Item Response Theory for the Analysis and Construction of Multidimensional Forced-Choice Tests

Susanne Frick

Inaugural Dissertation
Submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim

19 October 2021

Supervisors:
Prof. Dr. Thorsten Meiser
Prof. Dr. Eunike Wetzel

Dean of the School of Social Sciences:
Prof. Dr. Michael Diehl

Thesis Reviewers:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Christoph Klauer

Defense Committee:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Christoph Klauer
Prof. Dr. Thorsten Meiser

Thesis Defense:
9 December 2021

Soli Deo Gloria

Contents

Acknowledgments

Danke an

... Thorsten Meiser, für seine hilfreichen Ideen zu meinen Projekten und Papern und alle sonstige Unterstützung.

... Eunike Wetzel, die mich mit Thurstonian IRT und MFC in Berührung gebracht hat, mich seit meiner Hiwi-Zeit begleitet, die mir so Vieles beigebracht hat und immer bereit ist, mit mir alle großen und kleinen Fragen zu diskutieren.

... Christoph Klauer für die Begutachtung meiner Dissertation und seine klare, mathematische Perspektive auf meine Projekte.

... Edgar Erdfelder für die Begutachtung meiner Dissertation und die Diskussionen auf Retreats.

... Anna Brown for all her advice and help, for her clear understanding of IRT models and response formats and her applied perspective and experiences.

... Safir Yousfi für seinen R-Code, ohne den ich wahrscheinlich noch ein paar Monate länger beschäftigt gewesen wäre.

... Nils, Franziska, Mirka, Marcel und Viola für die gemeinsamen Mittagessen und (digitalen) Kaffeepausen und alle Gespräche über das Leben in der Wissenschaft.

... alle SMiPsters für alle wissenschaftlichen und nicht-wissenschaftlichen Unterhaltungen und die gemeinsame Zeit in Parks, Restaurants und auf Reisen.

... Nadja, Ruth-Maria, Flo, Lea und Joshua, die mit mir schöne Momente geteilt, mich abgelenkt und aufgemuntert und für mich gebetet haben.

... meine Eltern, die mich in allem unterstützt haben und immer für mich da sind.

... die Leute von Studenten für Christus, von der Freien Evangelischen Gemeinde Mannheim und aus Gabriele Hilsheimers Flötenensemble, die mir das Ankommen und Leben in Mannheim erleichtert und verschönert haben.

... meine Sprachengebetsfreunde Rebekka und Jonathan und den Jungakademiker-Hauskreis Karlsruhe, mit denen ich Glauben und Leben online teilen kann.

Abstract

The multidimensional forced-choice (MFC) format has been proposed as an alternative to rating scales. In the MFC format, respondents indicate their relative preference for items measuring different attributes within blocks. Test construction for the MFC format is complex because how the items are combined affects the properties of the test. The aim of this thesis was to investigate and further develop IRT methods for the MFC format that can help to improve MFC test construction, focusing on the Thurstonian IRT model and a ranking instruction.

In the first manuscript (Frick et al., 2021), we conducted an extensive simulation study on the normativity of Thurstonian IRT trait estimates. We investigated realistic test designs, removed a potential confounding with item parameter bias and compared recovery to that from classical test theory scoring and from rating scale and true-false formats. We found that with all positively keyed items, trait estimates showed ipsative properties. However, with mixed item keys, they were insensitive to otherwise suboptimal test designs. In an empirical study, we found that construct validity in the MFC format with three-item blocks was lower and criterion validity equal to the true-false format.

In the second manuscript (Frick, 2021b), I developed the Faking Mixture model, a model for faking in the MFC format that allows to estimate the fakability of individual MFC blocks. A simulation study showed good parameter recovery. An empirical validation showed that the model can capture expected differences in item desirability, but also that matched blocks were not fully fake-proof. Therefore, it is worth to apply the Faking Mixture model in order to reduce fakability by removing or modifying blocks during test construction.

In the third manuscript (Frick, 2021a), I proposed methods to estimate and summarize Fisher information for Thurstonian IRT models on the block level. Three simulation studies showed that the methods can accurately recover true information and are useful for test construction. It was examined how the proposed information summaries can be combined with algorithms for automated test assembly. Thus, block information can be used to assemble MFC tests that maximize reliability and have an ideal test design.

In summary, this thesis provided both new methods and guidelines for MFC test construction. Modeling the block level did and will help to adequately capture the relative response process and item interactions and it can provide avenues for further psychometric developments.

Articles

This cumulative thesis is based on the following three manuscripts:

Manuscript I

Frick, S., Brown, A. & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research.* Advance online publication. https://doi.org/10.1080/00273171.2021.1938960

Manuscript II

Frick, S. (2021). Modeling faking in the multidimensional forced-choice format – The Faking Mixture model. *Psychometrika.* Advance online publication. https://doi.org/10.1007/s11336-021-09818-6

Manuscript III

Frick, S. (2021). Block information in the Thurstonian item response model. *Manuscript submitted for publication to Psychometrika.*

This research deals with investigating and further developing item response theory methods for multidimensional forced-choice (MFC) tests. In the following, I will first give a short overview of the MFC format and its advantages in comparison to rating scales, of challenges in MFC test construction and of item response theory models for MFC tests, especially of the Thurstonian item response model. Then, I will summarize the three manuscripts. In the end, I will discuss implications and future research directions for MFC test construction and psychometric modeling. The full manuscripts are appended to this synopsis.

1 Introduction

Tests are frequently used to assess personality and draw inferences about respondents' trait levels. For example, employers use personality tests to assess whether applicants possess the characteristics needed for the job. Psychotherapists routinely use personality tests as part of the initial assessment. Since important life outcomes may depend on the results of personality tests, test scores should measure the intended construct precisely and free of irrelevant influences. In other terms, test scores should be reliable and valid. Most personality tests use a rating scale format (e.g., strongly disagree, disagree ...). However, rating scales often suffer from systematic influences on the response beyond the construct intended to measure, termed response biases (Paulhus, 1991). For example, respondents might show preferences for certain categories, called response styles (Henninger & Meiser, 2020; Wetzel, Böhnke, et al., 2016). Or, in a so-called high-stakes situation (e.g., when applying for a job), respondents might distort their responses in order to leave a certain impression, a response behavior called faking (MacCann et al., 2011). Response biases can diminish reliability and validity. For example, response styles can change correlations between scale scores (Moors, 2012). Faking can result in mean increases of trait scores of .1 to .6 SD when using rating scales (Birkeland et al., 2006; Viswesvaran & Ones, 1999). To prevent response biases emerging from the use of rating scales, the multidimensional forced-choice (MFC) format has been proposed as an alternative.

1.1 Multidimensional Forced-Choice versus Rating Scales

In the MFC format, several items measuring different attributes are combined into blocks and respondents indicate their relative preference for the items. In such, the MFC format is both an item and a response format. I refer to it as a response format in the following. Typical response instructions include ranking all items (for an example, see Figure 1) or selecting the items that describe oneself most and/or least. This research focuses on MFC blocks with a ranking instruction, because this response instruction (potentially) provides the largest amount of information and therewith the highest reliability (Brown & Maydeu-Olivares, 2011). Additionally, the number of items per block can vary, with two to four items being the most common.

Research interest in the MFC format has increased in recent years as evidenced by the growing number of articles published on this topic (Figure 2). Further, the MFC format

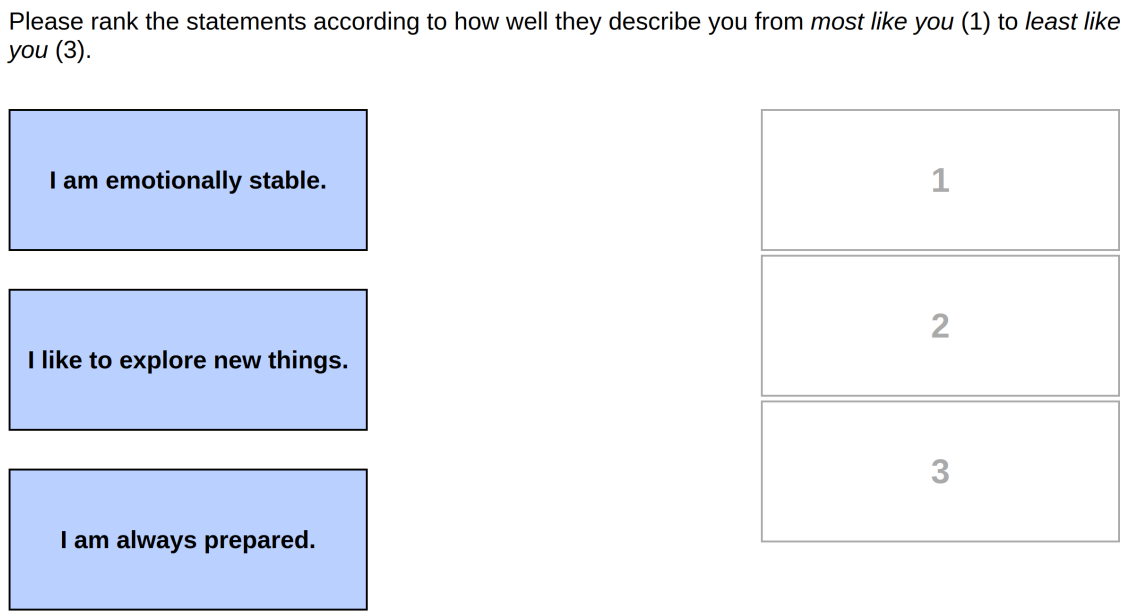

FIGURE 1: Example of a multidimensional forced-choice block from the Big Five Triplets (Wetzel & Frick, 2020).

has become popular in assessment which is reflected in several tests that use this format. For example, it is used to assess work-related personality in TAPAS (Drasgow et al., 2012), OPQ (Brown & Bartram, 2009–2011), and the personality test by TalentQ (Holdsworth, 2006).

The MFC format allows to prevent, or at least reduce, some of the response biases that occur with rating scales (Brown & Maydeu-Olivares, 2018a). From a theoretical perspective, uniform response biases, such as halo effects or acquiescence, are avoided, because the relative preferences remain the same if the preferences for all items increase to the same extent (Brown et al., 2017). This has been confirmed empirically: Halo effects (Brown et al., 2017) were reduced with an MFC as compared to a rating scale format. Furthermore, biases that arise from the use of rating scales, such as response styles, cannot occur (Brown & Maydeu-Olivares, 2018a).

The MFC format can prevent faking when the items within blocks are matched for their (social) desirability, as was first proposed by Edwards (1953). This is based on the assumption that respondents who want to fake would first try to rank the items according to how desirable they are. If this is not possible, because all items are equally desirable, they give an honest response instead (Berkshire, 1958; Gordon, 1951). Figure 3 shows examples of blocks with all socially desirable and all socially undesirable items. Empirically, faking was reduced with an MFC format, resulting in mean increases of only .06 SD on trait scores in a meta-analysis (Cao & Drasgow, 2019).

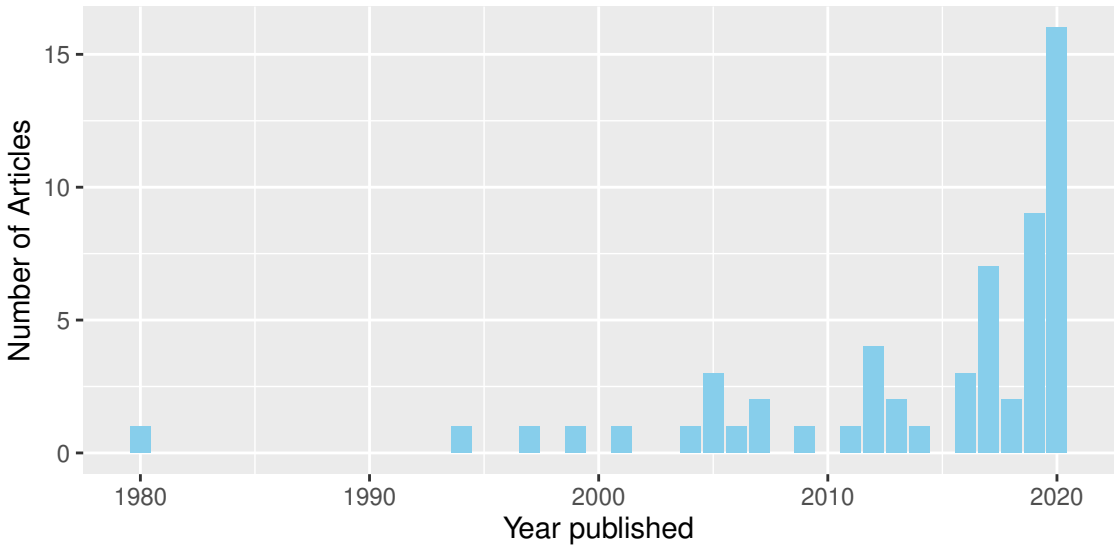

FIGURE 2: Number of new articles published in journals listed in the Web of Science Core Collection including the keywords "multidimensional" and "forced-choice" in any entry.

To address the issue of validity more directly, it is important to compare how well MFC and rating scale formats perform at predicting external constructs and criteria. Overall, similar (Lee et al., 2018; Wetzel & Frick, 2020; Zhang et al., 2019) or higher (Bartram, 2007; Salgado & Táuriz, 2014; Watrin et al., 2019) construct and criterion validities were observed with an MFC as compared to a rating scale format. Differences in validities probably depend on how the MFC responses were scored and on the type of criteria investigated (Wetzel et al., 2020). Moreover, the assessed constructs might slightly differ between the response formats: When the same items were presented in an MFC versus a rating scale format, correlations between traits slightly changed (Guenole et al., 2018; Wetzel & Frick, 2020). This could be explained by item interactions that occur in the MFC format: Item properties can change when items are presented together in blocks (Lin & Brown, 2017).

1.2 Challenges in the Construction of Multidimensional Forced-Choice Tests

Constructing MFC tests is a more complex endeavor than constructing rating scale tests, because the items must be combined into blocks. To give an example, it is usually preferable to have the same number of items per trait so that reliability is comparable. In a test measuring five traits with block size three, there are $\binom{5}{3} = 10$ possible combinations of traits. If we increase the number of traits to 15, this yields $\binom{15}{3} = 455$ combinations. How

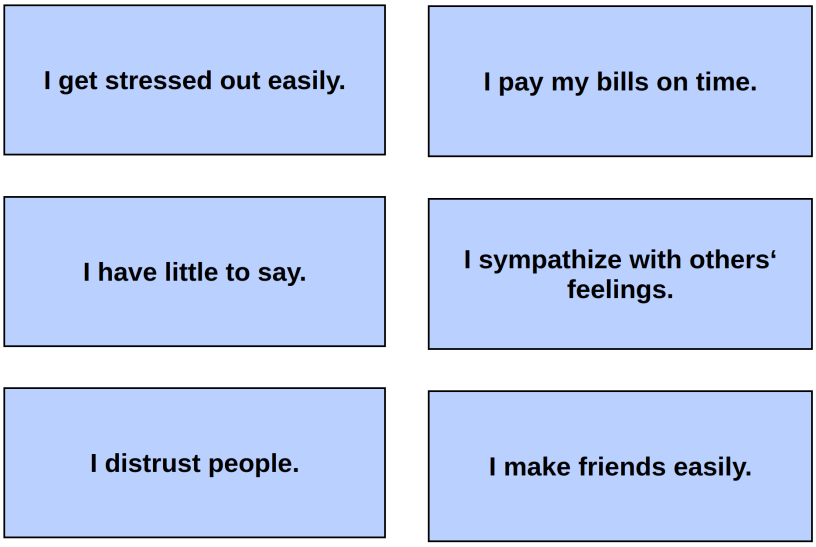

FIGURE 3: Examples of socially undesirable (left) and socially desirable (right) multidimensional forced-choice blocks from the Big Five Triplets (Wetzel & Frick, 2020).

the items are combined affects the properties of the test, both in terms of measurement and response behaviors. In the following, I outline three important aspects of MFC test construction that motivated the present research.

Normativity

When trait scores can be compared between different persons they are called normative. The opposite of normative is ipsative. Ipsative scores arise when the sum of scores across different traits (or attributes) is constant across persons (Clemans, 1966). It follows mathematically from this property that correlations with and between ipsative scores and correlation-based analyses, such as factor analysis, are distorted (Clemans, 1966; Hicks, 1970). MFC tests scored with classical test theory (CTT) yield *fully* ipsative scores when all items within blocks are ranked (ranking instruction) and all items are keyed in the same direction. To illustrate, for blocks of size $B = 3$, respondents assign ranks 1 to 3 to the items, which sum to 6. Across K blocks and all traits, this results in a total sum score of $K \times 6$ for each respondent. MFC tests scored with CTT yield *partially* ipsative scores when items are keyed in different directions or when the instruction is to select only some items. With partially ipsative scores, there is some variance in the total score. However, they are said to retain characteristics of ipsative scores (Hicks, 1970).

Item response theory (IRT) models, however, allow deriving normative scores from MFC data (Brown, 2016; Brown & Maydeu-Olivares, 2011, 2013; McCloy et al., 2005). In IRT, normative scores can be derived when the scale origin for the latent traits is identified.

For this to be the case, the test design must meet certain conditions, which depend on the item type (Brown, 2016). There are two common item types in personality psychology: For dominance items, the preference for an item increases monotonically with increasing trait levels. This idea is expressed, for example, in a linear factor model. For ideal-point (or unfolding) items, the preference for an item is highest at one point of the trait continuum (the item location) and decreases with increasing distance from it. To identify the scale origin for MFC tests with dominance items, the matrix of factor loadings for pairwise comparisons must be full-rank. With ideal-point items, the general conditions have not been examined so far. In the special case of equal weights for all items (i.e., all items correlate with the trait to the same extent), the item locations must differ between blocks.

The results of simulation studies complement these theoretical conditions: With dominance items, trait scores showed ipsative properties and trait recovery was decreased when all items were keyed in the same direction, that is, when all factor loadings were positive (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2020). The same was found for ideal-point items with equal locations (Hontangas et al., 2015; Hontangas et al., 2016). Hence, MFC tests should be scored and constructed in such a way that normative trait scores can be derived.

Item Matching and Fakability

If the test should reduce faking, the items within blocks must be matched for desirability. When matching items, several issues should be considered: First, an estimate of item desirability is needed. Some researches use item intercepts or differences in item intercepts between honest responding and faking instructions for this (e.g., Lee et al., 2018; Ng et al., 2020). Others use ratings of item desirability (e.g., Heggestad et al., 2006; Jackson et al., 2000). Second, to combine items of equal desirability requires defining which differences in item desirability estimates are considered negligible. If the differences are too large, the blocks might still be fakable. A recent study showed that agreement on which rank order was desirable was higher with larger differences in item desirability (Hughes et al., 2021). Third, item desirability might differ between assessment contexts. For example, desirability ratings for agreeableness items differed between the scenarios of applying for a job as a manager versus as a nurse (Pauls & Crost, 2005). Fourth, item interactions can occur in the form of item desirability changing in the context of item blocks because the relative response format might trigger more fine-grained distinctions of desirability (Feldman & Corah, 1960; Hofstee, 1970).

Reliability

A further issue to consider when constructing MFC tests is reliability. With the same number of items, MFC tests are theoretically less reliable than rating scale tests. This

can be illustrated by recoding rankings into binary outcomes of pairwise comparisons (Table 1). As can be seen from Table 1, a block of size $B = 3$ is approximately equally informative as the same three items presented in a dichotomous true-false format. More generally, a block of size B yields $B(B-1)/2$ pairwise comparisons. In comparison, rating scales with C categories yield $C - 1$ pieces of information per item. Moreover, binary outcomes of pairwise comparisons involving the same item, e.g., between items 1 and 2 and between items 1 and 3, are locally dependent given the latent traits. Thus, for block sizes $B > 2$, information is slightly lower than it would be expected if the binary outcomes were independent (Brown & Maydeu-Olivares, 2011, 2018b; Yousfi, 2018). Hence, achieving sufficient levels of reliability is an important issue in MFC test construction.

Table 1: Example of recoding rankings into binary outcomes

Item	Content	Ranking	Comparison	Outcome
i_1	I am emotionally stable.	1	$i_1 > i_2$	1
i_2	I like to explore new things	3	$i_1 > i_3$	1
i_3	I am always prepared.	2	$i_2 > i_3$	0

Note. This is a sample block from the Big Five Triplets (Wetzel & Frick, 2020).

Beyond the specific aspects described, the preceding overview reveals some overarching issues that research on the MFC format should address: First, it is important to investigate which (item) properties actually matter for the resulting trait scores. Second, in order to account for potential item interactions, the block level should be modeled. And third, methods for the construction of MFC tests should be developed that allow all relevant aspects to be considered simultaneously. The three manuscripts in this thesis each incorporate one or more of these issues.

1.3 Item Response Models for Multidimensional Forced-Choice Tests

Following Brown (2016), IRT models for MFC tests can be classified according to three axes: (a) whether block sizes $B > 2$ can be modeled, (b) whether the model assumes a dominance or an ideal-point relationship between item and trait and (c) whether the decision model for choice behavior is based on the ideas of Thurstone (Thurstone, 1927, 1931) or Bradley and Terry (Bradley, 1953; Bradley & Terry, 1952). Thurstonian models imply a probit link function whereas Bradley-Terry models imply a logit link function. As to my knowledge, two additional models have been proposed since the work by Brown (2016): The multi-unidimensional pairwise preference two-parameter logistic model (MUPP-2PL, Morillo et al., 2016), which can be classified as a Bradley-Terry model for dominance items and block size $B = 2$ and the generalized graded unfolding model for ranks (GGUM-

RANK, Lee et al., 2019), which can be classified as a Bradley-Terry model for ideal-point items and any block size, with a ranking instruction.

The present research employs the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011), which is a Thurstonian model for dominance items and any block size, for two reasons: First, the Thurstonian IRT model is the most broadly applicable in terms of response formats and ranking instructions. Second, it is a model for dominance items which are currently most common in personality psychology (Brown & Maydeu-Olivares, 2010). Moreover, research interest in this model is currently high: Half of the 28 articles about this model were published in the past two years (2019 and 2020), as evidenced by a search for articles including the keywords "Thurstonian item response theory" or "Thurstonian IRT" in any entry published in journals listed in the Web of Science Core Collection after the introduction of the Thurstonian IRT model in 2011.

Thurstonian Item Response Model

In the Thurstonian IRT model, there is a latent value underlying each item response called *utility*. The utility t of item i for person j is a linear function of a latent trait η_j, weighted with an item loadings λ_i and having an intercept μ_i and an error term ε_{ij}:

$$t_{ij} = \mu_i + \lambda_i \eta_j + \varepsilon_{ij} \tag{1}$$

The latent traits are assumed to follow a multivariate normal distribution: $\mathbf{H} \sim N(\mathbf{M_H}, \mathbf{\Sigma_H})$. The errors follow independent normal distributions: $\boldsymbol{\varepsilon}_i \sim N(0, \psi_i)$. According to Thurstone's Law of Comparative Judgment (Thurstone, 1927, 1931), respondents rank the items within each block according to the magnitude of their utilities.

The Thurstonian IRT response probabilities are usually expressed for binary outcomes of pairwise comparisons (Table 1) instead of rank orders, which enabled model estimation in the first place (Maydeu-Olivares, 1999; Maydeu-Olivares & Brown, 2010). The response probability for outcome l comparing items i and m that measure traits c and d, respectively, can be expressed as:

$$P\left(y_{lj} = 1 | \eta_{cj}, \eta_{dj}\right) = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_{cj} - \lambda_m \eta_{dj}}{\sqrt{\psi_i^2 + \psi_m^2}}\right) \tag{2}$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function evaluated at x. Typically, instead of separate intercepts μ_i and μ_m for the items, a threshold $-\gamma_l$ for the outcome is estimated (i.e., the restriction $\gamma_l = \mu_i - \mu_m$ is not imposed).

Since binary outcomes of pairwise comparisons involving the same item are locally dependent given the latent traits, the same applies to the response probabilities in Equation 2. Consequently, if these response probabilities are multiplied, the likelihood of the response

pattern is overestimated for block size $B > 2$. Therefore, instead of using a likelihood-based approach, the item parameters and trait correlations are usually estimated using limited information methods and a two-step procedure. First, the tetrachoric correlations and thresholds for the binary outcomes are estimated. Second, the results from the first step are used as input to limited information methods such as unweighted or diagonally weighted least squares, accounting for error covariances of the outcomes. For a tutorial on how to estimate Thurstonian IRT models in Mplus (Muthén & Muthén, 1998–2017) using this procedure, see Brown and Maydeu-Olivares (2012). Trait scores are then estimated given the previously obtained item parameters and trait correlations in a maximum-likelihood approach, such as maximum a posteriori (MAP) or weighted likelihood estimation (WLE). Thus, for trait estimation, the local dependencies for block size $B > 2$ are neglected. This yields unbiased point estimates but underestimated standard errors and overestimated reliability (Brown & Maydeu-Olivares, 2011; Yousfi, 2018), although the extent of the reliability overestimation was deemed negligible (Brown & Maydeu-Olivares, 2011).

Alternatively, following Yousfi (2018), the response probability for the full rank order can be expressed by first sorting vectors of utilities $\boldsymbol{t}_k$ and of error variances $\boldsymbol{\psi}_k^2$ within each block k in descending order, according to the selected rank order r. For example, if the rank order 3-1-2 was selected by person j, we would sort the vector of utilities as $\boldsymbol{t}_{jk} = \begin{pmatrix} t_{3j} & t_{1j} & t_{2j} \end{pmatrix}'$. For estimation, differences between consecutive utilities are calculated. In the example, the area where $t_{3j} > t_{1j} > t_{2j}$ is equivalent to the area where $t_{3j} - t_{1j} > 0 \cap t_{1j} - t_{2j} > 0$. The differences between consecutive utilities are calculated with a comparison matrix $\boldsymbol{A}$. For example, if block size $B = 3$:

$$\boldsymbol{A}_{B=3} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{3}$$

Then, the probability to select rank order r is the area under the multivariate normal density where each difference between two consecutive utilities $\boldsymbol{A}\boldsymbol{t}_{jk}$ is positive:

$$P(X_{jk} = r) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty N\left(\boldsymbol{A}\boldsymbol{t}_{jk}(r), \boldsymbol{A}\boldsymbol{\psi}_k^2(r)\right) d\boldsymbol{A}\boldsymbol{t}_{jk}(r) \tag{4}$$

The multiple integral in Equation 4 can be numerically approximated with methods developed by Genz (2004) and Genz and Bretz (2002). For equivalent variants of expressing the response probability, see Maydeu-Olivares (1999). To compute Equation 4 from estimated item parameters, the item intercepts have to be estimated or the restriction on the thresholds for the binary outcomes must be imposed.

To illustrate the effect of neglecting local dependencies, I conducted a small simulation on standard error accuracy for block size $B = 4$, because the effect of local dependencies increases with block size. Traits and their observed standard errors were estimated

based on the formulation neglecting local dependencies (Equation 2) and the true response probability (Equation 4). The test design was identical to the condition with block size $B = 4$, five traits and 1/2 mixed keyed comparisons in Frick et al. (2021). Besides that, the simulation design was identical to simulation study 1 on standard error accuracy in Frick (2021a) for the condition with high loadings and the short test. Figure 4 shows that when neglecting local dependencies, standard errors were underestimated both for the maximum likelihood (ML) and the MAP estimator. The bias was smaller for extreme trait levels and it showed high variance for the ML estimator in these areas. This might have occurred because the estimation procedure and the box constraints were not optimized for the formulation neglecting local dependencies.

In comparison to the scale of the latent traits ($SD = 1$) and the range of true SEs (Figure 5), the bias of observed SEs was small but not negligible. As expected, the bias of the point estimates of the latent traits was comparable between the true likelihood and the one neglecting local dependencies (Figure 5). When neglecting local dependencies, it was slightly higher for the MAP estimator, because the likelihood is given too much weight in relation to the prior (Yousfi, 2020).

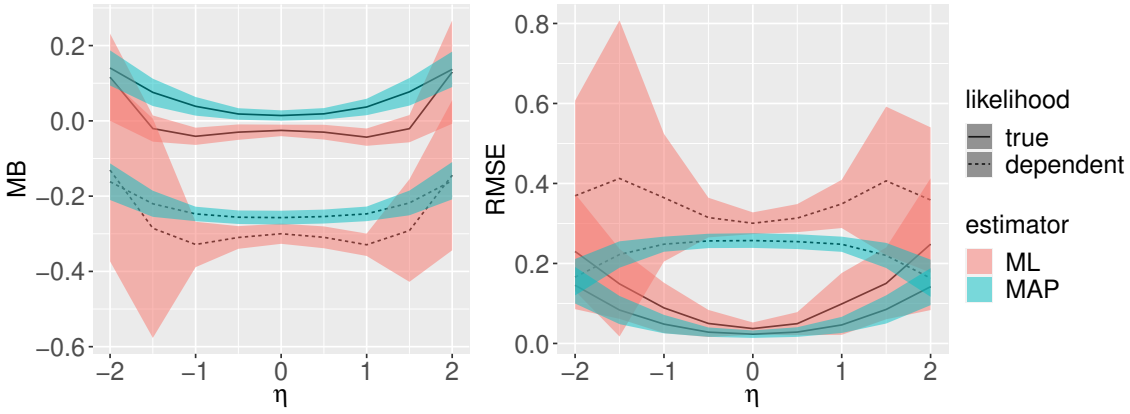

FIGURE 4: Bias of observed standard errors in the simulation on local dependencies. Shaded areas show $\pm 1SD$ around the mean (line). MB = Mean Bias, RMSE = Root Mean Square Error, true = true likelihood, dependent = likelihood neglecting local depencencies, ML = Maximum Likelihood, MAP = Maximum a Posteriori.

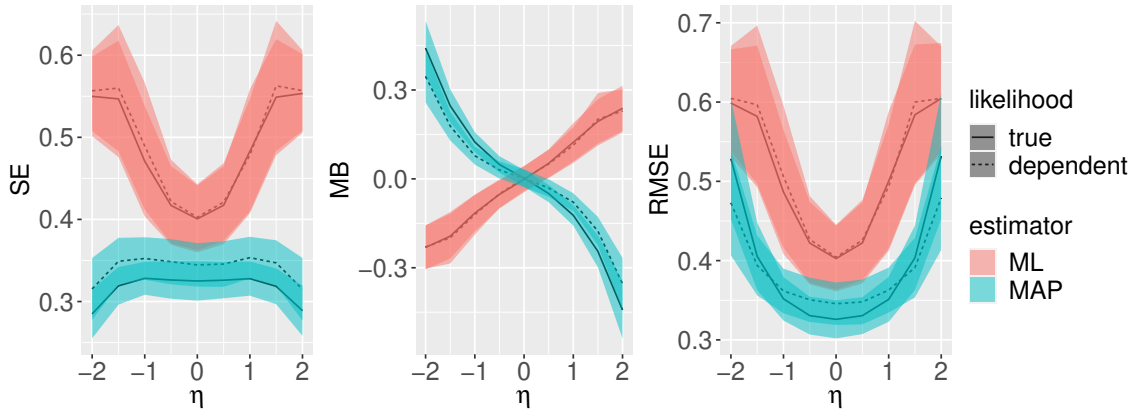

FIGURE 5: Trait recovery and empirical *SE*s in the simulation on local dependencies. Shaded areas show $\pm 1SD$ around the mean (line). SE = empirical Standard Error, MB = Mean Bias, RMSE = Root Mean Square Error, true = true likelihood, dependent = likelihood neglecting local depencencies, ML = Maximum Likelihood, MAP = Maximum a Posteriori.

1.4 Overview of Manuscripts

The present research addresses challenges in MFC test construction by investigating and developing IRT methods for this response format, focusing on the issues of normativity, fakability, and information. Although I used the Thurstonian IRT model throughout the three manuscripts, some findings transfer to and some methods could be applied to other IRT models for MFC tests as well. In this synopsis, I highlight where this is the case.

Since the theoretically derived conditions for normativity differ from the results of simulation studies, in the first manuscript (Frick et al., 2021), we conducted an extensive simulation study on this issue. We investigated the interplay of various test design factors with normativity, eliminated bias in item parameters as a potential confound, and compared Thurstonian IRT trait recovery to that from CTT scoring and from rating scale and true-false formats. The empirical counterpart of normativity/ipsativity is the relative response process. Therefore, the simulation study was complemented with an empirical study investigating the effect of a relative versus an absolute response process on validity while controlling for reliability.

In light of item interactions within blocks and the variety of methods to assess item desirability and to match items, in the second manuscript (Frick, 2021b), I developed a mixture IRT model that allows to assess fakability on the block level—the Faking Mixture model. As a post-hoc method, this model accounts for item interactions and is a useful complement to a priori methods of matching. The model results can be used to remove

or modify blocks so that the fakability of the whole test is reduced. Moreover, to my knowledge, this is the first IRT model for the MFC format that can capture response processes in addition to those triggered by the content trait.

Given that reliability with an MFC format is usually lower than with conventional rating scales, it is essential to construct MFC test in a way that maximizes reliability/information. So far, information in Thurstonian IRT models was calculated for binary outcomes which comes with empirical, practical and statistical disadvantages. Therefore, in the third manuscript (Frick, 2021a), I proposed methods to estimate and summarize Fisher information on the block level (block information) and investigated their performance in three simulation studies. Moreover, I combined algorithms for automated block selection with information summaries from the optimal design literature. These algorithms allow to automatically assemble MFC tests with maximum reliability while considering restrictions on test design such as item keying or trait balancing.

2 Investigating the Normativity of Trait Estimates from Multidimensional Forced-Choice Data

Frick, S., Brown, A. & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2021.1938960

2.1 Simulation Study

Motivation

The first aim of the simulation study was to examine Thurstonian IRT trait recovery under realistic conditions. An ideal MFC test would have the same number of items per trait. Item keys would be structured such that at least half of the pairwise comparisons across the test would be between items keyed in different directions. Previous simulation studies examined these ideal designs and, in addition, all positively keyed items (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2020). However, ideal test designs might not be representative of existing tests. For example, the Big Five Triplets (Wetzel & Frick, 2020) are an MFC test with 20 blocks and block size $B = 3$ measuring the Big Five traits. All blocks are matched for desirability. However, item matching resulted in unbalanced numbers of items per trait: There are 16 neuroticism, 13 extraversion, ten openness, seven agreeableness, and 14 conscientiousness items. All blocks except one contain at least one negatively keyed item. However, when neuroticism is defined in the opposite direction, as emotional stability, the item keys obviously change. Then, there are only four blocks containing a negatively keyed item. From previous simulation studies, it is unclear to what extent deviations from ideal test designs affect trait recovery.

The second aim of the simulation study was to investigate Thurstonian IRT trait recovery with unbiased item parameters. Previous studies reported convergence issues when all items in the test were positively keyed (Brown et al., 2017; Bürkner et al., 2019; Guenole et al., 2018). Although it is possible that the matrix of factor loadings for pairwise comparisons is of full rank with all positively keyed items, empirical underidentification might still

occur. Empirical underidentification can lead to bias in item parameters which propagates to the trait scores. Therefore, we examined trait recovery with item parameters fixed to their true values.

The third aim of the simulation study was to compare Thurstonian IRT trait recovery to that from (partially) ipsative CTT scoring, from rating scales and from true-false data. Previous comparisons between those scoring methods and response formats used empirical data (Brown & Maydeu-Olivares, 2013) or did not include single-stimulus formats (e.g., rating scale or true-false formats; Hontangas et al., 2015; Hontangas et al., 2016). We kept the amount of information across MFC block sizes approximately equal to the true-false version. To accomplish this, the number of pairwise comparisons over the test was kept equal for different block sizes, while in turn the number of items varied. In this way, we could investigate the effect of local dependencies because any reliability differences between block sizes would be attributable to local dependencies.

Methods

In the simulation study, the following factors were varied and completely crossed: Number of traits, trait correlations, item keying, number of items per trait, and block size. MAP estimates for the latent traits were obtained based on the true item parameters and with the true trait correlations as prior covariances. Trait recovery was evaluated for single traits and for sums and differences of two traits each. Further, bias in mean correlations was calculated. The bias in mean correlations can be regarded as an indicator to ipsativity (Hicks, 1970).

Results

Figure 6 shows the correlation between true and estimated traits, averaged across traits, block sizes and numbers of items per trait. Regarding test design, positively keyed items were found to be detrimental to trait recovery, as, for example, evidenced by lower correlations between true and estimated traits in Figure 6. With positively keyed items, trait recovery was lower for five as compared to 15 traits and for positive as compared to mixed positive and negative trait correlations or uncorrelated traits (Figure 6). The other factors of test design, namely, unequal numbers of items per trait, varying levels of item keying and block size, had negligible effects on trait recovery. The mean trait correlation was negatively biased, indicating ipsativity (Clemans, 1966; Hicks, 1970) due to the condition with all positively keyed items. Similarly, the recovery of sums of traits (i.e., absolute trait levels) was affected by item keying, but not that of differences of traits (i.e., relative trait levels). Thus, the lower recovery with all positively keyed items could be attributed to ipsativity. Reliability was comparable to the true-false format, but lower than that of rating scales (Figure 6), as to be expected by the amount of information. With CTT scoring

of MFC responses, recovery was markedly worse and ipsativity was present in all conditions besides the one with uncorrelated traits and half of pairwise comparisons between differently keyed items, which was ideal for CTT scoring.

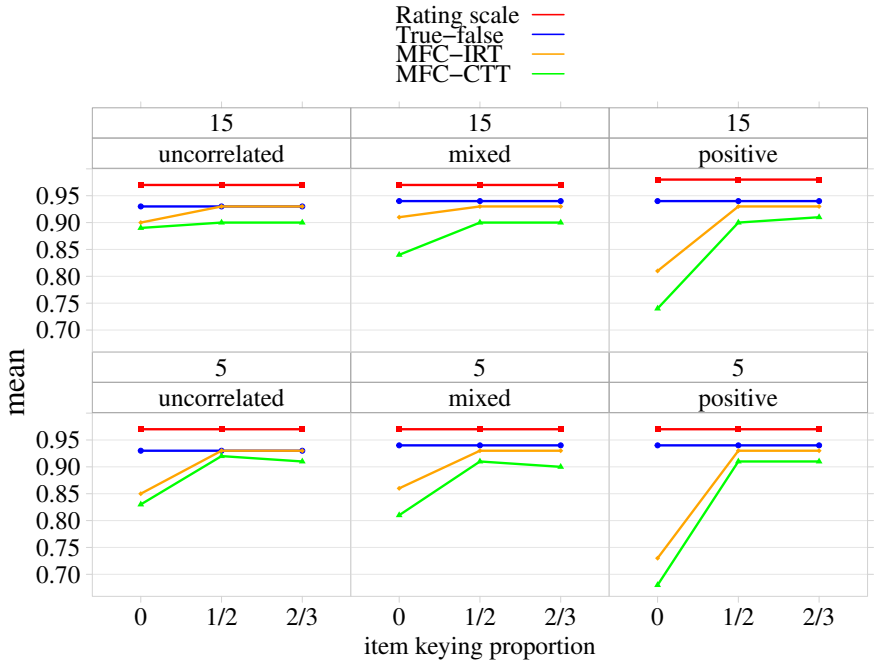

FIGURE 6: Mean correlation between true and estimated traits (i.e., $r(\eta, \hat{\eta})$) by condition. The results were averaged across traits, across block sizes two to four and across equal and unequal numbers of items per trait. MFC = multidimensional forced-choice format; IRT = item response theory scoring, CTT = classical test theory scoring, mixed = mixed positive and negative trait correlations, positive = all positive trait correlations, 5 = 5 traits, 15 = 15 traits.

2.2 Empirical Study

Motivation

The empirical study compared construct and criterion validity between the MFC format with block size three and the true-false format. The true-false format was chosen as a comparison because the amount of information is comparable to an MFC format with block size three (see also Table 1). Moreover, the true-false format is free from response styles arising from the use of rating scales such as midpoint and extreme responding. Assuming that a relative response process leads to higher differentiation between behaviors (Kahnemann, 2011), we expected validities to be higher in the relative (MFC) than in the absolute (true-false) response format.

Methods

$N = 999$ respondents filled out both an MFC and a true-false version of the Big Five Triplets (Wetzel & Frick, 2020), with an interval of two weeks in between and in counterbalanced order. Further, they answered questions on criterion variables focusing on the areas of employment (e.g., ability to supervise people at work; yes/no), social (e.g., having Facebook; yes/no), health (e.g., exercising regularly (at least once a week); yes/no) and relationships (e.g., being married; yes/no). Further, the constructs quality of life, satisfaction with life and depression/mental health were assessed with the World Health Organization Quality of Life BREF (WHOQOL group, 1996, WHOQOL-BREF,), the Satisfaction with Life Scale (SWLS; Diener et al., 1985) and the Center for Epidemiologic Studies-Depression Scale short form (SWLS; Cole et al., 2004), respectively. Based on meta-analyses and studies with large samples, we formulated and preregistered which Big Five traits and constructs/criteria were expected to correlate and only tested for differences in these correlations between MFC and true-false. Each construct (modeled with a graded response model; Samejima, 1969) and each criterion was regressed on the Big Five latent traits, separately for the MFC (modeled with the Thurstonian IRT model) and the true-false version (modeled with the two-parameter normal ogive model).

Results and Discussion

Figure 7 shows correlations with the constructs and with exemplary criteria. For all constructs, the differences in correlations between MFC and true-false were small to medium and in favor of true-false. For the criteria, all differences in correlations were negligible, besides one statistically insignificant difference in favor of MFC. Thus, our expectation of higher differentiation in the MFC format leading to higher validity was not confirmed. Possible explanations for this include: Method biases common to absolute response formats, such as acquiescence, might have increased the correlations between the Big Five traits assessed with the true-false format and constructs assessed with rating scales. Moreover, it is unclear which criteria actually value differentiation, because previous research was done with absolute response formats that allow to compensate for low levels on one trait with high levels on another trait. Last, the MFC format might not always trigger deeper retrieval. For example, in a recent think-aloud study, sometimes the response process could be sufficiently described by absolute evaluations of the items (Sass et al., 2020).

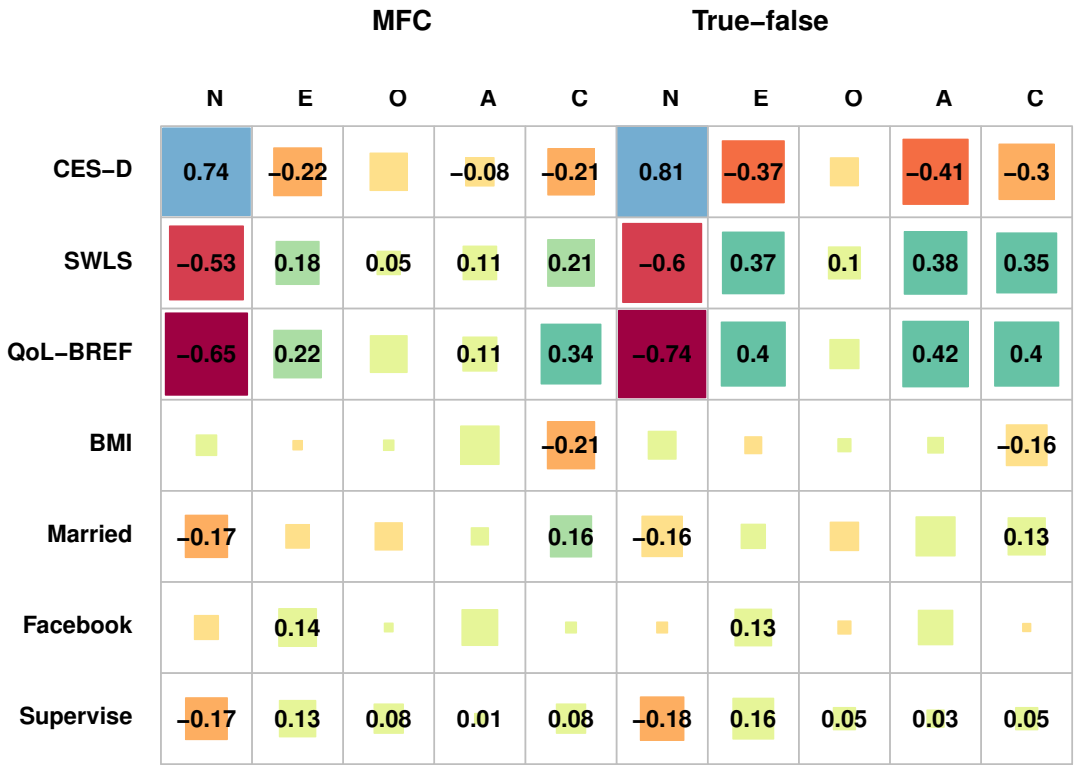

FIGURE 7: Correlations between the Big Five Triplets in the multidimensional forced-choice (MFC) and true-false (TF) version. The size of the square indicates the magnitude of the correlation. Positive correlations are depicted in green and blue and negative correlations in orange and red. Only correlation coefficients for correlations that were predicted are shown. N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness, CES-D short form = Center for Epidemiologic Studies-Depression Scale, SWLS = Satisfaction with Life Scale, WHO-QoL BREF = World Health Organization Quality of Life BREF, BMI = Body Mass Index.

3 Modeling Faking in the Multidimensional Forced-Choice Format - The Faking Mixture Model

Frick, S. (2021). Modeling faking in the multidimensional forced-choice format – The Faking Mixture model. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-021-09818-6

3.1 Motivation

In this manuscript, I introduced the Faking Mixture model, an IRT model for faking in MFC tests. Previous modeling approaches are limited in their usefulness for the MFC format or they cannot be applied to it. First, previous modeling approaches for faking in MFC tests focus on changes in trait scores, on the test level (e.g., Pavlov et al., 2019; Wetzel et al., 2021). The Faking Mixture model is the first one that allows to estimate the fakability of individual MFC blocks. Hence, its results can inform modifications of the test, such as removing items or blocks, with the aim of reducing fakability. Second, to apply the IRT models currently available for faking or socially desirable responding in rating scales (Böckenholt, 2014; Leng et al., 2019), it is necessary to know a priori which response options are desirable. In the MFC format, response options are rank orders. However, responses to MFC blocks are needed in order to know which rank orders are more desirable, because the relative response process might change evaluations of item desirability (Feldman & Corah, 1960; Hofstee, 1970). By modeling responses on the block level, the Faking Mixture model can capture such item interactions. Moreover, the Faking Mixture model reflects assumptions and empirical findings about the process of faking, some of which are specific to the MFC format. This will be outlined in the following part.

3.2 Model Properties

Respondents do not necessarily fake all items (MacCann et al., 2011). But when they fake they might not even consider their content traits (Robie et al., 2007). This is captured in the Faking Mixture model by conceptualizing responses in a high-stakes situation as a

mixture of responses based on the content trait and faked responses (Figure 8).

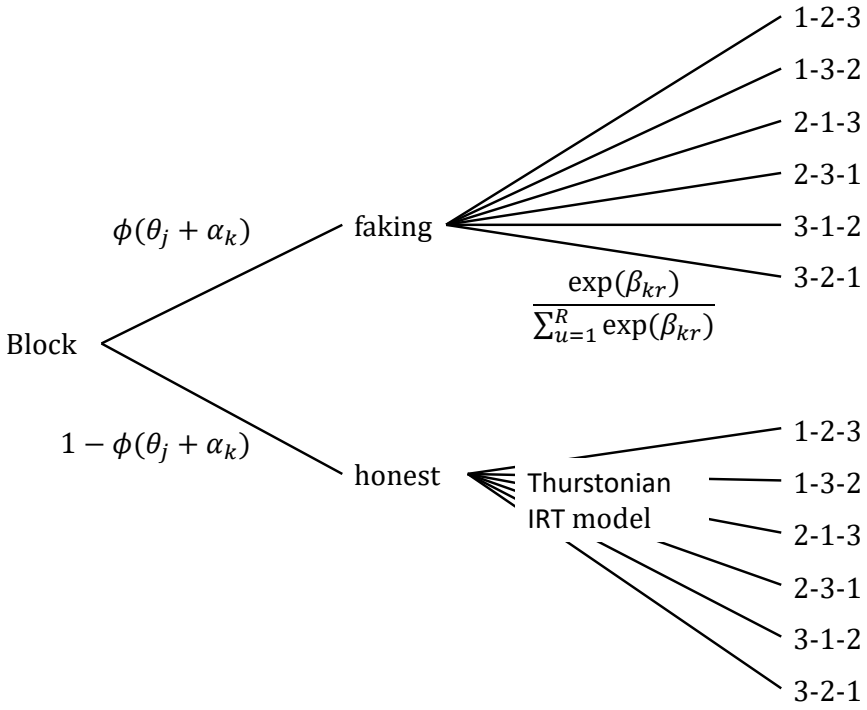

FIGURE 8: The Faking Mixture model depicted as a multinomial processing tree model.

For each person j and each block k, there is a probability to fake on this block $P(F_{jk} = 1)$ or to respond based on the content traits $P(F_{jk} = 0)$. In both cases, faking ($F_{jk} = 1$) or responding based on the content traits ($F_{jk} = 0$), there is a probability for each rank order r to be selected. Thus, the probability of observing rank order r for person j on block k is the sum of these two response probabilities:

$$P(X_{jk} = r) = P(F_{jk} = 1)P(X_k = r|F_{jk} = 1) + P(F_{jk} = 0)P(X_{jk} = r|F_{jk} = 0) \quad (5)$$

Not all respondents fake when they are in a high-stakes situation (MacCann et al., 2011). But a respondent highly motivated to fake might even do so on closely-matched blocks. To capture this in the Faking Mixture model, a faking tendency θ_j is introduced. The probability of faking a block increases both with the person's faking tendency θ_j and the block fakability α_k:

$$P(F_{jk} = 1) = \Phi\left(\theta_j + \alpha_k\right) \quad (6)$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function, evaluated at x, and Φ^{-1} its inverse.

The probability to select a rank order when faking $P(X_k = r|F_{jk} = 1)$, called rank

order probability, is modeled by rank order parameters β_{kr} via the softmax function (like a multinomial IRT model without a person parameter):

$$P(X_k = r|F_{jk} = 1) = \frac{\exp(\beta_{kr})}{\sum_{u=1}^{R} \exp(\beta_{ku})} \tag{7}$$

The rank order probabilities are constant across persons in order to reflect item desirabilities, which depend on the situation but not on the person. (Therefore, the person subscript j is dropped.) More precisely, the rank order probabilities reflect *differences* in item desirabilities because they are not linked to the individual items. Further, they are not related within traits. This facilitates the estimation of the rank order parameters while at the same time being flexible to account for differential desirabilities of the items and traits in the context of item blocks.

The block fakability α_k is obtained from the sum of squares of the rank order probabilities across all $R = B!$ rank orders:

$$\alpha_k = \Phi^{-1}\left(\sum_{r=1}^{R} \left(P(X_k = r|F_{jk} = 1) - M\left[\boldsymbol{P(X_k|F_{jk} = 1)}\right]\right)^2\right) \tag{8}$$

Thus, the more respondents agree about which rank order to prefer when faking, the more likely they are to fake on this block. This captures the idea underlying matching in the MFC format, namely, that respondents are more likely to base their response on their own content trait levels when items are closely-matched and vice versa (Berkshire, 1958; Gordon, 1951).

The response probabilities when responding honestly $P(X_{jk} = r|F_{jk} = 0)$ follow the Thurstonian IRT model as formulated in Equation 4. Currently, there is no computer software availabe that can estimate both the Thurstonian IRT model for rank orders and the within-block mixture of the Faking Mixture model at once. Therefore, the response probabilities when responding honestly are estimated with low-stakes data from the same respondents and treated as fixed in the estimation of the Faking Mixture model. The parameters of the Faking Mixture model are estimated in a Bayesian modeling framework (for details, see Frick, 2021b). Note, that the Faking Mixture model is theoretically not limited to the Thurstonian IRT model or to the MFC format; the response probabilities when responding honestly could potentially follow any other IRT model.

3.3 Simulation on Parameter Recovery

I conducted a simulation study to examine how well the parameters of the Faking Mixture model could be recovered. The simulation study investigated possible conditions from minimum to extreme faking and fakability, varying the faking trait mean and variance,

and the variance of the rank order parameters (i.e., the mean fakability across the test). The results showed that the parameters were generally well recovered. Both the faking trait θ_j and the rank order parameters β_{kr} were recovered best when they had a high variance. In addition, the faking trait θ_j was recovered better when its mean was medium, so that there were no floor or ceiling effects. The rank order parameters β_{kr} were recovered better when the faking trait mean was high, because this allowed to observe more instances of faking.

3.4 Empirical Validation

For the empirical validation, I re-analyzed a dataset from Wetzel et al. (2021). In this dataset, $N = 1244$ respondents were randomly assigned to either the original version of the Big Five Triplets (Wetzel & Frick, 2020), which is matched for social desirability, or a version in which one item in seven triplets was replaced by a clearly more desirable one. I fitted the Faking Mixture model to (a) the matched version and (b) to both versions allowing the rank order parameters for the different items to differ between groups and estimating differences in the block fakability parameters α_k.

Applying the Faking Mixture model to the matched version showed that the blocks had intermediate to high fakability. Figure 9 shows the rank order probabilities for two exemplary blocks. In the matched version, for Block 3, it was *undesirable* to rank the item "I am often sad" first, whereas the preferences for the other four rank orders were approximately equal. For Block 5, ranking the item "I love big parties" last was *desirable*, so that the probabilities were high only for the two rank orders where this was the case. Therefore, Block 5 was more fakable than Block 3. Comparing the results for the mixed and the matched version showed that the mixed blocks were more fakable than the matched blocks (in all seven cases). Moreover, the clearly more desirable items were preferred when faking. For example, when replacing "I act without thinking" with "I treat my belongings with care", the probabilities for rank orders in which this item was ranked first increased (Figure 9). Indeed, for all mixed blocks, the rank order probabilities were different from zero only for two or three rank orders, always including the ones in which the highly desirable item was ranked first.

Thus, this re-analysis validated the Faking Mixture model by showing that mixed blocks were more fakable than matched blocks and that more desirable items were preferred in mixed blocks. Moreover, it showed that matching alone was not sufficient, because even the matched blocks were still fakable. Probably, item desirability was evaluated differently in the context of item blocks. Hence, the Faking Mixture model is worth using, because it can capture such item interactions.

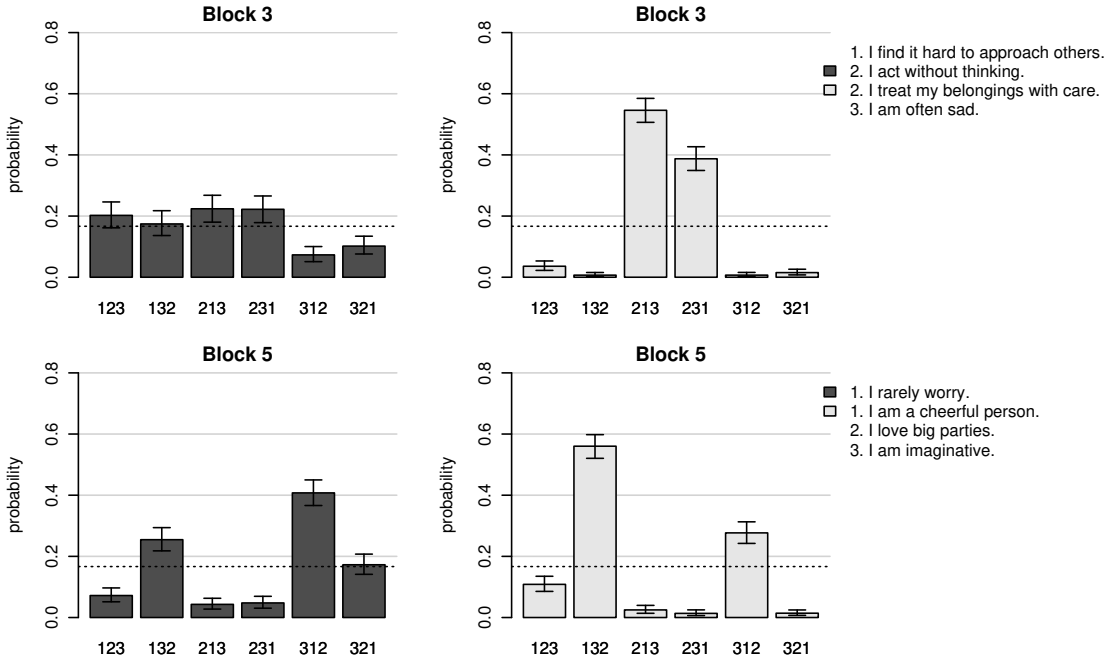

FIGURE 9: Probabilities for rank orders when faking in MFC-matched versus MFC-mixedfor two selected blocks. The dotted line indicates where all rank orders are equally probable. Results for MFC-matched are depicted in dark-grey on the left side, for MFC-mixed in light-grey on the right side.

4 Block Information in the Thurstonian Item Response Model

Frick, S. (2021). Block information in the Thurstonian item response model. *Manuscript submitted for publication to Psychometrika.*

4.1 Motivation

Currently, information in the Thurstonian IRT model is calculated for binary outcomes of pairwise comparisons (Brown & Maydeu-Olivares, 2011, 2018b). This procedure has several disadvantages: First, possible item interactions are not fully accounted for. Indeed, some authors reported that item properties differed depending on which items were combined to blocks (Lin & Brown, 2017; Wetzel & Frick, 2020). Second, for a test constructor, it is unclear which item to select if the item properties differ depending on which items are compared. Third, the information for binary outcomes of pairwise comparisons is locally dependent for block sizes $B > 2$ (Brown & Maydeu-Olivares, 2011, 2018a). Thus, test information and estimates of standard errors and reliability based on pairwise comparisons are biased. Therefore, I argue that information should be computed on the block level instead (henceforth called block information).

Yousfi (2018) formulated the response probability on the block level (Equation 4) and proposed to estimate it via numerical integration (using methods developed by Genz, 2004; Genz & Bretz, 2002). He investigated how this formulation can be used to estimate the person parameters without local dependencies and showed that it yields unbiased Fisher information on the test level (Yousfi, 2020). However, to my knowledge, this procedure was not used to compute Fisher information on the block level so far.

Fisher information for a block and a single rank order r is obtained as the negative of the Hessian of the response probability $P(X_{jk} = r)$ in Equation 4:

$$\mathbf{I}_{kr} = -\mathbf{H}\left(P(X_{jk} = r)\right) \tag{9}$$

where $H(f)$ denotes the Hessian of function f. Expected block information $\mathbf{I}_k$ is obtained

by weighting with the probability for all $R = B!$ possible rank orders:

$$\mathbf{I}_k = \sum_{r=1}^{R} \mathbf{I}_{kr} P(X_{jk} = r) \tag{10}$$

Block information in Thurstonian IRT models comes with several challenges: First, there is no closed-form expression for it, so that numerical approximation must be used, both for the response probability (Equation 4) and for its hessian (Equation 9). Second, because the Thurstonian IRT model is only identified with multiple blocks (Brown, 2016), block information is not invertible. Third, block information is a matrix, because in an MFC test, each block measures multiple traits. Information in matrix form again presents a challenge for test constructors.

To address these challenges, first, the accuracy of the estimation procedure was evaluated in several simulation studies. Second, information summaries were proposed that transform the block information matrix into a scalar or vector. Third, I examined how these information summaries can be used for automated test assembly (ATA). In ATA, items or blocks are selected from a pool to maximize some criterion (in this case, information) and to simultaneously fulfill certain restrictions on test design, such as test length, item keying, trait balancing, or fakability (for an introduction to ATA, see van der Linden, 2005). Thus, ATA can be used to integrate the diverse aspects of MFC test construction investigated in the three manuscripts of this thesis. Several types of algorithms are available for ATA. Therefore, I explained which information summaries and algorithms can be combined. Last, in two ATA simulations, it was investigated how the information summaries perform in test assembly. For this purpose, each information summary was combined with an exemplary algorithm and their performance was compared.

4.2 Block Information Summaries

The first information summary proposed was called block R^2. Block R^2 is computed from the sampling variances of traits based on the test (or pool) including this block σ^2_T and excluding this block $\sigma^2_{T\setminus k}$:

$$\mathbf{R}^2_k = 1 - \frac{\boldsymbol{\sigma}^2_T}{\boldsymbol{\sigma}^2_{T\setminus k}} \tag{11}$$

Thus, block R^2 summarizes block information on the level of traits, in the familiar R^2 metric, and relative to the set of reference blocks T. Figure 10 shows an example of how block R^2 varies across trait levels for a block from a simulated test measuring five traits with 20 blocks of size $B = 3$.

The other information summaries proposed are so-called optimality criteria originating

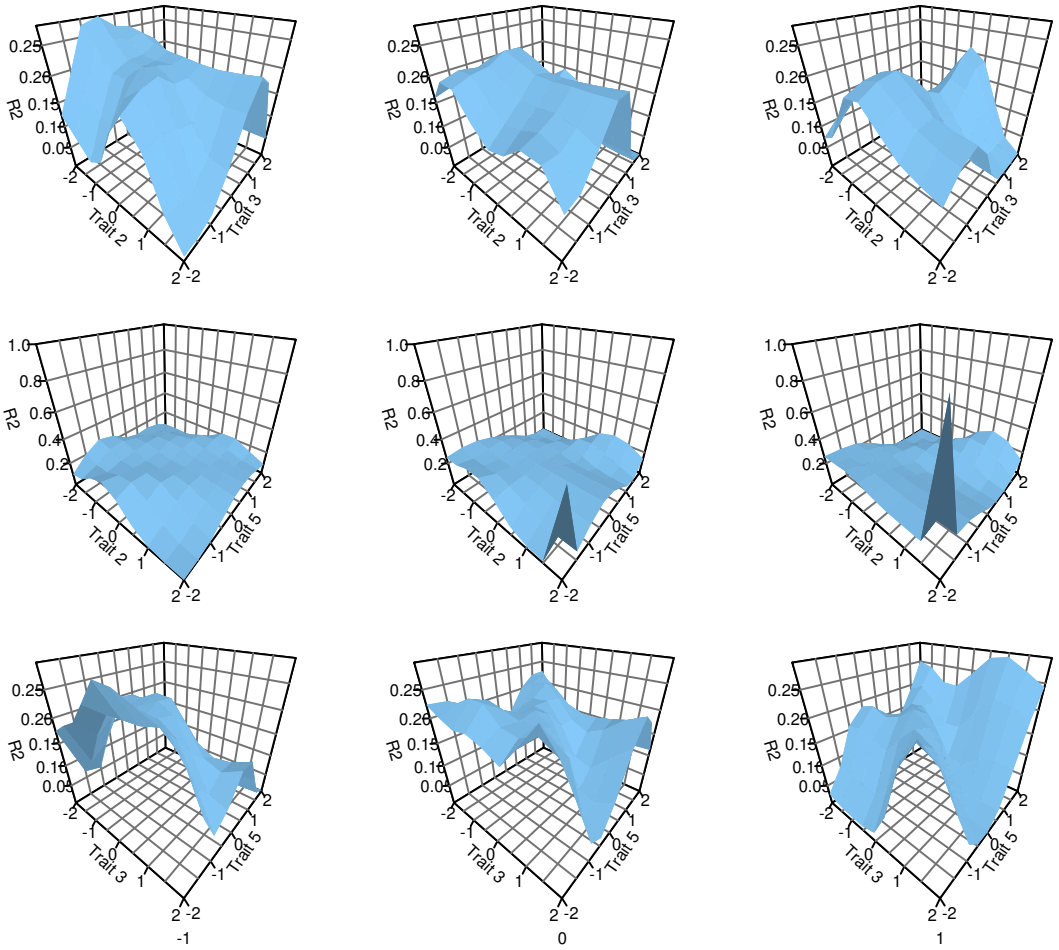

FIGURE 10: Block R^2 for Trait 5 from a simulated test block. Items 1-3 measured traits 2, 3, and 5, respectively. The simulated item parameters were $\mu_1 \approx 0.73$, $\mu_2 \approx -0.89$, $\mu_3 \approx -0.62$, $\lambda_1 \approx 0.92$, $\lambda_2 \approx -0.90$, and $\lambda_3 \approx 0.94$.

from the optimal design literature. They summarize an information matrix into a scalar, that is, for block information, across traits. Optimality criteria have been used for ATA and for computerized adaptive testing. For example, Debeer et al. (2020) investigated how well linear approximations to A- and D-optimality perform in multidimensional ATA. A- and D-optimality performed best in a simulation of computerized adaptive testing in which items were adaptively combined to MFC blocks of size $B = 2$ (Lin, 2020). Therefore, A- and D-optimality were also proposed to be used as block information summaries in this manuscript. A-optimality is the sum of the sampling variances (i.e., the trace of the inverse of the information matrix) and D-optimality is the determinant of the information matrix. To calculate A- and D-optimality, the information matrix must be invertible. As previously explained, for the Thurstonian IRT model, this is the case only for multiple blocks (i.e., for test information).

If an ATA problem can be framed as a (constrained) linear optimization problem, the optimal solution can be found by mixed integer programming (MIP; Debeer et al., 2020; van der Linden, 2005). A- and D-optimality are not linear (additive) across blocks and therefore cannot be used in MIP algorithms, but T-optimality can. For this reason, I additionally proposed to use T-optimality as a block information summary, although it performed worst in the computerized adaptive testing simulation by Lin (2020). T-optimality is the trace of the information matrix. Thus, it can be computed on a non-invertible matrix, but it is not affected by trait correlations.

4.3 Block Information for Test Construction - Simulation Studies

The first simulation study examined the accuracy of standard errors (*SE*s). Three types of *SE*s were computed: Empirical *SE*s served as true *SE*s. Empirical *SE*s were defined as *SD*s of MAP estimates across responses for the same trait levels (persons). Expected and observed *SE*s were based on Fisher information. To compute expected *SE*s, the Hessian for each rank order was weighted by its probability (Equation 10). To compute observed *SE*s, the Hessian was calculated only for the observed rank orders (Equation 9). Across blocks, this is equivalent to the Hessian at the likelihood of the trait estimate. Both ML and MAP estimates were obtained. Additionally, the size of factor loadings and test length were varied. The results showed that empirical *SE*s were smaller for the MAP estimator than for the ML estimator, especially with small loadings. However, this gain in accuracy was not detected by the information-based (expected and observed) *SE*s, i.e., they were overestimated for the MAP estimator with small loadings. Overall, expected and observed *SE*s were similarly accurate. Hence, if block-level information is not needed, researchers can obtain observed *SE*s directly with the trait estimate and save computational time and resources.

Since A- and D-optimality can only be computed for multiple blocks, I conducted two ATA simulations, one on test construction and one on test extension. When extending a test, information for multiple blocks is already available and therefore it is invertible. Note, however, that as few as three blocks were sufficient in the current simulations. In the simulation on test construction, the target information curve (flat vs. proportional to information in the pool) and restrictions (only test length vs. additional restrictions on trait balancing and item keying) were varied. The performance of T-optimality was compared to that of block R^2 and the mean of loadings within a block (mean loadings). Mean loadings represent the procedure of using the size of factor loadings as the main criterion for item or block selection. Block R^2 was averaged across traits to obtain a scalar. For the simulation on test extension, A- and D-optimality were added. Developing

a sophisticated algorithm for ATA with a non-linear optimization criterion would require a separate research project (e.g., Kreitchmann et al., 2021; Olaru et al., 2015). Therefore, A- and D-optimality were combined with a simple (so-called greedy) algorithm and the condition with more complex restrictions on test design was dropped. For details on the algorithms, see the main manuscript (Frick, 2021a).

The results of both ATA simulations showed that all criteria performed better than random block selection, but on par with each other. Therefore, the decision for an information summary and an ATA algorithm should be based on other aspects such as whether trait-level information is of interest or how accurately a target information surface should be approximated. In sum, the three simulation studies showed that and illustrated how block information can be used for test construction.

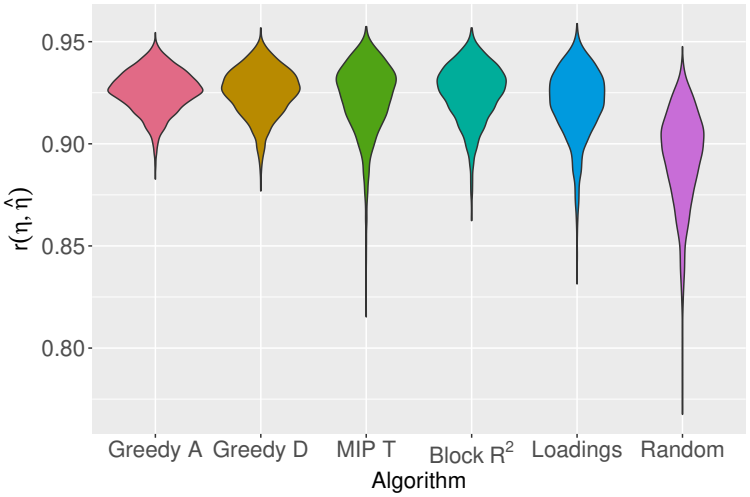

FIGURE 11: Correlation between true and estimated traits ($r(\eta, \hat{\eta})$) by algorithm in the simulation study on test extension for target information proportional to the block pool. A = A-optimality, D = D-optimality, T = T-optimality, MIP = Mixed Integer Programming.

5 General Discussion

In this cumulative thesis, I have developed and investigated IRT methods that can help to improve the construction of MFC tests. We investigated the effect of test design on the normativity of trait scores. We found that all positively keyed items were detrimental, but that suboptimal designs only affected trait recovery with all positively keyed items. I developed the Faking Mixture Model, which allows to assess the fakability of MFC blocks. An empirical application showed that it is useful to apply the Faking Mixture model in addition to matching, due to item interactions. Last, I investigated methods to estimate and summarize block information and showed how they can be used to automatically assemble MFC tests. I found that the estimation bias of expected and observed Fisher information was comparable and small, and that all proposed summaries can be used to construct MFC tests.

5.1 Recommendations and Methods for MFC Test Developers

According to the results of our simulation (Frick et al., 2021), it is recommended that MFC tests include at least some comparisons between items keyed in different directions. This is in accordance with other simulations that found that trait recovery decreased drastically with all positively keyed items (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2020). The exact proportion of items keyed in different directions is likely of minor importance, since it had a negligible effect in our simulation study. If all items are positively keyed, assessing a high number of traits and assessing traits that are uncorrelated or negatively correlated can yield better trait recovery. If the numbers of items per trait are unequal (unbalanced), this will naturally lead to smaller recovery for traits assessed with fewer items. However, we found no additional decrease in recovery due to the inseparable design of MFC tests. Moreover, keeping the amount of information equal, the decrease in precision due to local dependencies with block sizes larger than two was negligible. If the test should reduce faking, based on the application of the Faking Mixture model (Frick, 2021b), it is recommended to match items for desirability and in a second step to examine fakability of the resulting MFC blocks.

Several new methods were developed that can aid MFC test developers: The Faking Mixture model (Frick, 2021b) allows to estimate fakability on the block level, thereby

accounting for item interactions. Block information in the Thurstonian IRT model can be estimated and summarized (Frick, 2021a). Last, blocks can be selected automatically based on their information while simultaneously taking into account other restrictions on test design. Selecting blocks instead of items is more expensive in terms of respondent time. Future MFC test development will show in which cases this is necessary and practical.

5.2 Statistical Analysis of Simulation Studies

Throughout the simulation studies in this thesis, I conducted statistical analyses of the results to investigate which factors matter for the outcome of interest. For other examples of this technique, see Plieninger (2017) or Lin (2020). It has been advocated since quite some time (Harwell et al., 1996; Skrondal, 2000) that simulation results should be analyzed statistically instead of only visually by examining tables of means and variances across conditions.

Specifically, I summarized the simulation results in terms of variance explained by the main factors and by orthogonal contrasts within an ANOVA framework. For example, in the simulation study on normativity (Frick et al., 2021), I investigated how much variance in trait recovery was explained by the difference between all positively keyed items and various levels of mixed keyed items. Several properties of explained variance make it particularly suited for analyzing simulation studies: It is descriptive and therefore insensitive to sample size. In simulation studies, sample size (i.e., the number of replications) can be increased arbitrarily (up to the computational resources available). Further, in contrast to inferential tests for ANOVA results, explained variance is insensitive to heterogeneous variances across conditions, which can easily occur in simulation studies. For example, when test length is manipulated, trait recovery will show higher variance in conditions with shorter test lengths.

On the downside, explained variance yields only relative information about the comparison of conditions. Therefore, throughout the simulation studies, I additionally reported means and variances within conditions to evaluate the absolute level of recovery. Alternatively, for example, the number of replications can be planned a priori (Feinberg & Rubright, 2016) so that the design is not over-powered. Or, equivalence testing could be used to overcome the power problem by including effect sizes of interest in the testing procedure.

Another issues in the analysis of simulation studies is how to correctly analyze the results from Bayesian simulation studies (Boykin, 2020). In this thesis, only the second manuscript (Frick, 2021b) used a truly Bayesian estimation procedure. To summarize the simulation results, I used coverage rates, which carry the full distributional information, but also measures such as mean bias, which originate from a frequentist view. Using frequentist

statistics to summarize Bayesian simulation studies is not uncommon in psychometrics (e.g., Leng et al., 2019). However, it could be argued that one should be consistent in the use of inferential frameworks and analyze simulations of Bayesian models in a Bayesian way (Boykin, 2020).

5.3 About the Relative Nature of MFC Responses

The MFC format is a relative response format: In contrast to single-stimulus formats such as a rating scale or a true-false format, the response process for the MFC format involves relative comparisons between the items (Sass et al., 2020). In this thesis, the relative nature of MFC responses was observed and accounted for in several instances.

The relative response process can result in item interactions: Item properties from single-stimulus items do not necessarily translate to MFC blocks. Moreover, item properties might not even be invariant across different block compositions. For example, some authors observed that estimates of item parameters differed depending on which items were combined into blocks (Lin & Brown, 2017; Wetzel & Frick, 2020). By focusing on the block level, both the Faking Mixture model (Frick, 2021b) and block information (Frick, 2021a) allow to capture item interactions. The empirical validation of the Faking Mixture model (Frick, 2021b) contributes to evidence of item interactions: MFC blocks that were matched for social desirability were still fakable. Thus, in the context of MFC blocks, item desirability differed from that assessed through ratings of the individual items. Future research could compare block information (Frick, 2021a) between different block compositions or response instructions. This would allow to summarize all parameter differences on the block level and to illustrate at which trait levels (or combinations thereof) item interactions impact measurement precision.

Moreover, MFC test construction would benefit from being able to predict how items interact when combined into blocks. Lin and Brown (2017) discussed how item interactions could be predicted from the item content. In the context of faking and item matching, block fakability estimates obtained from the Faking Mixture model could be compared to item desirability estimates and it could be investigated which matching procedures yield smaller fakability.

Moreover, due to the different response processes, the MFC format and single-stimulus formats might measure (slightly) different constructs (Guenole et al., 2018; Wetzel & Frick, 2020; Wetzel, Roberts, et al., 2016). This raises the question which construct researchers actually aim to assess. To better compare validities, future research should use designs that can represent the specifics of both formats. Two limitations of our empirical study (Frick et al., 2021) can guide this: First, future research could investigate construct validity when both constructs are assessed with the same type of response format. For example,

Wetzel and Frick (2020) found higher correspondence between self- and other-ratings when both were assessed with an MFC as compared to a rating scale format. Second, future research could compare criterion validities between absolute and relative response formats using criteria that truly value differentiation between behaviors. The question of "ipsative" criteria is not new to MFC research (e.g., Hicks, 1970). However, recent validity research with normative IRT scoring did not explicitly address the type of criteria investigated (Brown & Maydeu-Olivares, 2013; Lee et al., 2018; Walton et al., 2019; Watrin et al., 2019; Wetzel & Frick, 2020; Zhang et al., 2019).

The Faking Mixture model (Frick, 2021b) integrates assumptions and empirical findings about faking in the MFC format into a formal statistical model. In this way, this thesis contributed to theories on the nature of faking in the MFC format. The Faking Mixture model makes the assumption that item desirability is perceived by individuals in the same way. When respondents disagree about which item to prefer when faking, the response probability for each rank order is approximately equal and the block fakability is low. However, empirically, individuals could be strongly convinced that a certain rank order is desirable and be likely to fake. Future research could empirically investigate the assumptions underlying the Faking Mixture model. Moreover, faking good and faking bad can lead to quite different response patterns (Bensch et al., 2019). This cannot be captured by the current model formulation. Future research could extend the Faking Mixture model or develop other modeling approaches to account both for faking good and faking bad.

5.4 Avenues for Psychometric Developments

The Faking Mixture model (Frick, 2021b) is an example of cognitive psychometrics. The field of cognitive psychometrics tries to bridge the gap between psychometrics and cognition research by modeling heterogeneity in persons and items (stimuli) in cognitive (response) processes (Batchelder, 1998; Riefer et al., 2002). In cognition research, this means to model IRT-like heterogeneity in cognitive experiments and in assessment to understand IRT models as models of the response process. Multinomial processing tree models are a class of models that is especially suited for cognitive psychometrics (Batchelder, 1998). In these models, nominal outcomes of responses are modeled by splitting the response process into multiple sub-processes (Erdfelder et al., 2009). Different strategies exist to account for heterogeneity in persons and/or items in these models (e.g., Klauer, 2010; Matzke et al., 2015).

Any IRT model that can be represented with a tree structure can be conceived of as a multinomial processing tree model (Plieninger & Heck, 2018). This applies to the Faking Mixture model, as depicted in Figure 8. Other examples for IRT models with a tree structure are item response tree models (Böckenholt, 2012; De Boeck & Partchev, 2012),

the acquiescence model (Plieninger & Heck, 2018), and the retreive-deceive-transfer model (Leng et al., 2019). To my knowledge, the Faking Mixture model is the first model for response biases or - more generally - response processes in the MFC format that has a tree structure. Future research could develop multinomial processing tree models for other biases in the MFC format such as careless responding, which is the tendency to respond without regard to the item content (Meade & Craig, 2012).

Moreover, response process data could be incorporated into IRT models for the MFC format. Both multinomial processing tree models (e.g., Heck & Erdfelder, 2016; Klauer & Kellen, 2018) and certain IRT models (e.g., Ulitzsch et al., 2020; van der Linden et al., 2010) have been extended to incorporate response times. In addition, there are approaches to modeling response sequences in computerized testing (e.g., Ulitzsch et al., 2021). In a recent think-aloud study, it was found that respondents used different strategies to respond to MFC blocks (Sass et al., 2020). Information about the sequence and timing of rankings could be used to improve trait estimation and its reliability or to better disentangle processes related to faking or careless responding.

In the third manuscript (Frick, 2021a), I investigated methods to automatically assemble MFC tests. Such methods might prove particularly useful, since constructing an MFC test is a complex combinatorical endeavor that requires considering several aspects simultaneously. Hence, future research should further develop algorithms and optimization criteria for the automatic assembly of MFC tests. In the manuscript, for the criteria of A- and D-optimality, I used a very simple greedy heuristic that sequentially selects the next item or block that is optimal at this point. However, the resulting combination of items or blocks might not be optimal. Alternatively, local search heuristics that introduce randomness to keep the search from being trapped in a sub-optimal space can be used. They are often inspired by natural processes, such as genetic algorithms (e.g., Kreitchmann et al., 2021) or ant colony optimization (e.g., Olaru et al., 2015). Future research could develop a local search heuristic, adapt a more sophisticated greedy heuristic (e.g., Luecht, 1998) to MFC blocks or investigate optimization algorithms for non-linear criteria (e.g., Masoudi et al., 2019; Masoudi et al., 2017).

Moreover, future research could investigate how block information can be used for CAT. Two CAT algorithms for the assembly of MFC pairs, based on the Thurstonian IRT model (Lin, 2020) and based on the generalized graded unfolding model for rank data (Joo et al., 2020), already exist. Both algorithms make the assumption that item properties are invariant across block compositions. A CAT algorithm that uses MFC block information would be a useful complement because it can capture item interactions.

Throughout this thesis, I focused on the Thurstonian IRT model. However, it would be interesting to compare and investigate other IRT models for the MFC format as well. The main finding of the simulation study was that trait recovery decreased due to ipsativity

when all items were keyed in the same direction (i.e., they all had positive factor loadings, Frick et al., 2021). Similar effects were found with models for ideal-point items when the item locations were identical (Hontangas et al., 2015; Hontangas et al., 2016). Future research could develop the theoretical conditions for identifying the scale origin with ideal-point items in an MFC format and investigate them in simulation studies. As previously described, the Faking Mixture model could be populated with other IRT models for the MFC format or for single-stimulus formats. Moreover, the block information summaries proposed in the third manuscript (Frick, 2021a) could be adapted to other IRT models for MFC data and it could be examined which algorithms for automated test assembly they can be combined with.

5.5 Conclusion

In this thesis, I investigated and developed item response theory methods for the multidimensional forced-choice format. I focused on three aspects which are relevant for test construction: normativity, fakability and reliability. The research presented provides both guidelines and new tools for MFC test developers. The empirical studies led to new insights about the response process for MFC blocks and highlighted open research questions in this area. The psychometric developments are a starting point for future research on modeling response processes and biases and on automated test assembly. In sum, I hope that the research presented in this thesis will prove valuable for the future construction and psychometric modeling of tests in both multidimensional forced-choice and other response formats.

6 Bibliography

Bartram, D. (2007). Increasing Validity with Forced-Choice Criterion Measurement Formats. *International Journal of Selection and Assessment*, *15*(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*(4), 331–344. https://doi.org/10.1037/1040-3590.10.4.331

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

Berkshire, J. R. (1958). Comparisons of Five Forced-Choice Indices. *Educational and Psychological Measurement*, *18*(3), 553–561. https://doi.org/10.1177/001316445801800309

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. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika*, *79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

Boykin, A. (2020, July). *Simulation studies in psychometrics: State of the practice*. International Meeting of the Psychometric Society.

Bradley, R. A. (1953). Some Statistical Methods in Taste Testing and Quality Evaluation. *Biometrics*, *9*(1), 22–38. https://doi.org/10.2307/3001630

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. https://doi.org/10.2307/2334029

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. (2010). Issues That Should Not Be Overlooked in the Dominance Versus Ideal Point Controversy. *Industrial and Organizational Psychology*, *3*(4), 489–493. https://doi.org/10.1111/j.1754-9434.2010.01277.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. (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

Brown, A., & Maydeu-Olivares, A. (2018a). Modeling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 523–570). Wiley-Blackwell.

Brown, A., & Maydeu-Olivares, A. (2018b). 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., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 1–28. 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

Clemans, W. V. (1966). *An analytical and empirical examination of the properties of ipsative measurement*. Psychometric Society. http://www.psychometrika.org/journal/online/MN14.pdf

Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychological Assessment*, *16*(4), 360–372. https://doi.org/10.1037/1040-3590.16.4.360

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01

Debeer, D., van Rijn, P. W., & Ali, U. S. (2020). Multidimensional Test Assembly Using Mixed-Integer Linear Programming: An Application of Kullback–Leibler Informa-

tion. *Applied Psychological Measurement*, *44*(1), 17–32. https://doi.org/10.1177/0146621619827586

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions*. Drasgow Consulting Group.

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

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial Processing Tree Models: A Review of the Literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *Journal of Consulting Psychology*, *24*(6), 480–482. https://doi.org/10.1037/h0042687

Frick, S. (2021a). Block information in the Thurstonian item response model. *Manuscript submitted to Psychometrika*.

Frick, S. (2021b). Modeling Faking in the Multidimensional Forced-Choice Format - The Faking Mixture Model. *Psychometrika*, Advance online publication. 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*, Advance online publication. https://doi.org/10.1080/00273171.2021.1938960

Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, *14*(3), 251–260. https://doi.org/10.1023/B:STCO.0000035304.20635.31

Genz, A., & Bretz, F. (2002). Comparison of Methods for the Computation of Multivariate *t* Probabilities. *Journal of Computational and Graphical Statistics*, *11*(4), 950–971. https://doi.org/10.1198/106186002394

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology*, *35*(6), 407–412. https://doi.org/10.1037/h0058853

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

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*(5), 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

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

Hofstee, W. K. B. (1970). Comparative Vs. Absolute Judgments of Trait Desirability. *Educational and Psychological Measurement*, *30*(3), 639–646. https://doi.org/10.1177/001316447003000311

Holdsworth, R. F. (2006). *Dimensions Personality Questionnaire*. Talent Q Group.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I., & de la Torre, J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, (28.1), 76–82. https://doi.org/10.7334/psicothema2015.204

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the "Ideal" Personality Response: Effects of Item Matching in Forced Choice Measures for Personnel Selection. *Journal of Personnel Psychology*, *20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

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

Joo, S.-H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods*, *52*, 761–772. https://doi.org/10.3758/s13428-019-01274-6

Kahnemann, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux. https://books.google.de/books?id=ZuKTvERuPG8C&printsec=frontcover&dq=kahneman+thinking+fast+and+slow&hl=de&sa=X&ved=0ahUKEwj3-9CO6-XfAhVJkMMKHSLMAfoQ6AEILzAB#v=onepage&q=kahneman%20thinking%20fast%20and%20slow&f=false

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, *82*, 111–130. https://doi.org/10.1016/j.jmp.2017.12.003

Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2021). A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods*, Advance online publication. https://doi.org/10.3758/s13428-021-01677-4

Lee, P., Joo, S.-H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK Statement and Person Parameter Estimation With Multidimensional Forced Choice Triplets. *Applied Psychological Measurement*, *43*(3), 226–240. https://doi.org/10.1177/0146621618768294

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

Leng, C.-H., Huang, H.-Y., & Yao, G. (2019). A social desirability item response theory model: Retrieve–deceive–transfer. *Psychometrika*, *85*, 56–74. https://doi.org/10.1007/s11336-019-09689-y

Lin, Y. (2020). Asking the Right Questions: Increasing Fairness and Accuracy of Personality Assessments with Computerised Adaptive Testing. *Doctoral Dissertation*. https://doi.org/10.1177/0013164416646162

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

Luecht, R. M. (1998). Computer-Assisted Test Assembly Using Optimization Heuristics. *Applied Psychological Measurement*, *22*(3), 224–236. https://doi.org/10.1177/01466216980223003

MacCann, C., Ziegler, M., & Roberts, R. D. (2011, August 22). Faking in personality assessment. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New Perspectives on Faking in Personality Assessment* (pp. 309–329). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0087

Masoudi, E., Holling, H., Duarte, B. P. M., & Wong, W. K. (2019). A Metaheuristic Adaptive Cubature Based Algorithm to Find Bayesian Optimal Designs for Nonlinear Models. *Advances in Sampling and Optimization*, *28*(4), 861–876. https://doi.org/10.1080/10618600.2019.1601097

Masoudi, E., Holling, H., & Wong, W. K. (2017). Application of imperialist competitive algorithm to find minimax and standardized maximin optimal designs. *Computational Statistics & Data Analysis*, *113*, 330–345. https://doi.org/10.1016/j.csda.2016.06.014

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

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., & 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., Heggestad, 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

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

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

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework:

Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén.

Ng, V., Lee, P., Ho, M.-H. R., Kuykendall, L., Stark, S., & Tay, L. (2020). 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

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

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

Pauls, C. A., & Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and Individual Differences*, *39*(2), 297–308. https://doi.org/10.1016/j.paid.2005.01.003

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

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32–53. https://doi.org/10.1177/0013164416636655

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633–654. https://doi.org/10.1080/00273171.2018.1469966

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*(2), 184–201. https://doi.org/10.1037//1040-3590.14.2.184

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology*, *21*(4), 489–509. https://doi.org/10.1007/s10869-007-9038-9

Salgado, J. F., & Táuriz, 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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4.2), 1–97. https://doi.org/10.1007/BF03372160

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. (2020). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement*, Advance online publication. https://doi.org/10.1177%2F0013164420934861

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. https://doi.org/10.1207/S15327906MBR3502_1

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

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining Clickstream Analyses and Graph-Modeled Data Clustering for Identifying Common Response Processes. *Psychometrika*, *86*(1), 190–214. https://doi.org/10.1007/s11336-020-09743-0

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A Multiprocess Item Response Model for Not-Reached Items due to Time Limits and Quitting. *Educational and Psychological Measurement*, *80*(3), 522–547. https://doi.org/10.1177/0013164419878241

van der Linden, W. J. (2005). *Linear models of optimal test design*. Springer.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

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. (2019). 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*, 134–148. https://doi.org/10.1027/1614-0001/a000285

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford University Press. https://kar.kent.ac.uk/49093/1/Response_biases_Final_accepted_version.pdf

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

Wetzel, E., Frick, S., & Greiff, S. (2020). The Multidimensional Forced-Choice Format as an Alternative for Rating Scales: Current State of the Research. *European Journal of Psychological Assessment*, *36*(4), 511–515. https://doi.org/10.1027/1015-5759/a000609

Wetzel, E., Roberts, B. W., Fraley, R. C., & Brown, A. (2016). Equivalence of Narcissistic Personality Inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality*, *61*, 87–98. https://doi.org/10.1016/j.jrp.2015.12.002

WHOQOL group. (1996). *WHOQOL-BREF. Introduction, administration, scoring and generic version of assessment*. World Health Organization. https://www.who.int/mental_health/media/en/76.pdf

Yousfi, S. (2018). Considering Local Dependencies: Person Parameter Estimation for IRT Models of Forced-Choice Data. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 175–181). Springer International Publishing. https://doi.org/10.1007/978-3-319-77249-3

Yousfi, S. (2020). Person Parameter Estimation for IRT Models of Forced-Choice Data: Merits and Perils of Pseudo-Likelihood Approaches. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: 84th Annual Meeting of the Psychometric Society, Santiago, Chile, 2019* (pp. 31–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-43469-4

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2019). 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