should be linked so that  $\mathbf{A}$  has rank  $n - 1$ . It is also recommended to use at least 500 respondents and item sets of 12 items per trait. Finally, the construction of multidimensional blocks is advised, but it is not detrimental for the results to use some unidimensional blocks.


## 4.2 Manuscript

### **Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying**

Markus T. Jansen and Ralf Schulze  
University of Wuppertal

#### **Author Note**

Markus T. Jansen  <https://orcid.org/0000-0002-5162-4409>

Ralf Schulze  <https://orcid.org/0000-0001-5780-8973>

Supplement material available at

[https://osf.io/x3rdw/?view\\_only=caec86375b42492598de02fdf7fb5ec6](https://osf.io/x3rdw/?view_only=caec86375b42492598de02fdf7fb5ec6)

The author(s) declared no conflicts of interest with respect to the authorship or the publication of this article. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence concerning this article should be addressed to Markus Thomas Jansen, Institute of Psychology, University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany. E-mail: [mjansen@uni-wuppertal.de](mailto:mjansen@uni-wuppertal.de)



#### **4.2.1 Abstract**

The use of the self-report method in psychological research and applications is prone to response biases and faking. Reducing response distorting influences is an important goal in psychological assessment. The Thurstonian forced-choice block designs can eliminate some response biases and, with careful assessment design, intentional response distortions such as faking. Recent advances in Thurstonian modeling with the linked block design offer enhanced estimation precision of item and person parameters as compared to traditional non-linked block designs. The recent discussions about the use of both positively and negatively keyed items to construct mixed-keyed blocks is extended in the present paper. The focus is laid on the information that a paired comparison provides for the estimation of person parameters and the item position in the full ranking. It is highlighted that mixed-keyed blocks contribute viable information for model estimation but they are highly fakeable, especially in high-stakes situations. As an alternative, it is shown how less fakeable forced-choice questionnaires can be constructed with linked block designs by including blocks that are homogeneous regarding social desirability along with some mixed-keyed blocks. The results of simulation studies show that the inclusion of mixed-keyed blocks enhances model estimation in linked block designs. Recommendations for forced-choice test construction based on the recent advances on Thurstonian modeling are provided.

*Keywords:* Thurstonian modeling, forced-choice format, faking, Thurstonian linked block design, mixed-keyed blocks

---

#### 4.2.2 Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

A major part of the psychological assessment toolbox are self-reports in the form of rating scales, where respondents are asked to rate their agreement with or endorsement of a given stimulus (e.g., a statement). A common issue with such Likert-type items is the occurrence of response biases like the tendency to agree with statements irrespective of content (acquiescence bias), the tendency to very strongly agree- or disagree and avoid more nuanced response options (extreme responding), or to agree with statements solely because of their socially desirable content (social desirability bias, e.g., Cronbach, 1946; Jackson & Messick, 1958; Paulhus, 2002). Additionally, individuals may be motivated and able to fake responses in high-stakes situations in order to present themselves in a more favorable light (e.g., Ziegler et al., 2012). Both response biases and faking can be intentional or unintentional and significantly undermine the validity of the results obtained from rating scales.

In recent years, the multidimensional forced-choice (MFC) format (Brown & Maydeu-Olivares, 2011) has become increasingly popular as it proposes an effective way to reduce the effects of faking and response biases (see e.g., Bürkner et al., 2019; Cao & Drasgow, 2019; Frick et al., 2021; Schulte et al., 2021). In forced-choice (FC) formats, respondents must choose between stimuli according to some given criterion, no matter how hard a choice might be for a respondent (hence *forced* choice). This idea can be applied to two stimuli, as is common for classical FC. An example of such a paired comparison is provided in the top panel in Figure 4.1. The idea may also be extended to FC blocks of three or more items as has become standard in recently developed models and designs. In the latter case, respondents often only have to identify the stimulus that satisfies the criterion best and the one that is worst. An example with a triplet of three stimuli is shown in the middle panel in Figure 4.1. Alternatively, respondents can be required to give a full ranking on blocks of

**Figure 4.1**

*Examples for the Forced-Choice Format.*

Paired comparison	Please select the option that describes you the best. <input type="radio"/> In difficult situations I stay calm. <input type="radio"/> I like partys.														
Triplet	Select one statement that describes you MOST accurately and one that describes you LEAST accurately. <table border="1" data-bbox="574 537 1219 705"> <thead> <tr> <th></th> <th>Most Like Me</th> <th>Least Like Me</th> </tr> </thead> <tbody> <tr> <td>I like art.</td> <td></td> <td></td> </tr> <tr> <td>I seldom am in a good mood.</td> <td></td> <td></td> </tr> <tr> <td>I talk a lot.</td> <td></td> <td></td> </tr> </tbody> </table>				Most Like Me	Least Like Me	I like art.			I seldom am in a good mood.			I talk a lot.		
	Most Like Me	Least Like Me													
I like art.															
I seldom am in a good mood.															
I talk a lot.															
Quad	Please rank the options according to how well they describe you. <table border="1" data-bbox="574 739 1219 929"> <tbody> <tr> <td>I like art.</td> <td>1.</td> </tr> <tr> <td>I seldom am in a good mood.</td> <td>2.</td> </tr> <tr> <td>I talk a lot.</td> <td>3.</td> </tr> <tr> <td>I am orderly.</td> <td>4.</td> </tr> </tbody> </table>			I like art.	1.	I seldom am in a good mood.	2.	I talk a lot.	3.	I am orderly.	4.				
I like art.	1.														
I seldom am in a good mood.	2.														
I talk a lot.	3.														
I am orderly.	4.														

stimuli (see bottom panel in Figure 4.1). In any case, FC items do not require a response scale. This feature of FC items alone already eliminates many biases associated with response scales.

Using the FC method has its costs, though. The data collected with FC assessment is clearly only ordinal, which requires the use of appropriate specialized statistical methods to process the data. A general drawback of FC measurement is also that a single person's choices (e.g., rankings) only provides ipsative information, meaning that the results are only comparable within a person and not between individuals (for a discussion see Baron, 1996). However, recent methodological advances in the last two decades have made it possible to estimate normative trait scores allowing for score comparability between respondents using the Thurstonian IRT model for dominance response items (Maydeu-Olivares & Brown, 2010).

Despite all recent advances with the MFC format and the Thurstonian IRT model, there still remain some critical issues that require solutions in order to ensure the dependable

and efficient use of this FC approach in practical applications. Among the more concerning issues are, for example, that a) parameter and trait score estimates can be highly biased and b) it is nearly impossible to obtain interpretable item utilities (Jansen & Schulze, 2023a, 2023b). The latter aspect pertains to item (instead of person) scaling. Although this is an important use case of Thurstonian factor models, it has not drawn much attention in research and applications. This is somewhat surprising since item scaling methods can be used for social desirability scaling (Jansen & Schulze, 2023c; Maydeu-Olivares & Böckenholt, 2005), which in turn should be an integral part of FC assessment (see also Bürkner, 2022; Bürkner et al., 2019; Jansen & Schulze, 2023a; Schulte et al., 2021). With respect to the first concerning issue, it is noteworthy that some advances have recently been made. Of concern in the present study is the generalization of the Thurstonian block designs by introducing the Thurstonian linked block design (TLB; Jansen & Schulze, 2023b). It was demonstrated that the TLB design improves the usefulness, feasibility, and reliability of Thurstonian modeling. However, the use of both same and mixed-keyed blocks was neglected in previous research with the TLB design. Also, simulation studies with the TLB design only used sample sizes of 2000 respondents (Jansen & Schulze, 2023b). Both aspects reduce the generalizability of the simulation results to real world scenarios. The current study aims to address these limitations and to further investigate the usefulness of TLB models per se and in comparison to non-linked designs. Special attention will be paid to the use of same- and mixed-keyed blocks in combination with smaller sample sizes. The goal here is to extend and update extant recommendations on how to build an FC assessment and to propose an updated guideline for practitioners.

#### **4.2.3 Thurstonian Models**

While previous studies have predominantly focused on the MFC format and the IRT approach to Thurstonian modeling, the present study extends the discussion by also including

and focusing on item parameters, item utilities in particular. The models summarized in this section have already been extensively discussed by Brown and Maydeu-Olivares (2011; also see Jansen & Schulze, 2023a, 2023b; Maydeu-Olivares & Böckenholt, 2005).

Thurstonian models can be categorized by two dimensions: the general model class, such as factor versus IRT models, and the design such as full versus block designs. A full design includes all possible nonredundant paired comparisons of a given set of stimuli whereas block models include stimuli as subsets assembled into specific blocks as is shown for triplets and quartets in Figure 4.1. In theory, the possible designs that can be used in Thurstonian models are not restricted to the use of specific block sizes or even same-sized blocks. In practice, however, multidimensional triplets have by far been most often used to date. Some specific model classes like the simple Thurstonian model (without a factor structure) will not be discussed here. For details on this model class, the interested reader is referred to Maydeu-Olivares and Böckenholt (2005; see also Jansen and Schulze, 2023a). The general model of the Thurstone framework is given by

$$\mathbf{y}^* = \mathbf{A}\mathbf{t} + \mathbf{e}, \quad (4.1)$$

where  $\mathbf{y}^*$  is the vector of unobservable differences in discriminative processes (Thurstone, 1927),  $\mathbf{t}$  is a  $n \times 1$  vector of the latent utilities (the position of the stimuli on the trait scale),  $\mathbf{A}$  is a  $\tilde{n} \times n$  design matrix, where the rows of  $\mathbf{A}$  correspond to the paired comparisons, and the columns correspond to the choice alternatives. Finally,  $\mathbf{e}$  is a  $\tilde{n} \times 1$  vector of uncorrelated random errors terms. The design matrix  $\mathbf{A}$  is the fundamental element that defines which paired comparisons are included in a design. A specific example of a design matrix for a full design with  $n = 4$  stimuli is

Keying

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \quad (4.2)$$

It is generally assumed that in the population of all respondents, the latent utilities follow a multivariate normal distribution (Maydeu-Olivares & Böckenholt, 2005; Thurstone, 1927), that is

$$\mathbf{t} \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad (4.3)$$

with  $\boldsymbol{\mu}_t$  is the vector of expected values and  $\boldsymbol{\Sigma}_t$  represents the variance-covariance matrix of the latent utilities. The parameters of Thurstonian models are estimated from the thresholds and tetrachoric correlations (Maydeu-Olivares & Böckenholt, 2005).

#### 4.2.3.1 Thurstonian Factor Model

Let  $m$  be the number of latent traits (or factors), then the  $n$  latent utilities in  $\mathbf{t}$  can be expressed as

$$\mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (4.4)$$

In Equation (4.4),  $\boldsymbol{\Lambda}$  is the  $n \times m$  matrix of factor loadings of the latent utilities on the latent traits,  $\boldsymbol{\eta}$  is a  $m \times 1$  vector of the latent traits, and  $\boldsymbol{\varepsilon}$  is a  $n \times 1$  vector of unique factors (error term). The Thurstonian factor model is given by

$$\mathbf{y}^* = \mathbf{A}(\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}. \quad (4.5)$$

The thresholds of the model are represented by  $\boldsymbol{\gamma}$ . To identify the intercepts (latent utilities)  $\boldsymbol{\gamma} = -\mathbf{A}\boldsymbol{\mu}_t$  is set for Thurstonian factor models. If the latent utilities of the items are not of interest, the thresholds can be freely estimated. Figure 4.2 shows an example for a

Thurstonian factor model for  $n = 4$  stimuli and one factor, that is  $m = 1$ . A factor model as shown in Figure 4.2 would be used when the item utilities are of interest.

#### 4.2.3.2 Thurstonian IRT Model

In cases where it is argued that item utilities are not of interest (e.g., Brown & Maydeu-Olivares, 2011; Frick et al., 2021), a powerful reparameterization of the Thurstonian factor model is available (Maydeu-Olivares & Brown, 2010). This reparameterization is done by

$$\mathbf{y}^* = -\gamma + \mathbf{A}(\mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + \mathbf{e} = -\gamma + \mathbf{A}\mathbf{A}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e} = -\gamma + \tilde{\mathbf{A}}\boldsymbol{\eta} + \tilde{\boldsymbol{\varepsilon}} \quad (4.6)$$

with  $\tilde{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{e}$  and  $\text{cov}(\tilde{\boldsymbol{\varepsilon}}) = \tilde{\boldsymbol{\Psi}}^2 = \mathbf{A}\boldsymbol{\Psi}^2\mathbf{A}' + \boldsymbol{\Omega}^2$ , where  $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{A}$  is a  $\tilde{n} \times m$  matrix. Assuming only one single trait and three items, for example (i.e.,  $m = 1$  and  $n = 3$ ) the reparameterization would be

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_3 \\ \lambda_2 - \lambda_3 \end{pmatrix} \quad (4.7)$$

Assuming  $m = 3$  and  $n = 3$  it is

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \end{pmatrix} \quad (4.8)$$

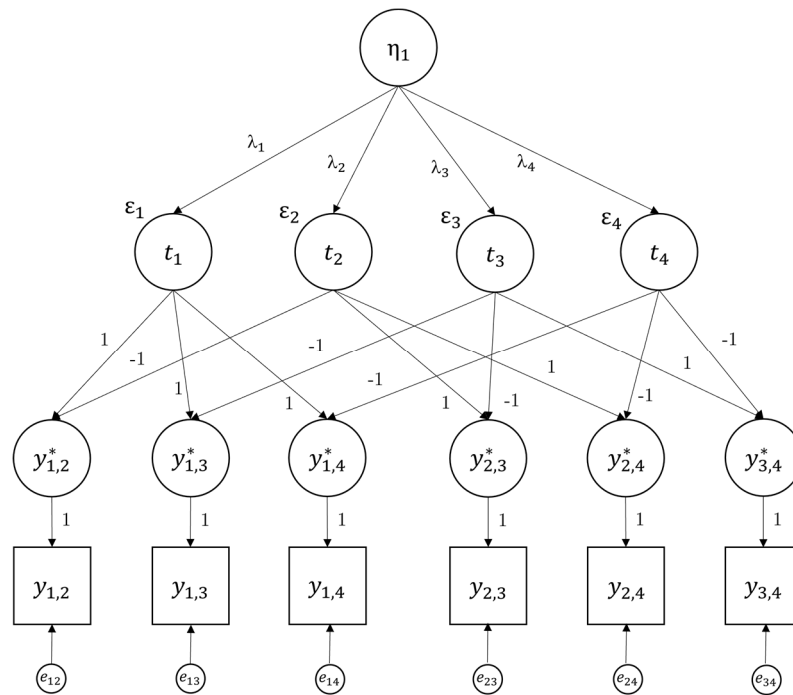
In both cases, the variance-covariance matrix of the unique pairwise errors is

$$\tilde{\boldsymbol{\Psi}}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 + \omega_1^2 & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 + \omega_2^2 & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2 + \psi_3^2 + \omega_3^2 \end{pmatrix}. \quad (4.9)$$

Figure 4.3 shows an example for a Thurstonian IRT model for  $n = 4$  stimuli and one factor, that is  $m = 1$ . Both models in (4.5) and (4.6) are equivalent, as (4.6) is simply a reparameterization of (4.5) and, therefore, both models have the same (but reparametrized) tetrachoric correlation matrix. The advantage of the reparametrized model is that it can be

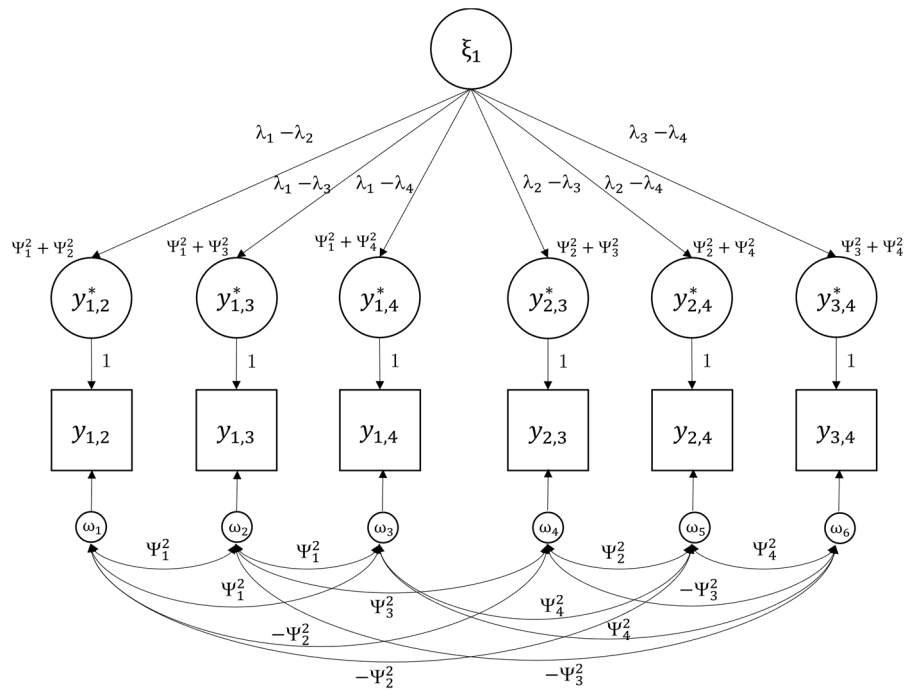
**Figure 4.2**

*Example of a Covariance Structure of a Thurstonian Factor Model for  $n = 4$  and  $m = 1$ .*



**Figure 4.3**

*Example of a Covariance Structure of a Thurstonian IRT Model for  $n = 4$  and  $m = 1$ .*





equivalently regarded as a normal ogive model with some special features. Let  $\Phi(x)$  be a standard normal distribution function at  $x$ ,  $\gamma_i$  the threshold for  $y_i$ ,  $\check{\lambda}_i$  the vector of factor loadings, and  $\check{\psi}_i^2$  the variance of the binary response. Then, for each binary response variable for items  $i$  and  $j$ , the item characteristic function (ICF) is given by

$$\Pr(y_i = 1 | \boldsymbol{\eta}) = \Phi\left(\frac{-\gamma_i + \check{\lambda}_i \boldsymbol{\eta}}{\sqrt{\check{\psi}_i^2}}\right). \quad (4.10)$$

Therefore, it is the ICF of a normal ogive model but with structured  $\check{\lambda}_i$  and  $\check{\psi}_i^2$ . Also, the ICFs are not independent (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). Given  $m = 1$  it is

$$\Pr(y_i = 1 | \eta) = P_i(\eta) = \Phi\left(\frac{-\gamma_i + \check{\lambda}_i \eta}{\sqrt{\check{\psi}_i^2}}\right) = \Phi\left(\frac{-\gamma_i + (\lambda_i - \lambda_j)\eta}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_i^2}}\right) \quad (4.11)$$

and if  $m > 1$ , for each comparison it is

$$\Pr(y_i = 1 | \eta_a, \eta_b) = P_i(\eta_a, \eta_b) = \Phi\left(\frac{-\gamma_i + \lambda_i \eta_a - \lambda_j \eta_b}{\sqrt{\psi_i^2 + \psi_j^2 + \omega_i^2}}\right). \quad (4.12)$$

#### 4.2.3.3 Thurstonian Model Designs

The design matrix  $\mathbf{A}$  defines the paired comparisons that are included in the assessment and analysis. The number of comparisons can be anything from one to  $\tilde{n} = n(n-1)/2$ . Two general designs are of particular interest with Thurstonian models: First, the full design ( $\mathbf{A}$  has  $\tilde{n}$  rows) is important because it yields the maximum empirical information about the relation of the items with respect to the decision criterion. Unfortunately, it is not easy and often not even possible to use a full design in practice. When the number of items exceeds a certain point (e.g.,  $n > 20$ ), gathering data by requiring respondents to make choices for all nonredundant paired comparisons is not feasible because

this would be too burdensome for the respondents and may also take more testing time than available. This is where the second general design type of interest, the block designs, come into play.

In block designs, only blocks of  $k$  items are presented (e.g.,  $k = 3$  with triplets as shown in Figure 4.1). This way of presenting stimuli and gathering responses is normally more efficient and less burdensome for respondents than paired comparisons. This may be the reason why almost exclusively MFC blocks are discussed and analyzed so far (e.g., Frick et al., 2021; Lin & Brown, 2017; Schulte et al., 2021). For example, when using triplets and setting the number of blocks  $p$  to 3 as well (i.e.,  $p = 3$ ; often  $p = n/k$ ), then the number of stimuli is  $n = 9$  and the structured design matrix  $\mathbf{A}$  would be

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad (4.13)$$

In the MFC context, blocks consist of items that are intended to assess different traits.

Nevertheless, all model equations presented as before still hold, the change is just with the design matrix  $\mathbf{A}$ .

Within block design there is another important classification: Linked and unlinked block designs. So far, Thurstonian unlinked block models (TB) are the more common type. In TB,  $\mathbf{A}$  is a block diagonal matrix with corresponding submatrices as shown in (4.13). It was shown that such models can result in biased item and person parameter estimates if the parameters of a full design are the target of estimation (Jansen & Schulze, 2023a, 2023b).

### Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

Estimating the parameters of a full design should always be the goal, as otherwise the number of possible models, given the same set of items, is immense (for a further discussion see Jansen & Schulze, 2023a).

To address the issues associated with TB models, a block design with linked blocks, the TLB, was proposed (Jansen & Schulze, 2023b). To illustrate, the TB example given above is changed into a TLB design by simply adding a fourth block  $B_4 = (1,4,7)$ . This fourth blocks links the other three blocks, so  $\mathbf{A}$  is changed to

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad (4.14)$$

It is instructive to view at the designs discussed here also from a different angle. Consider the rank of the design matrix  $\mathbf{A}$ . The rank of  $\mathbf{A}$  is  $n - 1$  in a full design because the comparative nature of the data results in a location indeterminacy that can be solved with identification constraints. For a TB design, each submatrix for each block has rank  $k - 1$  for the same reason. Hence, with  $p$  blocks the rank of  $\mathbf{A}$  is  $(k - 1)p = n - p < n - 1$ . The TLB design requires the addition of linking blocks until the location indeterminacy can be solved by one identification constraint, resulting in a rank of  $\mathbf{A}$  of  $n - 1$  again. It was shown that if only same-sized blocks of size  $k$  are used, a minimum of

$$\left\lceil \frac{p-1}{k-1} \right\rceil \quad (4.15)$$

additional blocks are needed, where  $\lceil \cdot \rceil$  is the ceiling function and corresponds to rounding up the result (Jansen & Schulze, 2023b).

The identification and estimation methods of the Thurstonian models, especially for latent trait estimation, are not of particular interest here and have been extensively discussed elsewhere. The interested reader is referred to Brown and Maydeu-Olivares (2011; see also Jansen & Schulze, 2023a, 2023b; Maydeu-Olivares & Böckenholt, 2005; Maydeu-Olivares & Brown, 2010) for details on these topics.

#### 4.2.3.4 Information Obtainable by Forced Choice Judgement

One of the main goals in psychological assessment is to accurately estimate person scores. In the forced-choice context this implies that it is essential to obtain as much information with paired comparisons as possible. To quantify the information that a specific paired comparison holds, the Fisher information matrix is useful. The Fisher information matrix is given by (see also Brown & Maydeu-Olivares, 2018; Appendix B)

$$I_l(\eta_a, \eta_b) = \frac{1}{\psi_i^2 + \psi_j^2 + \omega_l^2} \begin{pmatrix} \lambda_i^2 & -\lambda_i \lambda_j & \cdots & 0 \\ -\lambda_i \lambda_j & \lambda_j^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \frac{P_l(\eta_a, \eta_b)^2}{P_l(\eta_a, \eta_b)(1 - P_l(\eta_a, \eta_b))}. \quad (4.16)$$

As can be seen, the ICF as given by Equations (4.11) and (4.12) is part of the Fisher information. The ICF depends on the difference in loadings between items from a paired comparison and, therefore, contributes directly to the information of the corresponding paired comparisons (Brown & Maydeu-Olivares, 2018). In the extreme case of equal loadings (for unidimensional designs), the information would not depend on a person's trait scores at all. That is, no information for the estimation of that person's traits would be obtained in the case of equal loadings. In contrast, large differences in item loadings do contribute information for estimation. For multidimensional designs, a small loading difference offers a small amount of information only if also the trait difference is small. This is a main reason for the

attractiveness of the MFC format in comparison to one-dimensional blocks with a single underlying trait. The effect of the difference in loadings just described also explains the general recommendation to construct FC blocks of mixed-keyed items, that is, blocks that contain at least one positively and one negatively keyed item (Brown & Maydeu-Olivares, 2011). Mixed-keyed blocks can be regarded as extreme cases with respect to loading differences because they guarantee large differences in items loadings. The corresponding paired comparisons in mixed-keyed blocks should therefore yield much information on the latent trait scores. As discussed by Brown and Maydeu-Olivares (2011) and others (see also Bürkner 2022; Schulte et al., 2021), this effect is needed in order to improve convergence rates and trait score estimation as well as to reduce biases in parameter estimation.

An analogous discussion of the information on FC measurement was presented by Bürkner (2022). In this work, the question of what proportion of paired comparisons should be mixed-keyed was addressed. To provide an answer, the perspective was put on the design of experiments with a focus on optimal item design. Given standard criteria of optimality (A- and/or D-optimality) for an optimal design, the result was that half of the item pairs should contain mixed-keyed items and half of the pairs homogenously keyed items. In addition, each trait should be measured by the same number of paired comparisons (Bürkner, 2022).

Moreover, it was shown that for the mean of the trait scores, high factor loading differences yield the most information, while for trait score differences, the factor loading sums provide the most information (Bürkner, 2022). This leads to the need for both a) pairs with large item loading differences and b) pairs with small item loading differences, while at the same time, the standardized loadings themselves  $(1 - \lambda_i)$  should be large.

Focusing on the information given by the Fisher matrix and the trait scores is one way to shed light on the question of what makes a good FC design for applications. For a different approach assume there is indeed one true latent order of all stimuli for each respondent and

the order is transitive. In such cases, the most information about an item's position in the ranking should be obtainable from a full ranking of all items by all respondents. The more items an FC block contains, the more nonredundant paired comparisons are implied (one for pairs, three for triplets, six for quartets, ten for quintets, and so on). The most paired comparisons can be obtained by a single block of  $n$ . Hence, the single block design is highly attractive by this reasoning. However, there are three main reasons to use block designs instead of a full design. First, the number of paired comparisons grows quickly with every additional item, potentially resulting in prohibitively long assessment time and reduced motivation as well as effort for each respondent. Second, ranking a large number of items simultaneously is not only time-consuming but can easily exceed an individual's cognitive capacity (see Sass et al., 2020). Third, estimating the corresponding models requires significant computer resources and time. For instance, *Mplus* (Muthén & Muthén, 2022) currently only allows for about 300 binary indicators. This is already the number of indicators needed in a design with 25 items, if standard errors and the model fit have to be computed. Additionally, IRT model estimation in a case with 25 or more items may take weeks or months, with current computers and algorithms. For these reasons it appears to be advisable to use a more parsimonious block designs instead of a full design.

Given that it is rarely feasible to use the full design, it is useful to have another viewpoint on the information that a paired comparison provides that involves items  $i$  and  $j$ , both having a certain position within a ranking. To take this different viewpoint, the probability of choosing item  $i$  over  $j$  is taken into account by considering the utility distributions of both items. Under Thurstone's model, the difference in item utilities determines this probability. If the difference in utilities is large, then the probability of choosing item  $i$  over item  $j$  is very high when the difference favors item  $i$ , or very low vice versa. This is the case in both scenarios since the overlap between discriminative processes is

small. However, a comparison with very large utility differences provides only limited information on the ranking itself. Conversely, if the probability of choosing item  $k$  over  $j$  is near chance, then this comparison is very informative concerning the item order.

The information that a paired comparison provides on the items' positions (order of items) will be illustrated by the following example. Assume  $n = 4$  items, and assume the item order would simply be 1, 2, 3, 4. Comparing the item with the largest utility (item 1) to the item with the lowest utility (item 4), would only yield the information that item 1 has a higher utility than item 4. However, assuming transitivity this applies to all items, as item 1 has a higher utility and item 4 has a lower utility than every other item, therefore the information that this comparison has on the order of items is small. To extend the example, imagine that an item 5 is added and its position in the ranking should be determined. Further assume that item 5 has indeed a lower utility than item 4. When beginning to compare item 5 to item 1 (the largest difference in utilities), then to item 2 (second largest difference in utilities) and so forth, a total of four comparisons would be needed, to determine its position. This is the case since the large difference in item utilities in the first comparison, for example does not provide much information on the position of item 5. Directly comparing item 5 to item 4 (the smallest difference in utilities) instead would reveal the true ranking position of item 5 with only one comparison of maximal information. This is also true for any other true order as well. If a new items position would be in between two items in a ranking, then comparisons with these two adjacent items have minimal utility differences and therefore maximal information for the ranking order (i.e., the new item's position in the ranking). *Ceteris paribus*, for any comparison the information on the ranking is larger the smaller the utility differences of these items are. Note that choices involving items with small utility differences are also harder for a respondent to make. From a dominance response perspective, it is

assumed that item utilities are driven by trait utilities, which would result in a correlation between item utilities and the respective item loading on a trait.

In summary, the literature on Thurstonian FC modeling suggests that balancing same- and mixed-keyed item pairs within the FC assessment is necessary to accurately estimate item and person parameters. While this is true mathematically and statistically, it may not be true from a psychological perspective or in reducing response biases and faking with FC assessments.

#### ***4.2.3.5 Discussion on the Key Direction of Items***

The main reason why FC assessments have such an attraction for many users is the promise associated with this approach that a number of response biases and faking tendencies can be eliminated or reduced. Indeed, some response biases are eliminated by design (e.g., the acquiescence bias) but social desirability and plain faking are not as easily controlled for (see also Schulte et al., 2021). To substantially reduce the likelihood of faking, items in blocks have to be matched based on their social desirability. In order to successfully compose blocks that are homogenous with respect to the desirability of its items at least some reliable and valid information on the items' desirability properties is necessary. Such item desirability data is often lacking or based on questionable sources in practical applications. It appears that the possibility of estimating item desirability values with Thurstonian factor models has been overlooked in many studies. This is somewhat ironic because Thurstonian models were originally designed for item scaling (see Jansen & Schulze, 2023c; Maydeu-Olivares & Böckenholt, 2005). It is worth noting, though, that social desirability can vary by context (Ziegler et al., 2009), so matched blocks may differ depending on context. Balancing small differences in desirability with large differences in item loadings by including mixed-keyed blocks can result in a paradoxical situation, especially in high-stakes contexts when the motivation to fake responses is at its peak. For example, in a job application scenario, a



person may be presented with a FC assessment. A simple but extreme paired comparison would be the one between the items “I get my tasks done.” and “I fail even at easy tasks.” (with instruction as given in Figure 4.1). In this case the choice would be easy: the person would most likely choose the former statement because it is more desirable in this context even if choosing the latter statement was more truthful. While this is specifically true for high-stakes situations, the social desirability bias is not restricted to these cases. In theory, the large difference in loadings (and desirability) can result in a difference in response that provides much information on the trait means, according to the ICF and Fisher information. However, this information is likely to reflect a socially desirable or fake answer instead of a genuine dominance response on a scale rendering it essentially meaningless (as also noted by Bürkner, 2022). If this argument is valid for paired comparisons, then it also holds for blocks of size  $k$ , even worse so, as for blocks of three statements the ranking of a triplet results in three paired comparisons, two of which would be highly fakeable. Naturally, the desirability difference within blocks may not be so extreme in general. While blocks with statements of similar desirability but still mixed-keyed may exist (Wetzel & Frick, 2020), we argue against using this option except for some special cases because the difference in loadings is defined arbitrarily. A good example to illustrate is the well-known Big Five model of personality. For neuroticism the dominance responses are reversed, that is, if the loadings on the other four factors are positive, the loadings on neuroticism and correlations of the other facets with neuroticism are generally negative. Taking, for example, the Big Five Triplets (BFT; Wetzel & Frick, 2020), the first triplet matched by desirability is

1. I get stressed out easily. (N+)
2. I have little to say. (E-)
3. I distrust people. (A-)

### Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

---

Technically, both conditions for FC assessment are met: first, items within the block are similar in their desirability values (Wetzel & Frick, 2020) and second, the block contains mixed-keyed items. Hence, the information provided by two of the three paired comparisons yields high information based on the Fisher information. However, if we reverse the interpretation of neuroticism to emotional stability, then the item “I get stressed out easily.” should load negatively on emotional stability. The triplet would not change and the desirability values would remain the same and still be matched. However, with the reversed interpretation the block would contain only negatively-keyed items. The information from none of the paired comparisons would be high as a result. How is it possible that the unchanged triplet, presented to the same people yields much information in one interpretation but less so in the other interpretation of the factor? Well, it is not. To solve this seeming paradox, it is a necessary condition to exclude the influence of the interpretation direction of factors. The direction of interpretation must be the same for every factor. To achieve high differences in item loadings, the extreme case would be to have one item loading high on a trait and another item loading zero on all traits considered. Taking all this into account, it will be difficult, if not impossible, to construct desirability-matched blocks with mixed-keyed items (see also Bürkner et al., 2019; Jansen & Schulze, 2023a).

#### **4.2.3.6 Summary**

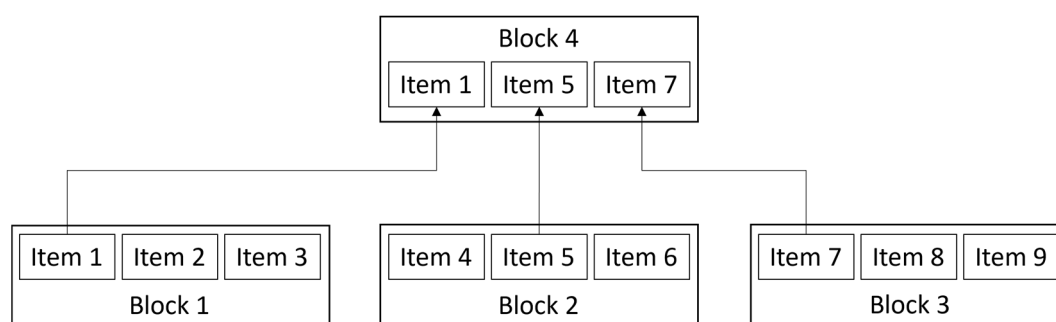
In summary, there are three main sources of information in paired comparisons: first, differences in item loadings on specific traits contributing to the trait scores, second, the optimal item design, and third, the position of each item within the full ranking. The influence of these sources implies that one should include desirability-matched blocks with similarly large item loadings to reduce faking and improve information on trait differences. A FC design should also include blocks with large differences in item loadings, including mixed-keyed blocks in extreme cases. Additionally, the number of same- and mixed-keyed

paired comparisons should be equal for optimal item design, if possible. However, including mixed-keyed blocks is problematic due to arbitrary interpretation of the direction of trait content, so emphasis should be placed on the first two recommendations.

Luckily, the Thurstonian linked block design (Jansen & Schulze, 2023b), makes it possible to construct FC tests that include both same- and mixed-keyed blocks. The strategy would be to start with unlinked item blocks that are all matched by social desirability to reduce fakeability and yield information about the ranking position and trait score differences. These initial blocks should contain same-keyed items, though it is reasonable to create blocks with homogeneously positively and negatively keyed items. Subsequently all blocks are linked by assembling items from the initial blocks into additional linking blocks. Linking naturally creates blocks with large item loading differences or even mixed-keyed blocks, as all blocks have to be interconnected. While the initial blocks may be harder to respond to, the mixed-keyed blocks should be easier and provide more information about the trait means, leading to less biased estimation of item, trait, and person parameters. Reducing the number of mixed-keyed blocks also reduces the potential for fake responses. Figure 4.4 shows an example of the linking procedure. Overall, the TLB design results in many blocks with same-keyed items and a few with mixed-keyed items.

#### Figure 4.4

*Example for the Linking Procedure of a Thurstonian Linked Block Design.*



#### 4.2.4 Simulation Studies

The goal of the simulation studies reported in this section is to compare TB and TLB designs by evaluating the bias of item parameters and the recovery of respondents' trait scores. Previous simulation study results on the TLB design will be extended by also including mixed-keyed blocks. The impact of sample size on the results will also be examined which will be particularly relevant for practical applications of the models and designs.

All simulation studies were done with R (R Core Team, 2022), the R package *ThurMod* (Jansen, 2023) and *Mplus* (Muthén & Muthén, 2022). For the simulation studies, the following conditions were used:

1. Model design (unlinked blocks vs. linked blocks)
2. Model type (factor vs. IRT model)
3. Number of traits  $m$  (one, three, five)
4. Number of items  $n$  (six, twelve, eighteen items per trait)
5. Correlations of the factors
  - a.  $m = 3$ : uncorrelated, (.20, .30, .50)
  - b.  $m = 5$ : uncorrelated, (-.21, .02, -.25, -.53, .40, .04, .27, -.02, -.02, .24)
6. Sample size (200, 500, 1000)
7. Number of negatively keyed items per trait (1 vs. 2)

If only one trait is present, the number of items was either twelve or eighteen. With the addition of three different block types, a total of 1008 conditions were simulated. Not all results are presented here to preserve clarity and to avoid repetition and redundancies. The full simulation results can be found in the OSF repository.

The TLB design is expected to show better parameter estimation due to the larger number of blocks. That is, with more blocks and corresponding paired comparisons, the

amount of information of these designs is larger by definition. However, previous simulation studies (Jansen & Schulze, 2023b) show that simply adding a number of blocks is not sufficient to improve parameter estimation.

Jansen and Schulze (2023a, 2023b) discussed that each design (given by design matrix  $\mathbf{A}$ ) in Thurstonian modeling leads to different results for the simulation of a sample. It is important that simulation results are not influenced by the design (the operationalization). To avoid design influences, it is important to simulate one dataset per iteration and estimate all models on the same data in each iteration. Consequently, all datasets were simulated with a full design making responses on all possible comparisons of all items available to draw from. Furthermore, to reduce complexity, data was generated from a Thurstonian factor model with restricted thresholds, assuming a ranking task ( $e = 0$ ). For each condition, 1000 repetitions were performed using triplets only. Results are shown and discussed for

1. Convergence rates
2. Utility estimates and correlations between estimated and true utilities
3. Loading estimates and correlations between estimated and true loadings
4. Correlation estimates between factors
5. Recovery of person parameters
6. Goodness of Fit

#### ***4.2.4.1 Simulation Study 1: A Forced-Choice Assessment Measuring Three Uncorrelated Traits***

Three uncorrelated traits were used in this first simulation study. In addition to the conditions listed above block compositions were also varied as follows:

1. Items are ordered by their loadings.
2. All blocks are purely multidimensional.
3. Initial blocks are purely unidimensional.

### Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

---

For comparability, blocks were created accordingly and then some item loadings were recoded to be negative to ensure that all block compositions had some negatively keyed blocks. Otherwise, if there is one negatively keyed item per trait, then block composition (1) would have no mixed-keyed blocks (the three lowest loadings would be in one triplet) and block composition (3) would have three mixed-keyed blocks (as all initial blocks are unidimensional). In all block compositions, linking was done such that the least number of mixed-keyed blocks was created. Block composition (1) should be less reliable than (2), as the influence of the trait is small by definition (see Equations (4.11) and (4.12)). The true factor loadings were drawn from a uniform distribution between .30 and .90 and utility means were drawn from a uniform distribution between -1 and 1. True uniquenesses were specified as  $\psi_i^2 = 1 - \lambda_i^2$ , similar to previous studies (Brown & Maydeu-Olivares, 2011; Jansen & Schulze, 2023b; Schulte et al., 2021). For conditions with twelve items, six items were equal to the six-item condition and six items were new. For conditions with 18 items, twelve items were equal to the 12-item condition and six items were newly added. For the TB designs, one factor loading and one uniqueness was fixed to unity for each block to estimate the IRT models. For the TLB design only one factor loading and one uniqueness was fixed. Additionally, to estimate the factor models one utility mean was fixed to zero, either per block or one for the whole design. If all initial blocks are unidimensional, then linking blocks necessarily include multidimensional blocks.

For each of the separate unlinked blocks of the TB design, parameter estimations (loadings/ latent utility means) were regressed on the true parameters and then rescaled so that the scale and origin were equal between estimations and true parameters with respect to the identification constraint. For the TLB design, parameters were rescaled by one regression parameter. However, this procedure conceals that the scale is redefined for each block for the TB design and is therefore incomparable. For the current simulation study, results from

rescaling the estimation for each block (bias) are presented, but correlations between estimates and the true parameters (recovery) are also reported. All results are compactly presented in Figures 4.5 and 4.6. For details, we refer again to the OSF repository. Overall, it can be easily gleaned from the figures that better results are observed as the number of items and the sample size increase.

#### ***4.2.4.1.1 Convergence Rates***

Rates of valid iterations (convergence; first two rows of Figures 4.5 and 4.6) are higher for IRT models than for factor models. For unlinked TB designs the convergence rates are only acceptable when 12 or 18 items per trait are used with multidimensional blocks and a sample size of 1000. For TLB designs, convergence rates are much better. They appear to be acceptable for sample sizes of at least 500 and with 12 or 18 items per trait. This applies to all block compositions. The exception is the TLB – factor model with block (3) and 12 items per trait (Figure 4.6 top-right cell) for which convergence rates are low even for a sample size of 1000. This result is likely an artifact of the specific simulated design and not a general trend, as we did not find this result anywhere else. Convergence rates are worst for unidimensional block designs.

#### ***4.2.4.1.2 Item Parameter Recovery***

All models yield precise estimates for the latent utilities if they are rescaled within each block, except for multidimensional blocks and a sample size of 1000 in a TB design. However, the need to rescale the utilities per block produces the effect that the scales between blocks are non-comparable. Somewhat naturally, the recovery is unacceptable for TB designs. In contrast, for TLB designs the recovery of item utilities is good if less than 1000 and excellent if at least 1000 respondents are included. Interestingly, for TB designs with blocks (1) the recovery is best for small item sets and gets worse for larger item sets. This can be explained by the number of fixed parameters. The larger the total number of items, given a

### Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

---

fixed number of items per block (triplets), the more constraints are needed. This reduces the correlation between true and estimated parameter values. The standard errors of the utility estimates are accurately estimated for all models except for TB designs with unidimensional blocks. As convergence rates are low in these cases, the interpretation of the bias is limited. The same principle is true for the correlations of the factors. While all models and designs can accurately recover correlation estimates, TLB designs are most accurate, especially with sample sizes of at least 500. Standard deviations between correlation estimates are smaller for TLB designs, resulting in more precise estimations. Interestingly, the unidimensional block compositions (3) yield considerably less biased results even for small sample sizes.

The bias of factor loadings estimates is acceptable for all models with a sample size of at least 500 respondents. However, similar to item utility estimates, it should be noted that the scale of loading estimated is redefined for every block with TB designs. The recovery of factor loadings is unacceptable whenever a TB design is used for estimation. In contrast, all TLB designs can accurately recover factor loading estimates with sample sizes of at least 500. For the standard errors of factor loadings, bias is acceptable if at least 500 respondents and 12 items per trait are used, but bias is generally lower for TLB designs. Lastly, thresholds are accurately estimated by all models and designs with a sample size of at least 500 respondents.

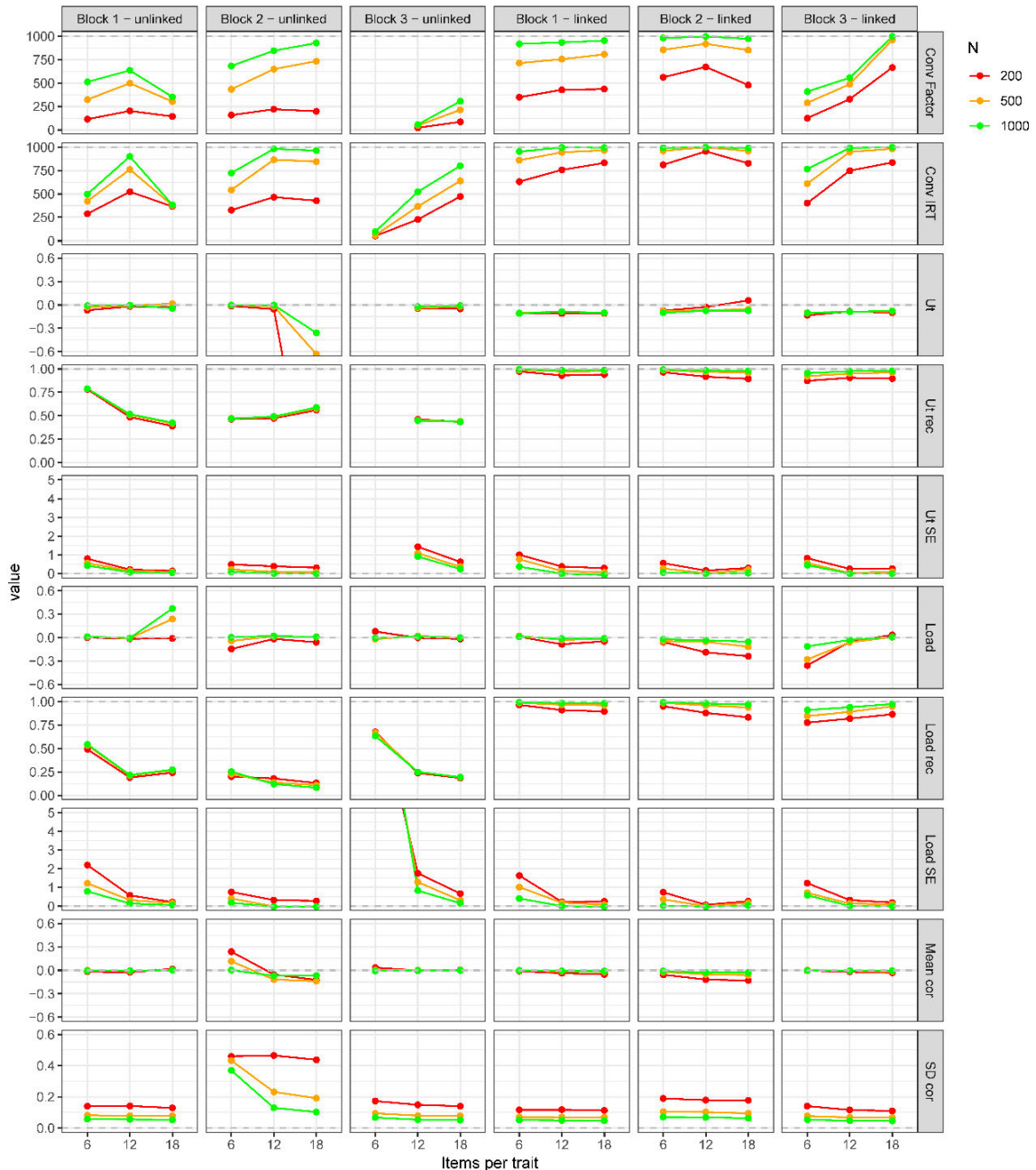
Comparing the results based on the number of negatively keyed items per trait, one may expect better results for conditions with more than one negatively keyed item per trait, but this is not the case in general. Slight reductions in bias and increased recovery rates are observed, but these tend not to be substantial for both TB and TLB designs.



Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

**Figure 4.5**

*Results for Convergence Rates, Bias and the Correlation Between True and Estimated Parameters for Three Uncorrelated Traits. One Item per Factor (Three Items Total) is Negatively Keyed.*

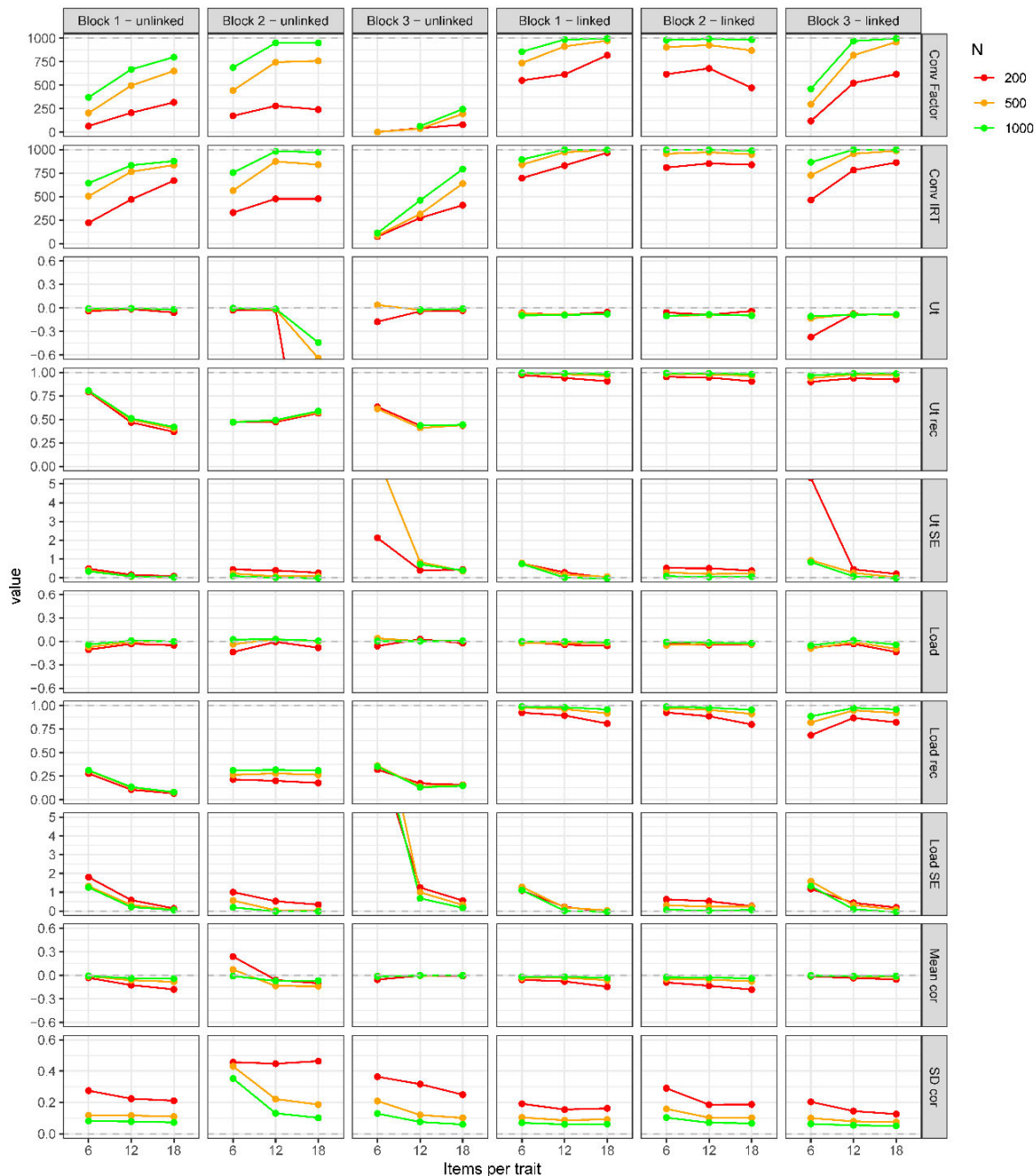


*Note.* Missing points correspond to missing values due to non-convergence or elimination of all iterations. Abbreviations: Conv: convergence rate; Ut: Utilities; Cor: Correlation; Load: Loadings.

## Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

**Figure 4.6**

*Results for Convergence Rates, Bias and the Correlation Between True and Estimated Parameters for Three Uncorrelated Traits. Two Items per Trait (Six Items Total) are Negatively Keyed.*



*Note.* Missing points correspond to missing values due to non-convergence or elimination of all iterations. Abbreviations: Conv: convergence rate; Ut: Utilities; Cor: Correlation; Load: Loadings.

#### 4.2.4.1.3 Latent Trait Score Recovery

For latent trait score recovery, real recovery is compared to the empirical recovery. Real recovery is given by the correlation between true scores (used for simulation) and estimated maximum a posteriori (MAP) scores using *Mplus*. Empirical recovery is determined as the square root of the empirical reliability given by

$$\rho = \sqrt{\delta} = \sqrt{\frac{\sigma^2 - \bar{\sigma}_{\text{error}}^2}{\sigma^2}}, \quad (4.17)$$

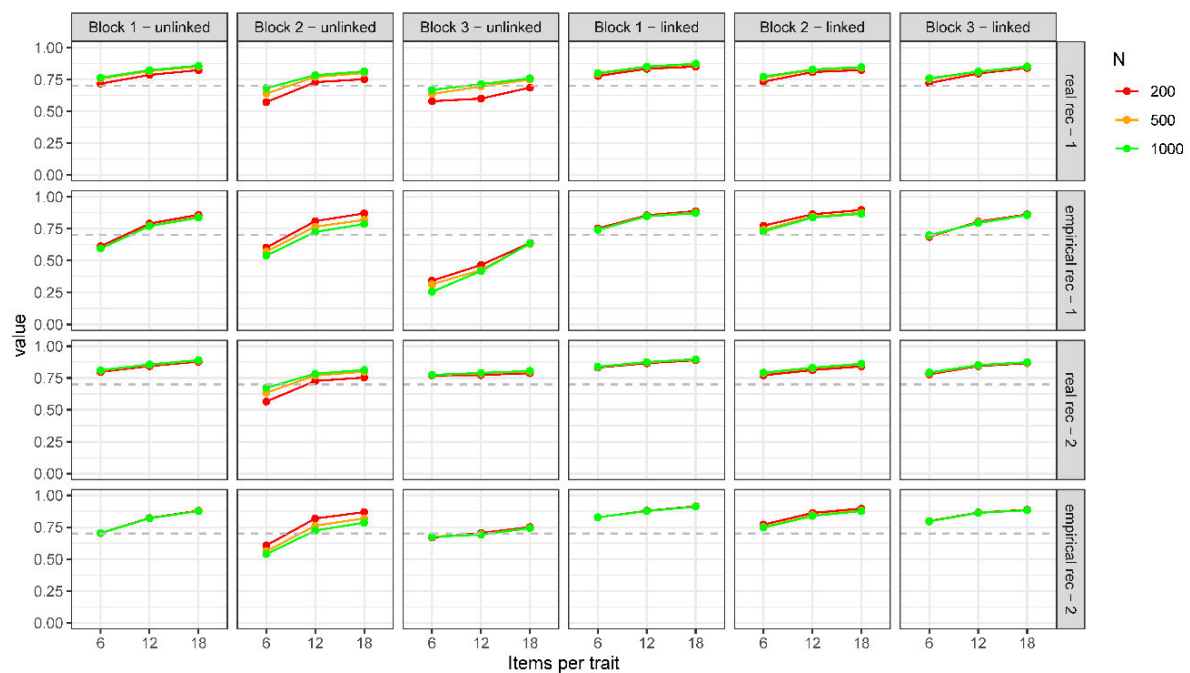
where  $\sigma^2$  is the observed score variance given by the variance of the estimated MAP scores and  $\bar{\sigma}_{\text{error}}^2$  is the average of the squared standard errors of the MAP scores (Brown & Maydeu-Olivares, 2011). Results for recovery are presented in Figure 4.7. Generally, recovery improves with more items per trait. The TLB designs have better recovery (above .70) as compared to TB designs, especially for blocks (2) and (3). For TB designs, more negatively keyed items per trait result in better recovery, but the benefit is rather small for TLB designs. Interestingly, recovery is larger for blocks (1) in comparison to (2) and (3), contrary to what equation (4.12) and previous simulation studies suggest. Real recovery increases with larger sample size even if the benefit is small in TLB designs. Surprisingly, the contrary is true for the empirical recoveries: the larger the sample size, the smaller the recoveries are. Studying the MAP results of *Mplus* shows that SEs of MAP scores are on average smaller for smaller sample sizes and larger for larger sample sizes. The smaller the SEs of the MAP scores, the smaller the error variance is and, hence, the larger the recoveries are.

The real trait scores are unknown in real data, so it is of interest to compare real and empirical recoveries. Real and empirical recoveries are similar for TLB designs, but empirical recoveries underestimate real recoveries for TB designs, especially when using only few items.

## Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

**Figure 4.7**

*Results for the Recovery of the Three Uncorrelated Traits Study (Simulation Study 1).*



Note. Abbreviations: rec: Recovery. The first two rows come from simulation with one (- 1) the second two rows from simulation with two (- 2) negatively keyed items per trait.

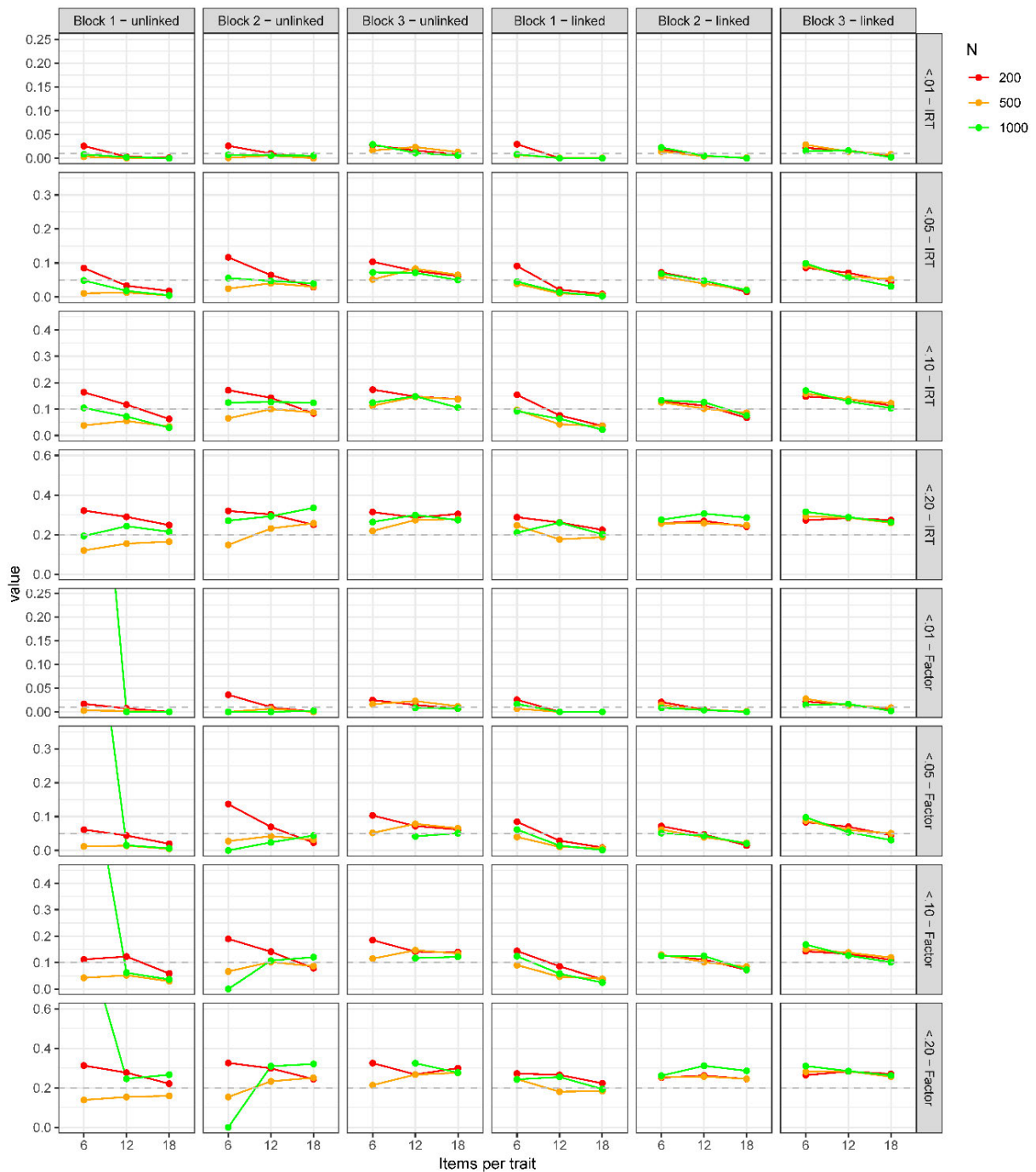
### 4.2.4.1.4 Goodness-of-Fit Tests

For all model results the degrees of freedom were corrected by the number of redundancies among the tetrachoric correlation matrix and thresholds. The empirical rejection rates are shown in Figure 4.8. The rejection rates do not vary fundamentally between item sets with one or two negatively keyed items per trait and are therefore averaged within each condition. For IRT and factor models as well as TB and TLB designs, the empirical rejection rates are higher than expected, but the discrepancy is smaller for TLB designs than for TB designs. With sample sizes of 200 or 500 respondents, the rates seldomly exceed the limits, but with a sample size of 1000 respondents and with small item sets, the target values are clearly exceeded even for TLB designs. Compared to the rejection rates reported by Brown

Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

**Figure 4.8**

*Empirical Rejection Rates for the Three Uncorrelated Traits Study (Simulation Study 1).*



*Note.* Degrees of freedom are adjusted by the number of redundancies per design. Missing points correspond to missing values due to non-convergence or elimination of all iterations.

and Maydeu-Olivares (2011), the rejection rates of the current study and for TLB designs are only slightly higher and, overall, mostly within an acceptable range.

#### ***4.2.4.1.5 Discussion***

The first simulation study shows that including differently keyed items (i.e., negatively keyed in this study) improves estimates of item and trait parameters as compared to prior simulation studies of TB and TLB designs (Jansen & Schulze, 2023b). The number of respondents included in the analyses affects the precision of estimation and convergence rates. As can naturally be expected, larger sample sizes result in lower bias and larger convergence rates, recoveries, and better empirical rejection rates.

The recovery of item parameters (loadings and utilities) in TB designs is low and unacceptable due to the numerous identification constraints per block. However, the TLB design with few identification constraints allows for estimation of both item and person parameters by including information between blocks. This leads to larger convergence rates, lower bias in estimates, and standard errors as well as especially high recovery of item parameters. The trait parameter recoveries for TLB designs are larger and the difference between real and empirical recoveries is smaller. Also, the empirical rejection rates for TLB designs are mostly acceptable, but not for TB designs. With regards to absolute performance, models with only 200 respondents produce largely unacceptable results, while those with sample sizes of at least 500 respondents are arguably acceptable. Therefore, such small sample sizes are not recommended. Finally, adding a second negatively keyed item per trait did not substantially improve item or trait parameter results.

---

#### ***4.2.4.2 Simulation Study 2: A Forced-Choice Assessment Measuring Three Correlated***

##### ***Traits***

In the second simulation study three correlated traits were considered. All other parameters and conditions were the same as in simulation study 1. The correlations between traits were  $r_{12} = .20$ ,  $r_{13} = .30$ , and  $r_{23} = .50$ , respectively.

##### ***4.2.4.2.1 Item Parameter Recovery***

The results are very similar to those of simulation study 1, and are not repeated here in detail (for detailed results see OSF). The convergence rates and bias estimated indicate better model performance for correlated traits as compared to uncorrelated traits. Of note, TLB designs show convergence rates for similar loading blocks (1) and unidimensional blocks (3) that are considerably higher for models with correlated traits in comparison to models with uncorrelated traits. However, this does not coincide with the less biased parameter estimated as was already the case in previous simulation studies (Jansen & Schulze, 2023b). As the data was simulated with correlated traits it is interesting to focus on the results for the trait correlations (Figure 4.9).

In general, bias and standard errors for the latent trait correlations are accurate only for the TLB designs and with a sample size of 1000. The inclusion of a second negatively keyed item per factor reduces the overall bias for both TB and TLB designs, but the bias is largely unacceptable for TB designs.

##### ***4.2.4.2.2 Latent Trait Score Recovery and Goodness-of-Fit***

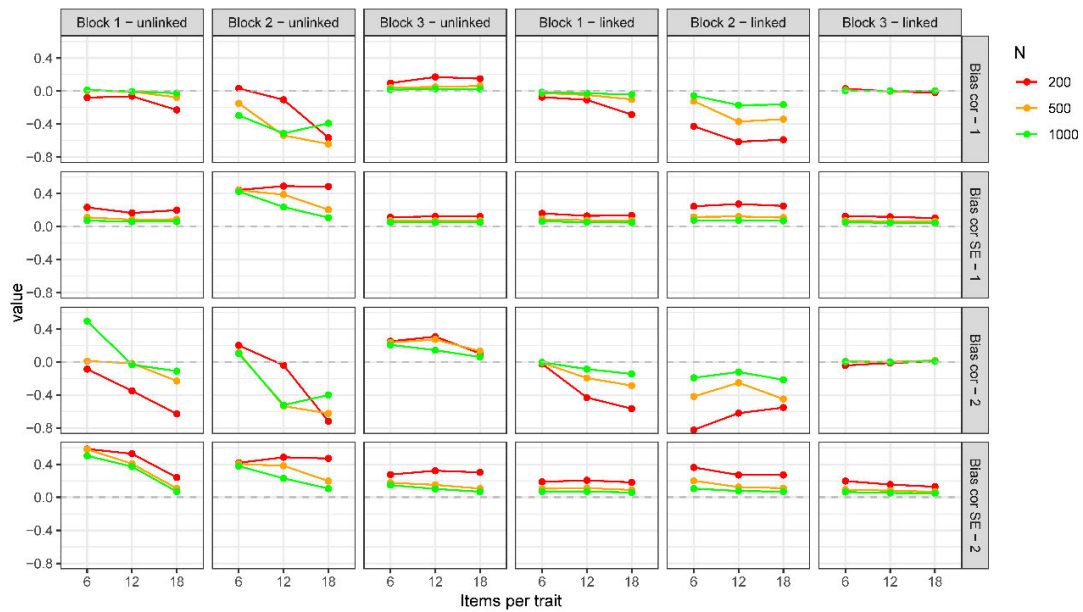
The results for the latent trait score recovery for TB and TLB designs are presented in Figure 4.10. Again, the results are similar to those in simulation study 1. The recoveries are slightly smaller for a correlated model with all positively correlated traits, as was also shown by Brown and Maydeu-Olivares (2011). However, the difference is more noticeable for TB designs, especially when using relatively few items.



Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

**Figure 4.9**

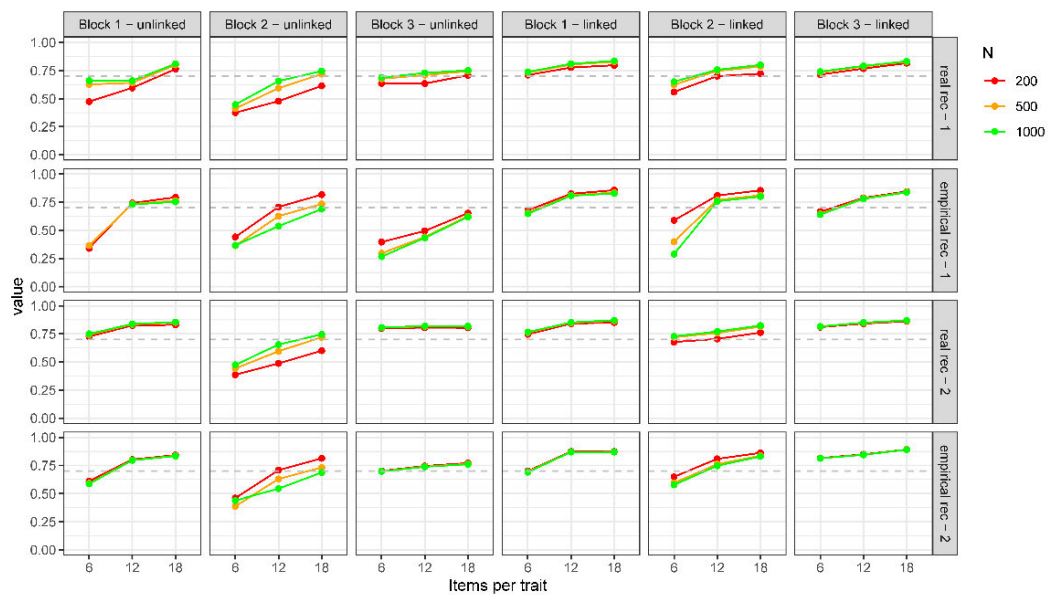
*Bias and SEs of Latent Correlations for Three Correlated Traits (Simulation Study 2)*



*Note.* Abbreviations: Cor: Correlation. The first two rows show results of simulations with one (- 1), the second two rows of simulation with two (- 2) negatively keyed items per trait.

**Figure 4.10**

*Results for the Recovery of Three Correlated Traits (Simulation Study 2).*

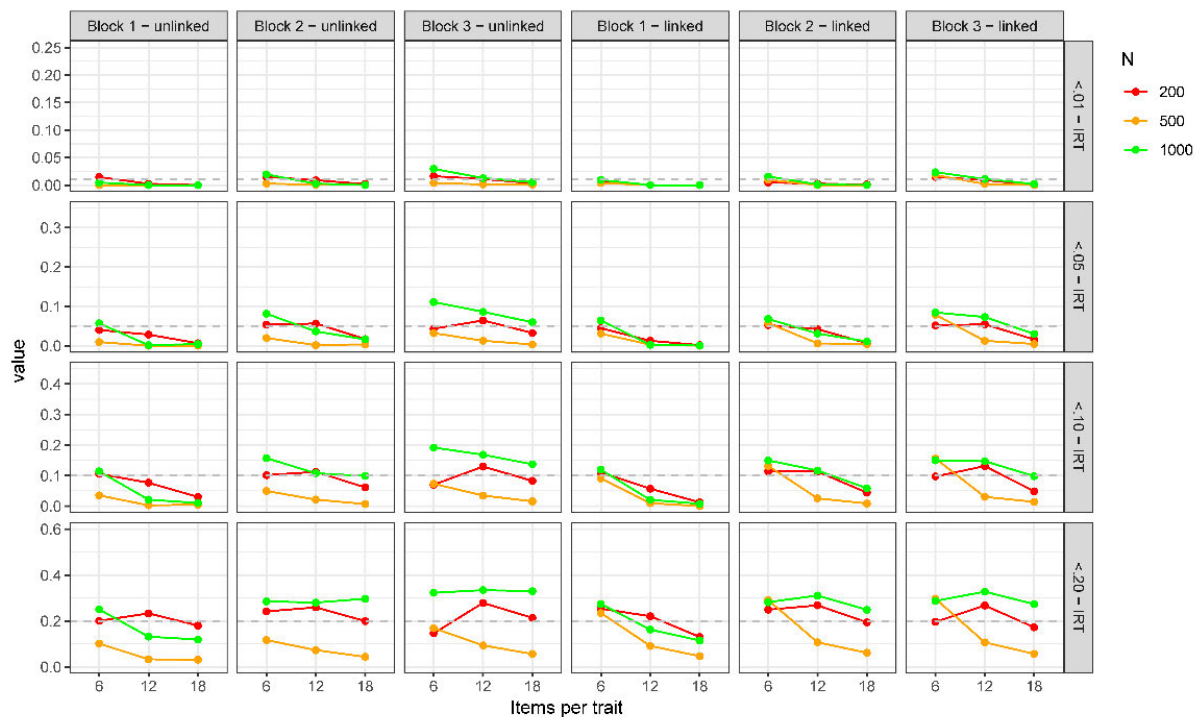


*Note.* Abbreviations: rec: Recovery. The first two rows come from simulation with one (- 1) the second two rows from simulation with two (- 2) negatively keyed items per trait.



**Figure 4.11**

*Empirical Rejection Rates of the IRT Models for the Three Correlated Traits Study.*



*Note.* Degrees of freedom are adjusted by the number of redundancies per design.

Considering the goodness of fit, models with correlated traits, again, yield slightly better results for all models and blocks. Especially for models with 12 or 18 items per trait, the empirical rejection rates are considerably smaller than expected for multidimensional blocks (see Figure 4.11 for IRT models).

#### 4.2.4.2.3 Discussion

The results of the second simulation study are mostly not substantially different from the results of the first simulation. For models with correlated traits, the relative bias shows a tendency to be smaller than for models with uncorrelated traits. Also, convergence rates are higher and recovery rates are slightly smaller, as all factors were positively correlated. However, these differences are not substantial for the TLB designs. More importantly,

recovery rates are better than when only same-keyed blocks are used (as was the case in Jansen & Schulze, 2023b), especially for latent trait scores.

Considering the simulation studies with one trait and five (uncorrelated/ correlated) traits, the results evince a similar pattern (again, see OSF). If only one trait is included, convergence rates are acceptable only for the TLB design and for the IRT model only when blocks are constructed with random items and at least 18 items are used. Note that it is not possible to construct multidimensional blocks here, which is why block composition (2) is treated as random. In all TLB designs, mean relative biases and recoveries are acceptable but for TB designs recoveries are not. With regards to the latent trait score recoveries, all TLB designs yield acceptable values if sample sizes of at least 1000 are used. The empirical rejection rates are exceeded in general.

For five traits, there are more items in comparison to the conditions with three traits. This higher number of traits and items enhances the quality of the overall results. Considering the recoveries of the item parameters, TB designs yield unacceptable results yet again, while TLB designs recover true values excellently. Once more, this follows the general pattern shown in conditions with three traits and supports the use of the TLB designs, ideally with multidimensional blocks.

#### **4.2.5 Overall Discussion**

In the current study, different designs for Thurstonian models – the classic TB design and the more recently developed TLB design – were presented and compared from both a methodological viewpoint and an empirical stance with comprehensive simulation studies. In addition, the analyses of the impact of different designs on the interpretation and quality of the results included the issue of the benefits of mixed-keyed blocks.

With respect to item keying, it is argued that the key direction of traits is somewhat arbitrary since the interpretation of an item direction is a matter of item content and

psychological theory. Unlike previous assessment strategies (Brown & Maydeu-Olivares, 2011, Wetzel & Frick, 2020), reversed scoring or interpretation of a factor (e.g., neuroticism and emotional stability) does not contribute to the information that can be gathered from paired comparisons for the estimation of latent trait scores. Three major sources of information that contribute to the quality of the latent trait scores estimation are discussed: the difference in loadings, the optimal item design perspective, and the relative position of each item in a ranking. To effectively use these various sources of information and control for fakeability in FC assessments, the use of the recently developed Thurstonian linked block design is recommended (Jansen & Schulze, 2023b).

The results of the simulation studies provided support for this recommendation and further shed light on the keying issue as well as some more practical aspects of designing a FC assessment suitable for modern Thurstonian modeling. The simulation study results clearly showed that – under the conditions instantiated – TB designs are not appropriate for item parameter estimation (loadings and utilities) because the recovery of item parameters  $r(\theta, \hat{\theta})$  is relatively poor and often only at about .50. This is due to the many identification constraints that need to be applied to each block with the TB design. In contrast, TLB designs have a rank of  $n - 1$  for their design matrix  $\mathbf{A}$  and therefore only the constraints for one block need to be applied. With a sample size of 1000 respondents, item parameter recoveries for the TLB design are considered to be of sufficient quality and appropriate for their use with  $r(\theta, \hat{\theta}) \geq .85$ . Latent trait recovery is generally better and more robust with TLB designs, but the relative benefit is reduced if more mixed-keyed blocks are used. Lastly, empirical rejection rates of a model test are generally more accurate for TLB designs, although they have a tendency to exceed the nominal significance values. Overall, the current results

support the findings of previous research comparing TB and TLB designs (Jansen & Schulze, 2023b).

The results also indicate, though, that with a rather typical sample size in psychological research of 200, convergence rates and biases are often unacceptable even for TLB designs. The difference in quality between 500 and 1000 respondents is smaller but the results suggest that even a sample size of 500 may not be sufficient for accurate model estimation in many cases relevant for practical applications. With 1000 respondents the quality of the results is generally at least acceptable, though there are a few exceptions where even this many respondents may not be sufficient as shown in the simulation studies. This clearly underscores the fact that the use of Thurstonian models is not available at a cheap rate and comes not only at the price of having to very carefully design the instrument but also to collect data from a significant number of participants.

With regards to the number of traits it was shown that multidimensional blocks yield the best results. The latent trait recovery was lower in general but the linking strategy in multidimensional linked blocks reduces the number of mixed-keyed blocks, which could explain this effect. Additionally, the linking of unidimensional blocks in block composition (3) results in some multidimensional blocks. This could explain the relatively better results for the linked blocks. The linking neutralizes the disadvantages of unidimensional blocks. The inclusion of some mixed-keyed blocks also reduces the disadvantage of blocks with the same loadings. If a second negatively keyed item per trait is included (resulting in a few more mixed-keyed blocks), the results improve but the difference does not appear to be substantial. Lastly, optimal item design would suggest that the number of same- and mixed-keyed pairs should be equal. This is not the case in the present simulation studies. However, including more negatively keyed pairs does not necessarily enhance the results. A design with unequal

numbers of same- and mixed-keyed pairs may not be optimal (Bürkner, 2022) but also not bad.

#### ***4.2.5.1 Recommendations for FC Test Construction***

The results and discussion from the present study suggest several recommendations for the construction of FC assessments (see Table 4.1). First, if possible and social desirability is of concern, then all traits should be defined in way such that high values on these traits are all socially desirable. For example, in the Big Five model of personality, this would require redefining neuroticism as emotional stability and keying the corresponding items accordingly. In other words, the goal is that the content definition and interpretation of each trait supports consistency in same- and mixed-keyed blocks. Second, to successfully construct FC assessments that are resistant to faking, it is crucial to obtain reliable information about the desirability of the items. Previous research has demonstrated how Thurstonian modeling can be used to estimate item utilities (Maydeu-Olivares & Böckenholt, 2005) and social desirability of an item pool in particular (Jansen & Schulze, 2023c). This may result in two-stage use of the Thurstonian models. In a first stage, a TLB factor model is used in order to estimate the social desirability item properties parsimoniously and within the same general model framework (FC) as the assessment to be constructed. The second stage would be to use these item utilities to design the assessment appropriately, which leads to the third recommendation: the desirability ratings should be used to create unlinked blocks with items of equal desirability. Blocks need not all have the same size, but for a consistent assessment presentation using blocks of the same size seems advisable. Although the strength of item desirability may correspond largely to the size of the loadings, this should not be problematic as the fourth recommendation is to link all initial blocks as required in a TLB design. The assessment should include both positively and negatively keyed items.

### Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item Keying

---

By following the four recommendations, the FC questionnaire will provide information about

1. the trait score means by high loading differences,
2. the trait score differences by high factor loading sums,
3. the relative position of each item in a full ranking, by matching items within a block.

**Table 4.1**

*Recommendations for a Stepwise Construction of FC Questionnaire.*

---

	Specific step
Step 0	Create item pool without any restriction (e.g., include positively and negatively keyed items).
Step 1	Define all constructs in the assessment in the same direction of social desirability.
Step 2	Reliably estimate the social desirability of each item.
Step 3	Match items in initial blocks by desirability.
Step 4	Link all initial blocks appropriately.
General	Use at least 500 respondents and 12 items per trait. Use multidimensional blocks or at least no completely unidimensional blocks.

---

Furthermore, the simulation study results demonstrated that the use of multidimensional blocks can enhance model and parameter estimation considerably. This should not be understood, though, as a recommendation to avoid including some same-trait paired comparisons at any cost because they are not detrimental to the quality of the results, especially with TLB designs. On the basis of the results reported it can be recommended to have at least about 12 items per trait and a sample size of at least 500 in order to justify the

expectation of an acceptable estimation precision. Of course, the truism applies here as well as in almost any assessment context: the more items and respondents, the better.

With reference to the number of items per block, there are several advantages of larger block sizes and only a few disadvantages. Matching the blocks by social desirability becomes more challenging with larger blocks, as more items with equal (or at least very similar) desirability are needed. Additionally, larger block sizes can result in increased cognitive load for the respondent leading to ranking errors and reduced motivation. On the other hand, there is some empirical evidence that blocks of four items have no effect on test motivation (Sass et al., 2020) and rankings can still be performed even with blocks of 15 items (Jansen & Schulze, 2023c). Among the advantages of larger blocks is, naturally, that larger blocks provide more information because more paired comparisons can be derived as compared to smaller blocks, even if blocks are not linked. From the perspective of optimal item design, blocks of four items can easily be constructed to have the same number of same- and mixed-keyed paired comparisons by including one item in one direction and three items in the other direction of a trait's interpretation. For example, one negatively and three positively keyed items could be put into one block.

#### ***4.2.5.2 Limitations***

There are some limitations of the current simulation studies that should be noted. First, the simulation studies only employ a single block size, namely triplets. Although it is clear that larger blocks do have advantages (and of blocks of  $k = 2$  disadvantages) and simulation study results on the TB design show better results for larger blocks overall (e.g., Brown & Maydeu-Olivares, 2011), there is still a lack of benchmark results for blocks of  $k \neq 3$  with a TLB design. A simulation study that considers block size as a variable would provide more insight into potential benefits of larger or varied block sizes. Furthermore, simulation studies on the TLB design have not presented reliability conditioned on the trait

values. The reliability approach used in the present study and by Brown and Maydeu-Olivares (2011) assumes that the reliability is uniform across all trait values. However, for the IRT approach, the precision of estimation depends on the trait values and could also be studied as such.

Another potential limitation of the present approach is the use of perfectly simulated data. Data was simulated from a homogeneous sample with relatively low error variances for sample sizes of 200, 500, or 1000 respondents. Due to the relative nature of comparative judgments and the relatively small amount of information in a dichotomous ordinal response, larger error rates and a more heterogeneous group of respondents can realistically be expected in many applications. Including more real data and empirical samples could provide further information on the performance of Thurstonian models and the number of respondents needed to achieve satisfactory convergence rates and appropriate quality of estimates.

#### ***4.2.5.3 Conclusions***

The present study provided further discussion and results on the recently developed Thurstonian linked block design. The findings show that the use of Thurstonian FC modeling with linked blocks is a significant advancement in normative scoring of FC questionnaires. It enables the integration of various types of trait information and the relative position of items in a respondent's ranking. The discussion includes the utilization of Thurstonian factor models for item scaling, specifically desirability scaling. Additionally, the estimation of latent trait scores and the important control for desirability per block are discussed. In conclusion, the results and discussion suggest that the TLB design is a significant improvement of previous designs and probably maybe also the design approach of choice for the development of less susceptible to faking yet reliable assessment procedures.



#### 4.2.6 References

- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, 69(1), 49–56.  
<https://doi.org/10.1111/j.2044-8325.1996.tb00599.x>
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3), 460-502.  
<https://doi.org/10.1177/0013164410375112>
- Brown, A., & Maydeu-Olivares, A. (2018). Modeling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 523–570). Wiley-Blackwell.
- Bürkner, P. C., (2022). On the information obtainable from comparative judgments. *Psychometrika*, 87, 1439-1472. <https://doi.org/10.1007/s11336-022-09843-z>
- Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, 79(5), 827-854. <https://doi.org/10.1177/0013164419832063>
- Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, 104(11), 1347–1368. <https://doi.org/10.1037/ap10000414>
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6(4), 475-494.
- Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, 1-29.  
<https://doi.org/10.1080/00273171.2021.1938960>
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55(4), 243–252. <https://doi.org/10.1037/h0045996>

Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item  
Keying

---

- Jansen, M. T. (2023). *Thurmod: An R package for Thurstonian modeling*. Package submitted.
- Jansen, M. T., & Schulze, R. (2023a). *Linear factor analytic Thurstonian forced-choice models: Current status and issues*. Manuscript submitted.
- Jansen, M. T., & Schulze, R. (2023b). *The Thurstonian linked bock model: Improving Thurstonian modeling for paired comparison and ranking data*. Manuscript submitted.
- Jansen, M. T., & Schulze, R. (2023c). *Item scaling of social desirability using conjoint measurement: A comparison of ratings, paired comparisons, and rankings*. Manuscript in preparation.
- Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389–414. <https://doi.org/10.1177/0013164416646162>
- Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10*(3), 285–304. <https://doi.org/10.1037/1082-989X.10.3.285>
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioural Research, 45*(6), 935-974. <https://doi.org/10.1080/00273171.2010.531231>
- Muthén, L. K., & Muthén, B. O. (1998-2022). *Mplus User's Guide*. Eighth Edition. Muthén & Muthén.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Routledge.
- R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Study 3: Forced-Choice with Thurstonian Models: Assessment Design and the Role of Item  
Keying

---

Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (2020). Taking the test taker's perspective:

Response process and test motivation in multidimensional forced-choice versus rating  
scale instruments. *Assessment*, 27(3), 572-584.

<https://doi.org/10.1177/1073191118762049>

Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires  
resolve the ipsativity issue of forced-choice response formats?. *Educational and  
Psychological Measurement*, 81(2), 262-289.

<https://doi.org/10.1177/0013164420934861>

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

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the  
multidimensional forced-choice format and the rating scale format. *Psychological  
Assessment*, 32(3), 239–253. <https://doi.org/10.1037/pas0000781>

Ziegler, M., MacCann, C., & Roberts, R. (2012). *New perspectives on faking in personality  
assessment*. Oxford University Press.

Ziegler, M., Toomela, A., & Bühner, M. (2009). A reanalysis of Toomela (2003): Spurious  
measurement error as cause for common variance between personality factors.  
*Psychological Test and Assessment Modeling*, 51(1), 65-75.

## 5. Thurmod: An R Package for Thurstonian Modeling

Jansen, M. T (2023a). *Thurmod: An R package for Thurstonian modeling*. Package submitted.

Implementing Thurstonian models for analysis is cumbersome and error prone, especially for the Thurstonian IRT models. This is due to the many model constraints that are to be defined. These include the structured loadings matrix  $\tilde{\Lambda}$  and the structured variance-covariance matrix of correlated error terms  $\tilde{\Psi}^2$ . Furthermore, especially for the unlinked block models, many identification constraints have to be set. The Thurstonian linked block design adds to these challenges, as the dependencies between blocks result in even more structured parameters to account for. Therefore, an automated way of processing and analyzing Thurstonian FC data is necessary. Previous software implementations, such as Excel scripts (Brown & Maydeu-Olivares, 2012) and the R package *thurstonianIRT* (Bürkner, 2019) are mainly limited by lacking the capability to include linked block designs. Furthermore, *thurstonianIRT* can only analyze the Thurstonian IRT block model, the analysis of Thurstonian factor models is not possible.

Thurmod (Jansen, 2023a; electronic supplement of the current work) has several functions that make the simulation of FC data easily available, produce syntaxes for analysis in *Mplus* (Muthén & Muthén, 2022) and *lavaan* (Rosseel, 2012), initiate analysis and read the results. The determination of the number of redundancies among the thresholds and tetrachoric correlations, determination of the rank of design matrix  $\mathbf{A}$ , and correcting the fit based on these redundancies is also implemented. Lastly, it is easy to count the number of additional blocks needed, create linking blocks based on an unlinked design, and produce all model components given a specific design. All these functionalities should make dealing with Thurstonian FC data time efficient and reduce the probability of errors in analyses.

## 6. General Discussion

The assessment of non-observable latent traits of individuals is a common task in both research and applied settings. One common method is to use self-report tests, such as rating scales, to gather information about an individual's personality, behaviors, opinions, and preferences. However, rating scale are often prone to several response biases, socially desirable responding and faking. Not only in situations in which faking is likely to be prevalent, gathering information about the test and its items, for example through item scaling, is an important research task.

To overcome these potential issues, the FC method is frequently proposed as a solution. The FC method eliminates many response biases per design, because there are no classical response options in the first place. However, one of the main disadvantages of the FC method is that the data is inherently ipsative and comparing individuals based on ipsative scores is not meaningful. In the past two decades, advances in FC measurement were made to overcome the ipsativity, with the Thurstonian factor model (Maydeu-Olivares & Böckenholt, 2005) and the Thurstonian IRT model (Maydeu-Olivares & Brown, 2010). These models allow for FC data of paired comparison and ranking data to be analyzed and latent trait scores to be estimated. These latent trait scores are estimated under the assumption of normally distributed latent traits, and therefore, can be interpreted as normative, meaning they can be compared between respondents. Thurstonian models add to the ipsative data the availability of normative data, and therefore solve the incomparability between respondents. Additionally, when items within a block are matched by desirability, the response on which item is preferred should depend solely on the traits of interest and some measurement error. FC data should therefore allow data analysis free of influences from faking and social desirability. The Thurstonian factor model can be used to obtain information about the desirability of

items by having respondents rank them based on the desirability, and subsequently estimating item utilities (for an example see Jansen & Schulze, 2023a).

In the introduction and in the manuscripts of the present thesis, several limitations of Thurstonian modeling were outlined, that limit the usefulness of the result provided with these models. These limitations include the need to work on many paired comparisons even in small item sets, the increasing complexity of a ranking task when more items are considered, and potential computational limitations. To address these issues, block designs, which involve only some item blocks and not all that are derivable, have been proposed (Brown & Maydeu-Olivares, 2011). However, these designs come with their own set of issues, including the seeming paradoxical relationship between mixed-keyed blocks and their fakeability, the many identification constraints needed to identify unlinked block designs, and the question about which design (full versus block) is being tested in the first place or from which design data should be simulated. Additionally, while the importance of matching items within a block by desirability is increasingly recognized (e.g., Bürkner, 2022; Bürkner et al., 2019; Frick et al., 2021; Pavlov et al., 2021; Schulte, Holling et al., 2021) it is often not implemented in practice, with few exceptions (Jansen & Schulze, 2023a; Wetzel & Frick, 2020). This is surprising as the Thurstonian models were introduced as means of item utility estimation that can be used for item scaling (see Böckenholt, 1993, Maydeu-Olivares, 1999). It is likely, that if someone had attempted to use the unlinked block designs for desirability scaling, they would have come to the conclusion, that this is not possible, as one utility mean needs to be fixed per block. Some items would be marked as equally desirable by design via the identification constraints, except if some prior knowledge about the rank among the fixed items is available, which generally is not the case. With all these limitations, the usefulness of the Thurstonian FC method for both, an alternative to the classic response scales and for a more or less fake proof assessment is limited.

The present work addresses the previously discussed limitations of Thurstonian modeling by proposing the linked block design for FC models. Instead of proposing an alternative model, this is done by changing the design of the corresponding FC models. The simple, but powerful, change is that all item blocks must be linked to one another.

The results of two simulation studies show, that the linked block design yields more precise and less biased results in almost all conditions simulated, as well good empirical model rejection rates, though, these rates are still larger than the target values. For block designs with small item loading or utility differences, the linking procedure reduces the disadvantage of low information, by adding blocks with larger loading or utility differences. For unidimensional blocks, the linking adds multidimensional blocks (if at least two traits are studied) with which the potential of low information on persons trait scores, if simultaneously loading differences are small, is reduced. Item and person parameter recovery is larger for linked designs compared to unlinked designs also due to the smaller amount of identification constraints that are needed in linked models. With only one set of identification constraints, all parameters are estimated on the same scale. If only one trait is considered, more items ( $n > 18$ ) are needed to achieve reliable results, even for linked block designs. Summarizing the results of the simulation studies, the following aspects should be considered: a) a full or a linked block design should be used, if possible, b) multidimensional blocks enhance convergence and estimation precision, but using at least some unidimensional blocks is not detrimental for the results, and c) at least 12 items per trait and 500 respondents should be considered; if  $m > 5$  the number of items per trait could be reduced (see Schulte, Holling et al., 2021). Furthermore, theoretical implications are that d) the desirability of items should be measured, e) initial unlinked blocks should be carefully matched by these desirability values, and f) only few mixed-keyed should be used if faking is expected to be a problem.

Furthermore, the linked block design eliminates some design constraints in comparisons to unlinked block designs. For unlinked block designs, typically it is  $n = k \times p$ , meaning that the number of items included in the test is a multiple of the number of items per block. However, with the linked block design, no such constraints are necessary. In the worst case, the addition of one item would result in the need of one additional block. Otherwise, any item can be added to the existing design by linking it in a block with already existing items. It should be noted that the linking procedure may result in some paired comparisons being presented multiple times. Assume the blocks are triplets and the interest is about the four items A, B, C and D. These items could be linked by two triplets, the first consisting of A, B, and C, and the second consisting of A, B, and D. This is in general not a problem, as the error term can be freed to be estimated to account for the potential of intransitive or inconsistent responding, for the comparison between A and B. However, it is not clear which of the potentially inconsistent responses should be used. To answer this question, additional research is needed, which is outside the scope of the current thesis. Lastly, the technical detail of the determination of the number of redundancies among the thresholds and tetrachoric correlations in ranking tasks has been generalized for any FC design.

Another advantage of generalizing the design is, that the fundamental model equations still hold, making it easy to adapt existing research and tests. In case of already existing tests, the inclusion of the linking blocks is easy to implement, however, while it is the simplest way to construct an initial unlinked block design, and then perform the linking of all blocks, the design proposed here is not limited to these cases. Also, the linked block design generalizes the applicability of a block design from the IRT model to the factor model, and thus makes item scaling possible in a more parsimonious representation than with all  $\tilde{n}$  comparisons. Item utilities can now be estimated with blocks by using only one set of identification constraints. Finally, a very important contribution related to the present work is



the development of a powerful R package, which allows for easy model syntaxes writing, analysis and more (Jansen, 2023a).

However, there are some limitations to the results of the current work in particular and the Thurstonian FC model in general. First, both simulation studies considered only a limited number of item sets and sample sizes. This was necessary to reduce simulation time of the already extensive simulation studies. For the sample size, one result is that for linked block designs 500 respondents can be sufficient for model estimation, while a sample size of  $N = 200$  is not. It is unclear where the sweet-spot is, except for it can be expected between these two levels. The same is true for the number of items, as 12 items per trait seem to be sufficient, but 6 items per trait are not. Also, the influences of the block size is ignored completely. It is assumed that larger blocks yield more information and therefore should contribute to better model estimation performance (as shown for unlinked designs by Brown & Maydeu-Olivares, 2011). However, the potential benefits of larger blocks for linked designs are yet unclear. Furthermore, the manuscripts focused exclusively on simulations as a means to study the Thurstonian linked block design. First, the analyses of simulation studies considered perfectly simulated data. In many real-world applications it can be expected that data include larger error rates and a more heterogeneous group of respondents. Second, with simulated data, aspects such as the cognitive effort and motivation (as in Sass et al., 2020) to work on larger blocks of potentially repeating item comparisons cannot be studied.

On the side of assessment time, the use of a linked block design comes at the cost of more blocks to present. Thurstonian modeling is not cheaply available, it comes at the price of collecting a significant number of respondents with the need to carefully design the assessment. However, especially in situations where faking is likely to occur, it is argued that the benefit in less biased and more reliable results is worth the price.

Still, the empirical and theoretical contributions and advances of the present work result in considerable improvements of Thurstonian FC measurement, strengthening the case for using the FC method for dominance response items in the future.

That being said, the Thurstonian FC method is far from being over-researched, in the contrary, there are many open research questions. First, some of the work already done for the MFC format has to be adopted for the linked block design. These include, but are not limited to, comparing the FC method with response scales (e.g., Dueber et al., 2019; Wetzel et al., 2021; Jansen & Schulze, 2023a) testing the normativity of latent trait scores (e.g., Frick et al., 2021), the examination of construct validity (Fischer et al., 2019; Lee et al., 2021; Ng et al., 2021; Wetzel & Frick, 2020) and, of course, testing the fakeability of FC tests (Ng et al., 2021; Pavlov et al., 2019; Wetzel et al., 2021). Furthermore, models such as the faking mixture model (Frick, 2022) can further be used to study the fakeability of constructed blocks. Comparisons in block-fakeability between linked and unlinked block design, especially in light of mixed-keyed blocks, can also enhance the understanding of Thurstonian FC modeling. Once when enough studies are done on the fakeability on FC measures with linked blocks, an updated meta-analysis would be fruitful (as in Cao & Drasgow, 2019). Additionally, questions commonly asked in test construction for both, IRT and classical test theory (CTT) should be considered, such as determining the number of traits of a given set of items. While previous research has focused on confirmatory analyses, Thurstonian models are not limited to this case, analyses can also be done exploratorily (e.g., Maydeu-Olivares & Böckenholt, 2005). Thus, the exploration of dimensionality and the identification of which items are meaningful for which trait, can also be performed.

Another interesting research question that should be considered in the future is what makes a good item in the context of the FC method. There is no reason to believe that an item has the exact properties within the FC context, as it has when applied with a response scale,

as the response to an item is not independent to the other items in the FC context. Research has shown that item properties may vary depending on the blocks they are used in (e.g., Lin & Brown, 2017). Therefore, even if an item has good difficulty values, is discriminative, fits the proposed model, contributes to high reliability (both in IRT and CTT) and validity on a response scale, it may not necessarily be true in the context of FC measurement. All these aspects and questions, which are important for test construction in general, also apply to the FC method.

The general questions of test construction named before can also be subsumed as *item selection*. In the context of FC and using a (linked) block design, this process has to be separated into two sub-problems: selecting useful items from the item universe and constructing blocks based on these items. This includes studying the dimensionality of each item and how closely it is related to a respective trait. From all items not eliminated in the first step, blocks need to be constructed in a way, that the FC test is reliable and valid. As discussed, this includes matching items by desirability, however, this is only one of many aspects. For faking there already exists the faking mixture model, which gives information on the fakeability per block, highly fakeable blocks should potentially be eliminated (Frick, 2022). But more generally, studying which blocks work well for the intended application context, and which do not, is necessary. As pointed out before, given a set of items, the number of combinations to create blocks is practically uncountable. To thoroughly investigate FC tests based on a set of items, it is not sufficient to test just one linked block design and eliminate some items or blocks if necessary. Some items may work well in one block, but not in others. To examine a large number of combinations, empirical information on many (if not all) possible items to block assignments is needed. If full information on all paired comparisons derivable by the set of items were available for a sample, different block combinations could be tested and cross-validated without the need for more empirical data.

Obviously, it would not be practical to construct many FC tests based on the same set of items and have tens of thousands of respondents take one or more of the tests. This approach would not be economical and there is a risk that a better block configuration, may not have been constructed and tested. So how to get to the desired information?

While there must still be research on what constitutes as a good item or block in the FC setting, work has been done to make the information on all paired comparisons available. The goal is to get the full amount of information, with the least amount of effort (number of blocks) from each respondent. In the FC setting, this is equivalent to a complete ranking of all items of interest. There are two possible approaches to achieve this. The first approach is to have respondents rank all items at once. This can be a complex task for respondents and it is not recommended to present large item sets at once (as discussed in Sass et al., 2020). If one big block is not suitable, the second approach is to use a sorting algorithm. Sorting algorithms perform the task of ordering a list of objects based on a certain criterion, often under the assumption of transitivity. Research on sorting algorithms focuses on computer science and thus, evaluates algorithms based on their computational complexity (worst-, average- and best-case number of comparisons or object changes) and use of computer resources in general (time and memory). A sorting algorithm is considered to be better than another if it uses fewer resources or is less complex (for detailed information and overview see Knuth, 1998). There are already some sorting algorithms that are almost optimal in the terms that the worst-case number of comparisons needed to sort a set of objects is near the information theoretical lower bound. However, as the main field of use is on computer science, these algorithms are also optimized corresponding to their application, that includes minimizing the time needed to run the algorithm and minimizing changes in the object list. Comparisons need to be fast, often in the range of milliseconds or even nanoseconds. However, for applications with human judges, such time dimensions are irrelevant, as comparisons need multiple seconds,

even if they are fast. There exists some yet unpublished work on an efficient sorting algorithm that is adapted to the use by human respondents and focuses mainly on limiting the worst-case number of comparisons a respondent has to work on (Jansen & Doeblner, 2023). The main disadvantage of any sorting algorithm is that they are only applicable for paired comparisons, that is, rankings of blocks of two objects. This makes sense, as any block with more than two items would include comparisons that are determinable by transitivity, and hence use more computational resources than needed. For human respondents, this is not the case. For test motivation, working on 20 triplets or 15 quads can be better than working on 30 paired comparisons. Also, with existing information about a ranking, the use of computer adaptive testing designs in FC measurement could be implemented. Therefore, it was fruitful to generalize the sorting algorithm by Jansen and Doeblner (2023) for any block size  $k$  (Jansen, 2023b). These algorithms present blocks of items to the respondent that are optimal to produce a full ranking, given the previous answers of that respondent. The number of blocks needed can be further optimized, if information on the item ranking for a respondent is available. However, a downside of the procedure is, that still many blocks need to be responded to, making this procedure, while promising, laborious and not trivial.

These algorithms could potentially make full data collection more efficient and can be easily combined with the Thurstonian linked block approach. The first step of both algorithms is to build a priority queue by fully linking all blocks, either by using a random starting point, resulting in different initial blocks for each respondent, or by a fixed manner, as in a fixed Thurstonian linked block design. If information on many or all blocks is available, a wider pool of item selection procedures should be available.

On a final note, alternatives to the “classical” FC method can be used and adapted, such as the Graded-Block design (Brown & Maydeu-Olivares, 2018). Here, the idea is to ask respondents to give a graded response on a scale of how much more they prefer one item over

another, after initial ranking. There are some potential advantages by a graded response FC design. For example, it addresses a common critique that respondents feel they do not have a real choice, especially when the items are equally desirable (e.g., Bartram & Brown, 2003; Sass et al., 2020). For respondents to clearly state that the choice was difficult by marking that the preference is small, potentially leads to a more engaged test taking and less participant dropout. Additionally, reliability can potentially be enhanced as a graded response on an ordered scale with, for example, five categories provide more information than a binary indicator. This could reduce the number of items needed per trait for reliable trait estimation. However, a Graded-Block design also requires more work as for each paired comparison, the degree of preference must be stated (Brown & Maydeu-Olivares, 2018). An idea for adaptation would be to rank a block of items on a slider, with the distance between items indicating the degree of preference. Some studies suggest that latent trait estimates are more reliable (Lingel et al. 2022) with the Graded-Block design and, again, highlight the importance of desirability matching, as the faking effect correlates highly with the social desirability differences between statements (Schulte, Kaup et al., 2021). The Graded-Block design is in structure equivalent to the Thurstonian models, but  $y_i$  is not binary anymore, which makes it a Thurstonian model for polytomous data (Brown & Maydeu-Olivares, 2018). This should make the use of a Graded-linked-block design easy to implement, and may enhance estimation even more.

In summary the present work proposes a new approach to Thurstonian modeling that outperforms previously studied Thurstone models and designs. The linked block design can easily be adapted to existing work. Further advancements in Thurstonian FC modeling, and future adaptations and extensions could make it a standard method for assessments in psychological science, applications, and beyond.

## References

- Aronoff, G. M., Mandel, S., Genovese, E., Maitz, E. A., Dorto, A. J., Klimek, E. H., & Staats, T. E. (2007). Evaluating malingering in contested injury or illness. *Pain Practice, 7*(2), 178–204. <https://doi.org/10.1111/j.1533-2500.2007.00126.x>
- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*(1), 49–56. <https://doi.org/10.1111/j.2044-8325.1996.tb00599.x>
- Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335–344. <https://doi.org/10.1016/j.jrp.2008.12.013>
- Bäckström, M., Björklund, F., & Larsson, M. R. (2012). Social desirability in personality assessment: Outline of a model to explain individual differences. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 201–213). Oxford University Press.
- Bartram, D., & Brown, A. (2003). *Test-taker reactions to online completion of the OPQ32i*. SHL Group.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. <https://doi.org/10.1111/j.1468-2389.2006.00354.x>
- Böckenholt, U. (1993). Applications of Thurstonian Models to Ranking Data. In Fligner, M. A., Verducci, J. S. (Eds.), *Probability models and statistical analyses for ranking data. Lecture Notes in Statistics, vol 80* (pp. 157–172). Springer. [https://doi.org/10.1007/978-1-4612-2738-0\\_9](https://doi.org/10.1007/978-1-4612-2738-0_9)

- Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. <https://doi.org/10.1007/s11336-014-9434-9>
- Brown, A., & Bartram, D. (2009–2011). *OPQ32r Technical Manual*. SHL group.
- Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148. <https://doi.org/10.1177/1094428116668036>
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. <https://doi.org/10.1177/0013164410375112>
- Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147. <https://doi.org/10.3758/s13428-012-0217-x>
- Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 516–529. <https://doi.org/10.1080/10705511.2017.1392247>
- Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662–1663.
- Bürkner, P. C., (2022). On the information obtainable from comparative judgments. *Psychometrika*, *87*, 1439–1472. <https://doi.org/10.1007/s11336-022-09843-z>
- Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827–854. <https://doi.org/10.1177/0013164419832063>



- Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. <https://doi.org/10.1037/ap10000414>
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307. [https://doi.org/10.1207/s15327043hup1803\\_4](https://doi.org/10.1207/s15327043hup1803_4)
- Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. [Doctoral dissertation]. Retrieved from <http://www.psychometrika.org/journal/online/MN14.pdf>
- Coombs, C. H. (1960). A theory of data. *Psychological Review, 67*(3), 143–159. <https://doi.org/10.1037/h0047773>
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475–494.
- Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measures decays when applicants fake. *Academy of Management Proceedings, 1*, 127–131. <https://doi.org/10.5465/ambpp.1996.4979062>
- Dueber, D. M., Love, A. M., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement, 79*(1), 108–128. <https://doi.org/10.1177/00131644177527>
- Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*(2), 13–24. <https://doi.org/10.1111/j.1744-6570.1962.tb01843.x>

- Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37*(2), 90–93. <https://doi.org/10.1037/h0058073>
- Edwards, A. L. & Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17*, 169–180. <https://doi.org/10.1007/BF02288780>
- Fisher, P. A., Robie, C., Christiansen, N. D., Speer, A. B., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus classical test theory scoring. *Personnel Assessment and Decisions, 5*(1), 49–61. <https://doi.org/10.25035/pad.2019.01.003>
- Frick, S. (2022). Modeling faking in the multidimensional forced-choice format: The faking mixture model. *Psychometrika, 87*, 773–794. <https://doi.org/10.1007/s11336-021-09818-6>
- Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research, 58*(1), 1–29. <https://doi.org/10.1080/00273171.2021.1938960>
- Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment, 25*(4), 513–526. <https://doi.org/10.1177/1073191116641181>
- Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9–24. <https://doi.org/10.1037/0021-9010.91.1.9>

- Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, 25(5), 560–576. <https://doi.org/10.1037/met0000249>
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74(3), 167–184. <https://doi.org/10.1037/h0029780>
- Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution?. *Human Performance*, 13(4), 371–388. [https://doi.org/10.1207/S15327043HUP1304\\_3](https://doi.org/10.1207/S15327043HUP1304_3)
- Jansen, M. T. (2023a). *Thurmod: An R package for Thurstonian modeling*. Package submitted.
- Jansen, M. T. (2023b). *N-ary Path Sort: Adaptively sorting items via forced-choice blocks of size N*. Manuscript in preparation.
- Jansen, M. T., & Doebler, P. (2023). *Binary Path Sort: An adaptive sorting algorithm for forced-choice designs with focus on a single respondent*. Manuscript in preparation.
- Jansen, M. T., & Schulze, R. (2023a). *Item scaling of social desirability using conjoint measurement: A comparison of ratings, paired comparisons, and rankings*. Manuscript in preparation.
- Jansen, M. T., & Schulze, R. (2023b). *Linear factor analytic Thurstonian forced-choice models: Current status and issues*. Manuscript submitted.
- Jansen, M. T., & Schulze, R. (2023c). *The Thurstonian linked block model: Improving Thurstonian modeling for paired comparison and ranking data*. Manuscript submitted.
- Jansen, M. T., & Schulze, R. (2023d). *Forced-choice with Thurstonian models: Assessment design and the role of item keying*. Manuscript submitted.

- King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–207. [https://doi.org/10.1007/978-3-531-91826-6\\_16](https://doi.org/10.1007/978-3-531-91826-6_16)
- King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 46–66. <https://doi.org/10.1093/pan/mpl011>
- Knuth, D. E. (1998). *Art of computer programming, volume 3: Sorting and searching*. Addison-Wesley Professional.
- Lanyon, R. I. (1993). Development of scales to assess specific deception strategies in the psychological screening inventory. *Psychological Assessment*, *5*(3), 324–329. <https://doi.org/10.1037/1040-3590.5.3.324>
- Lee, H., & Smith, W. Z. (2020). Fit indices for measurement invariance tests in the Thurstonian IRT model. *Applied Psychological Measurement*, *44*(4), 282–295. <https://doi.org/10.1177/0146621619893785>
- Lee, P., Joo, S. H., & Stark, S. (2021). Detecting DIF in multidimensional forced choice measures using the Thurstonian item response theory model. *Organizational Research Methods*, *24*(4), 739–771. <https://doi.org/10.1177/1094428120959822>
- Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389–414. <https://doi.org/10.1177/0013164416646162>
- Lingel, H., Bürkner, P. C., Melchers, K. G., & Schulte, N. (2022). Measuring personality when stakes are high: are graded paired comparisons a more reliable alternative to traditional forced-choice methods?. *PsyArXiv*.

- Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64*(3), 325–340.  
<https://doi.org/10.1007/BF02294299>
- Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*(3), 285–304.  
<https://doi.org/10.1037/1082-989X.10.3.285>
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*(6), 935–974.  
<https://doi.org/10.1080/00273171.2010.531231>
- McCloy, R. A., Heggstad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*(2), 222–248.  
<https://doi.org/10.1177/1094428105275374>
- Merten, T. (2013). *Beschwerdenvalidierung*. Hogrefe.
- Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298.  
<https://doi.org/10.1080/1359432X.2010.550680>
- Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, *125*(5), 656–688. <https://doi.org/10.1037/rev0000111>
- Muthén, B.O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. <https://doi.org/10.1007/BF02293813>
- Muthén, L. K., & Muthén, B. O. (1998-2022). *Mplus User's Guide*. Eighth Edition. Muthén & Muthén.

## References

---

- Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment, 103*(2), 224–237. <https://doi.org/10.1080/00223891.2020.1739056>
- Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires: Current issues in theory and measurement* (pp. 143–165). Springer.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press. <https://doi.org/10.1016/B978-0-12-590241-0.50006-X>
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Routledge.
- Paulhus, D. L. & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford Press.
- Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and Likert scores. *Organizational Research Methods, 22*(3), 710–739. <https://doi.org/10.1177/1094428117753683>
- Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*. <https://doi.org/10.1016/j.paid.2021.111114>

- Rosenthal, R. & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, 8(3), 183–189.  
<https://doi.org/10.1002/bs.3830080302>
- Rosseel Y (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, 23(1), 3–30. <https://doi.org/10.1080/1359432X.2012.716198>
- Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, 27(3), 572–584.  
<https://doi.org/10.1177/1073191118762049>
- Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats?. *Educational and Psychological Measurement*, 81(2), 262–289.  
<https://doi.org/10.1177/0013164420934861>
- Schulte, N., Kaup, L., Bürkner, P. C., & Holling, H. (2021). The fakeability of personality measurement with graded paired comparisons. *PsyArXiv*.
- 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(3), 184–203. <https://doi.org/10.1177/0146621604273988>
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>

- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 3(4), 529–554. <https://doi.org/10.1086/214483>
- Velicer, W. F. & Weiner, B. J. (1975). Effects of sophistication and faking sets on the Eysenck Personality Inventory. *Psychological Reports*, 37(1), 71–73. <https://doi.org/10.2466/pr0.1975.37.1.71>
- Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2), 197–210. <https://doi.org/10.1177/00131649921969802>
- Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment*, 27(4), 706–718. <https://doi.org/10.1177/1073191119843585>
- Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus Likert responses on an occupational big five questionnaire. *Journal of Individual Differences*, 40(3), 134–148. <https://doi.org/10.1027/1614-0001/a000285>
- Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, 32(3), 239–253. <https://doi.org/10.1037/pas0000781>
- Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, 33(2), 156–170. <https://doi.org/10.1037/pas0000971>
- Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1), 79–92. <https://doi.org/10.1348/000711099158973>



## References

---

Ziegler, M., MacCann, C., & Roberts, R. (Eds.). (2012). *New perspectives on faking in personality assessment*. Oxford University Press.

## List of Figures

<b>Figure 1.1</b> <i>The Discriminative Process Between Stimuli <math>i</math> (blue) and <math>j</math> (red).</i> .....	8
<b>Figure 1.2</b> <i>Covariance Structure of a Thurstonian Factor Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	12
<b>Figure 1.3</b> <i>Covariance Structure of a Thurstonian IRT Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	15
<b>Figure 2.1</b> <i>Examples for the Forced-Choice Format.</i> .....	31
<b>Figure 2.2</b> <i>Examples of the Discriminative Process of One Paired Comparison Between Stimulus <math>i</math> and Stimulus <math>j</math>.</i> .....	35
<b>Figure 2.3</b> <i>Examples of a Covariance Structure of a Simple Thurstonian Model for <math>n = 4</math> Items.</i> .....	40
<b>Figure 2.4</b> <i>Examples of a Covariance Structure of a Thurstonian Factor Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	42
<b>Figure 2.5</b> <i>Examples of a Covariance Structure of a Thurstonian IRT Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	48
<b>Figure 3.1</b> <i>Examples for the Forced-Choice Format.</i> .....	82
<b>Figure 3.2</b> <i>Example of a Covariance Structure of a Thurstonian Factor Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	89
<b>Figure 3.3</b> <i>Example of a Covariance Structure of a Thurstonian IRT Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	92
<b>Figure 3.4</b> <i>Example of Responses for Each Block Compared to the Utility Scale.</i> .....	98
<b>Figure 3.5</b> <i>Example of Responses for Each Block Compared to the Utility Scale in a Thurstonian Linked Block Design. In the Linking Block, the Order of Attractiveness is Item 7, Item 1, then Item 4.</i> .....	100
<b>Figure 3.6</b> <i>Empirical Rejection Rates for the Simulation Studies with Five Factors.</i> .....	118
<b>Figure 4.1</b> <i>Examples for the Forced-Choice Format.</i> .....	141
<b>Figure 4.2</b> <i>Example of a Covariance Structure of a Thurstonian Factor Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	146
<b>Figure 4.3</b> <i>Example of a Covariance Structure of a Thurstonian IRT Model for <math>n = 4</math> and <math>m = 1</math>.</i> .....	146
<b>Figure 4.4</b> <i>Example for the Linking Procedure of a Thurstonian Linked Block Design.</i> .....	157
<b>Figure 4.5</b> <i>Results for Convergence Rates, Bias and the Correlation Between True and Estimated Parameters for Three Uncorrelated Traits. One Item per Factor (Three Items Total) is Negatively Keyed.</i> .....	163

**Figure 4.6** Results for Convergence Rates, Bias and the Correlation Between True and Estimated Parameters for Three Uncorrelated Traits. Two Items per Trait (Six Items Total) are Negatively Keyed. .... 164

**Figure 4.7** Results for the Recovery of the Three Uncorrelated Traits Study (Simulation Study 1). .... 166

**Figure 4.8** Empirical Rejection Rates for the Three Uncorrelated Traits Study (Simulation Study 1). .... 167

**Figure 4.9** Bias and SEs of Latent Correlations for Three Correlated Traits (Simulation Study 2) ..... 170

**Figure 4.10** Results for the Recovery of Three Correlated Traits (Simulation Study 2). ..... 170

**Figure 4.11** Empirical Rejection Rates of the IRT Models for the Three Correlated Traits Study..... 171

## List of Tables

<b>Table 2.1</b> <i>The Relative Bias of the Three Block Designs Compared to the Full Design. ....</i>	62
<b>Table 2.2</b> <i>Results for the Latent Trait Recoveries and Reliabilities of the Simulation Study...</i>	67
<b>Table 3.1</b> <i>Thurstonian IRT Models: Results from Simulation Study 1.....</i>	109
<b>Table 3.2</b> <i>Thurstonian Factor Models: Results from Simulation Study 1.....</i>	110
<b>Table 3.3</b> <i>Results for Valid Iterations, Loading Estimates, and Standard Errors in the Simulation Study with Five Uncorrelated Traits.....</i>	115
<b>Table 3.4</b> <i>Results for Utility, Factor Correlation Estimates and Standard Errors in the Simulation Study with Five Uncorrelated Traits.....</i>	116
<b>Table 3.5</b> <i>Results for the Actual and Empirical Latent Trait Recovery for Five Uncorrelated (Simulation Study 2) and Five Correlated Traits (Simulation Study 3).....</i>	117
<b>Table 3.6</b> <i>Results for the Trait Correlations in Simulation Study 3. ....</i>	120
<b>Table 3.7</b> <i>Illustration of Redundancies for Different Scenarios. ....</i>	134
<b>Table 4.1</b> <i>Recommendations for a Stepwise Construction of FC Questionnaire.....</i>	176

## Nomenclature

<b>A</b>	Design matrix
$a, b$	Indices of latent traits
$\alpha, \beta$	Intercept and slope
$C_{n,k}$	Multinomial coefficient
<b>D</b>	Scaling matrix
$e, \mathbf{e}$	(Vector of) error terms of utilities
<b><math>\eta, \Phi</math></b>	Vector and covariance matrix of latent traits
$\gamma$	Intercepts
$h, i, j$	Indices of items
<b>I</b>	Identity matrix
<b>I(<math>\eta</math>)</b>	Information functions
$k$	Number of items per block
$l$	Index of paired comparisons
$\lambda, \Lambda$	(Matrix of) factor loadings for traits
$\tilde{\lambda}, \tilde{\Lambda}$	Reparametrized (matrix of) factor loadings for traits for the Thurstonian IRT model
$m$	Number of traits
$\mu_t, \boldsymbol{\mu}_t$	(Vector of) expected values of utilities t
$\boldsymbol{\mu}_y^*$	Vector of expected values of latent latent differences
$n$	Number of items
$\tilde{n}, \tilde{k}$	Number of paired comparisons derivable by $n$ and $k$ respectively
$\omega^2, \Omega^2$	(Covariance matrix of) error terms of utilities

## Nomenclature

---

$p$	Number of blocks
$\psi^2, \boldsymbol{\varepsilon}, \boldsymbol{\Psi}^2$	(Vector and covariance matrix of) error terms of latent traits (unique factors)
$\tilde{\psi}^2, \tilde{\boldsymbol{\varepsilon}}, \tilde{\boldsymbol{\Psi}}^2$	Reparametrized (vector and covariance matrix of) error terms of latent traits for the Thurstonian IRT model
$r$	Number of redundancies among thresholds and tetrachoric correlations
$\rho$	Reliability
$\sigma_t, \boldsymbol{\Sigma}_t$	Variance and covariance matrix of utilities $t$
$\boldsymbol{\Sigma}_y^*$	Covariance matrix of the latent differences
$t, \mathbf{t}$	(Vector of) utilities
$\tau$	Thresholds
$y, \mathbf{y}$	(Vector of) observed binary responses
$y^*, \mathbf{y}^*$	(Vector of) latent differences
$\mathbf{z}^*, \mathbf{P}_z^*$	Vector and correlation matrix of standardized latent differences

### Abbreviations

FC	Forced-choice
MFC	Multidimensional forced-choice
TB	Thurstonian block design
TLB	Thurstonian linked block design
TMB	Thurstonian multidimensional block design
C design	Completely linked design
P design	Partially linked design
U design	Unlinked design

## CRediT Authors Statement

Jansen, M. T., & Schulze, R. (2023b). *Linear factor analytic Thurstonian forced-choice measurement: Current status and issues*. Manuscript submitted for publication in Educational and Psychological Measurement.

Markus Thomas Jansen: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration

Ralf Schulze: Writing - Review & Editing, Resources, Supervision

Jansen, M. T., & Schulze, R. (2023c). *The Thurstonian linked block design: Improving Thurstonian modeling for paired comparison and ranking data*. Manuscript submitted for publication in Psychometrika.

Markus Thomas Jansen: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration

Ralf Schulze: Conceptualization, Writing - Review & Editing, Resources, Supervision

Jansen, M. T., & Schulze, R. (2023d). *Forced-Choice with Thurstonian models: Assessment design and the role of item keying*. Manuscript submitted for publication in Multivariate Behavioral Research.

Markus Thomas Jansen: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration

Ralf Schulze: Writing - Review & Editing, Resources, Supervision

## Statement of Originality

I hereby declare that I am the sole author of the thesis under the title “*Advances in Thurstonian Forced-Choice Modeling*” and that I did not use any other aids or resources than the ones stated. Those parts of the paper that were taken from other works, either as quote or paraphrase, are marked by respective statements of sources. Furthermore, I declare that this term paper has not been handed in for a different academic assessment by me or another person.

---

Markus Thomas Jansen

---

Place, Date