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

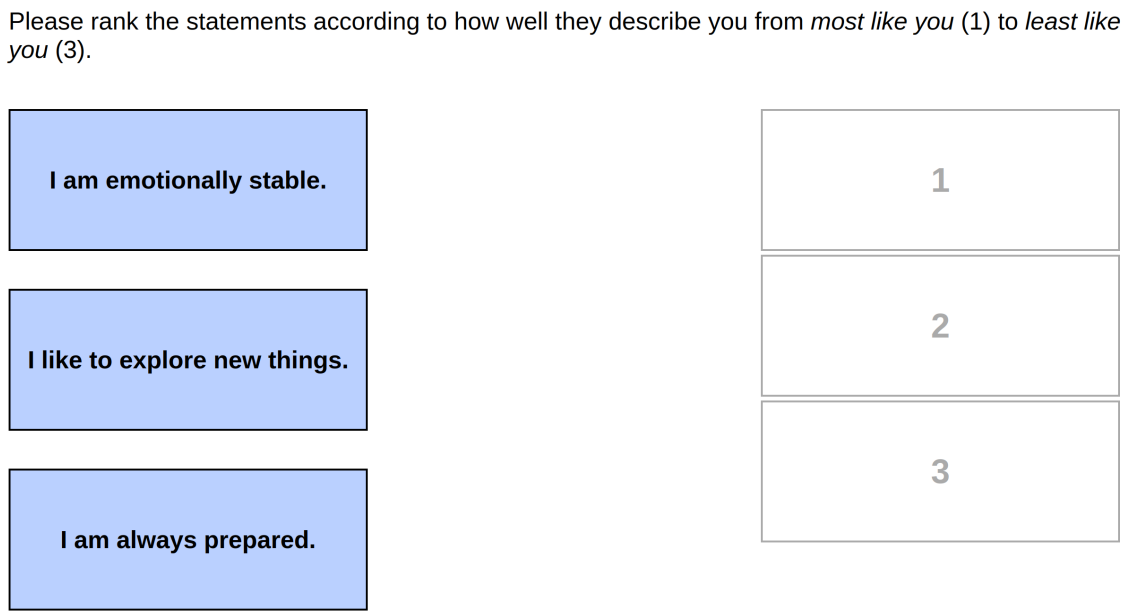


FIGURE 1: Example of a multidimensional forced-choice block from the Big Five Triplets (Wetzel & Frick, 2020).

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).

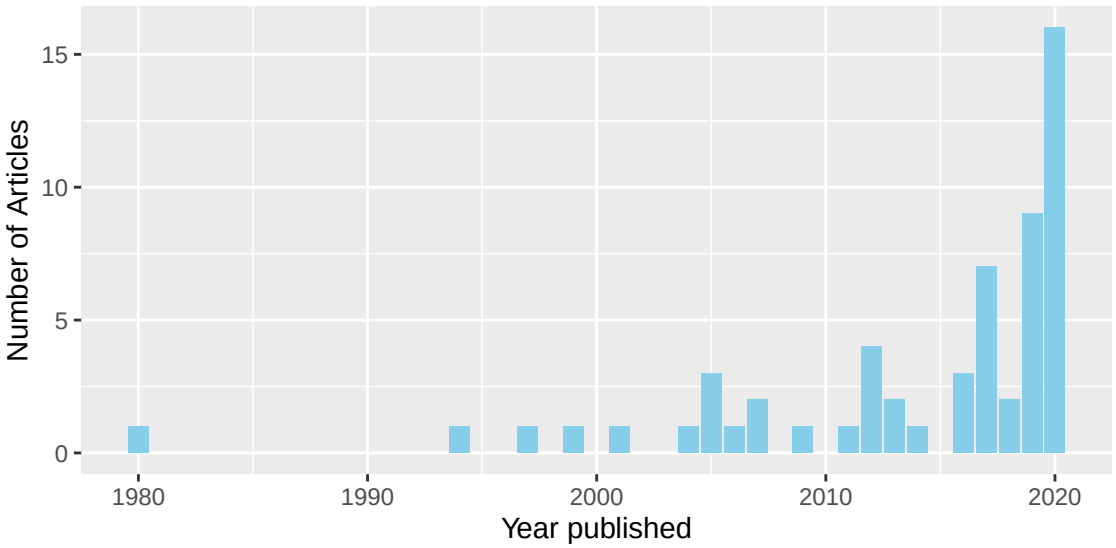


FIGURE 2: Number of new articles published in journals listed in the Web of Science Core Collection including the keywords "multidimensional" and "forced-choice" in any entry.

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

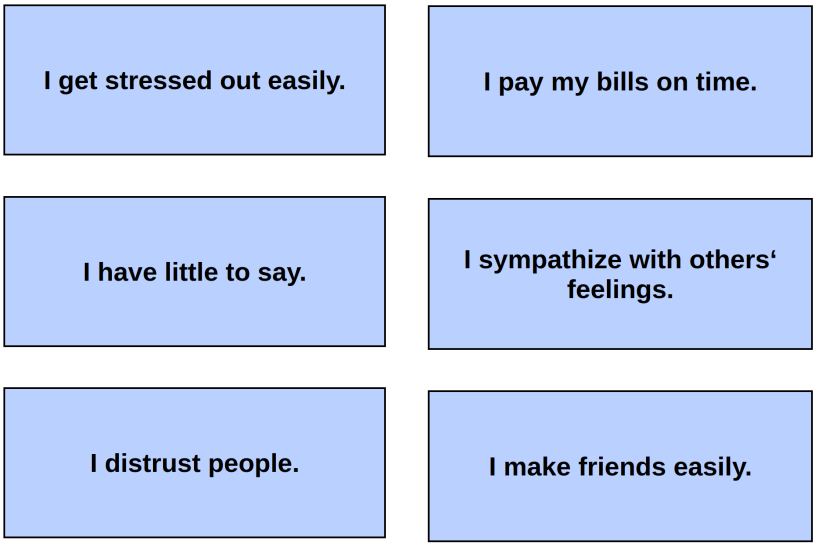


FIGURE 3: Examples of socially undesirable (left) and socially desirable (right) multidimensional forced-choice blocks from the Big Five Triplets (Wetzel & Frick, 2020).

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).

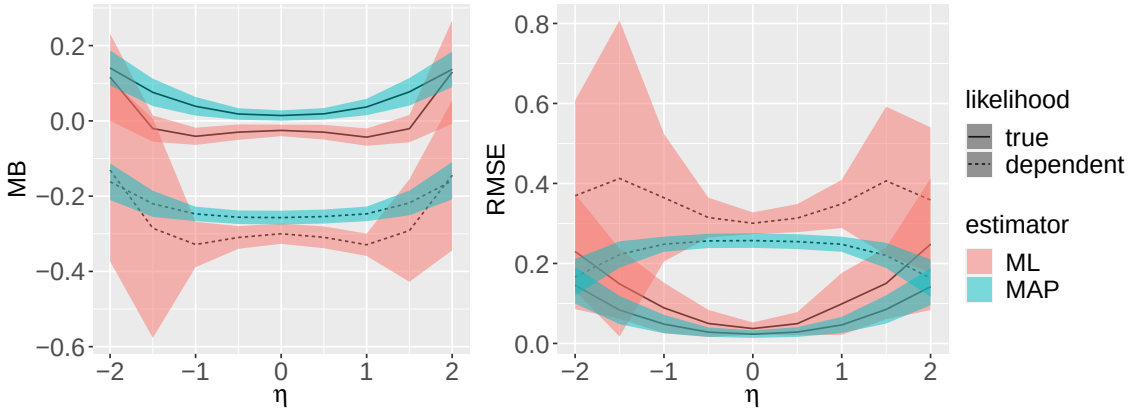


FIGURE 4: Bias of observed standard errors in the simulation on local dependencies. Shaded areas show $\pm 1SD$ around the mean (line). MB = Mean Bias, RMSE = Root Mean Square Error, true = true likelihood, dependent = likelihood neglecting local depencencies, ML = Maximum Likelihood, MAP = Maximum a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

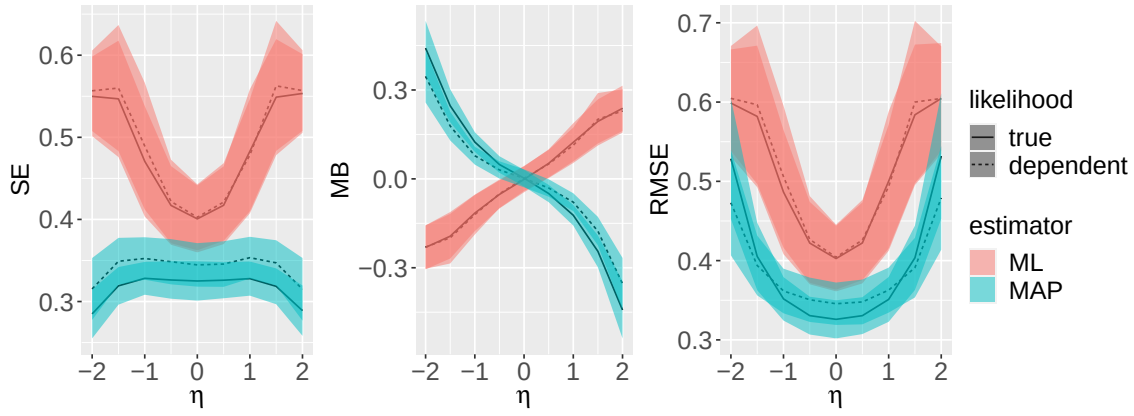


FIGURE 5: Trait recovery and empirical *SE*s in the simulation on local dependencies. Shaded areas show $\pm 1SD$ around the mean (line). SE = empirical Standard Error, MB = Mean Bias, RMSE = Root Mean Square Error, true = true likelihood, dependent = likelihood neglecting local depencencies, ML = Maximum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

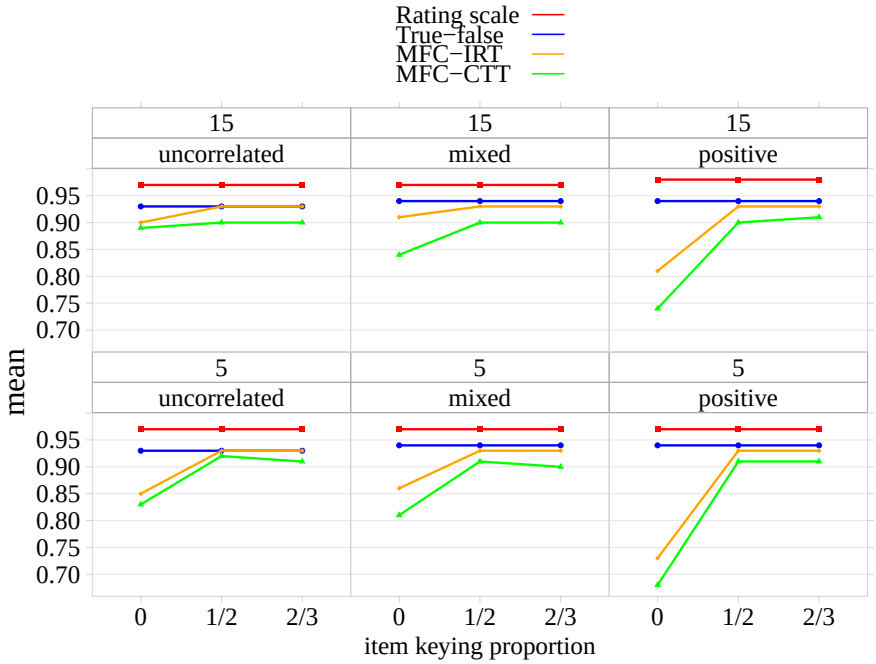


FIGURE 6: Mean correlation between true and estimated traits (i.e., $r(\eta, \hat{\eta})$) by condition. The results were averaged across traits, across block sizes two to four and across equal and unequal numbers of items per trait. MFC = multidimensional forced-choice format; IRT = item response theory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

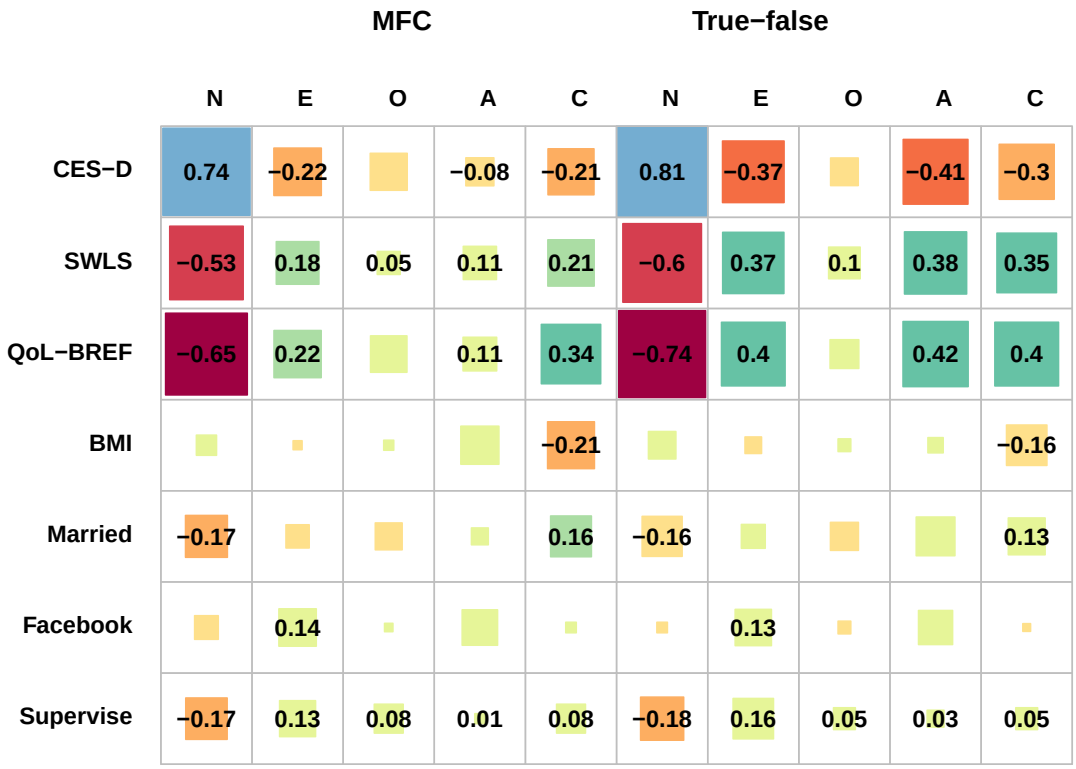


FIGURE 7: Correlations between the Big Five Triplets in the multidimensional forced-choice (MFC) and true-false (TF) version. The size of the square indicates the magnitude of the correlation. Positive correlations are depicted in green and blue and negative correlations in orange and red. Only correlation 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).

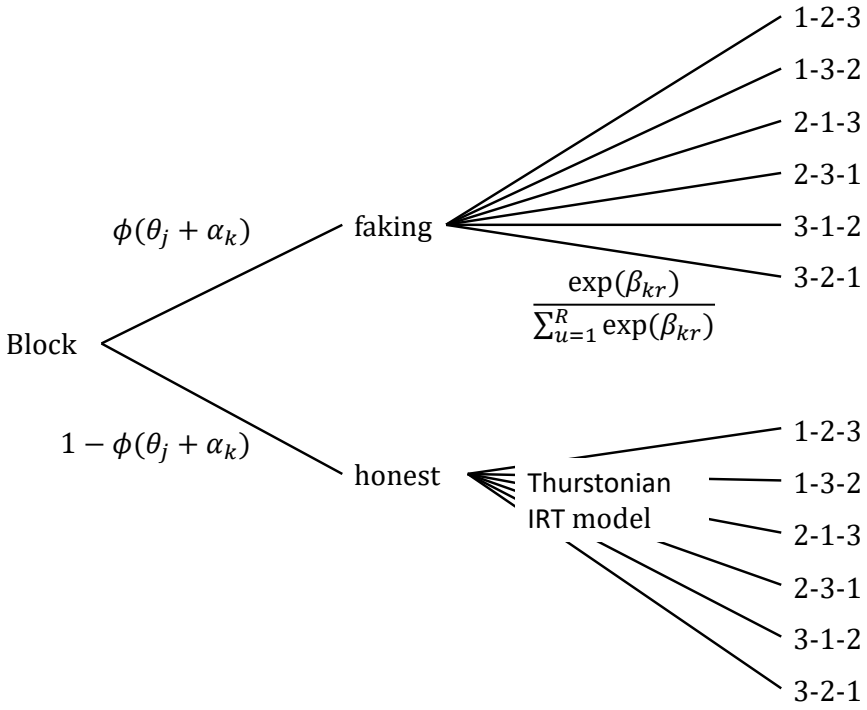


FIGURE 8: The Faking Mixture model depicted as a multinomial processing tree model.

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.

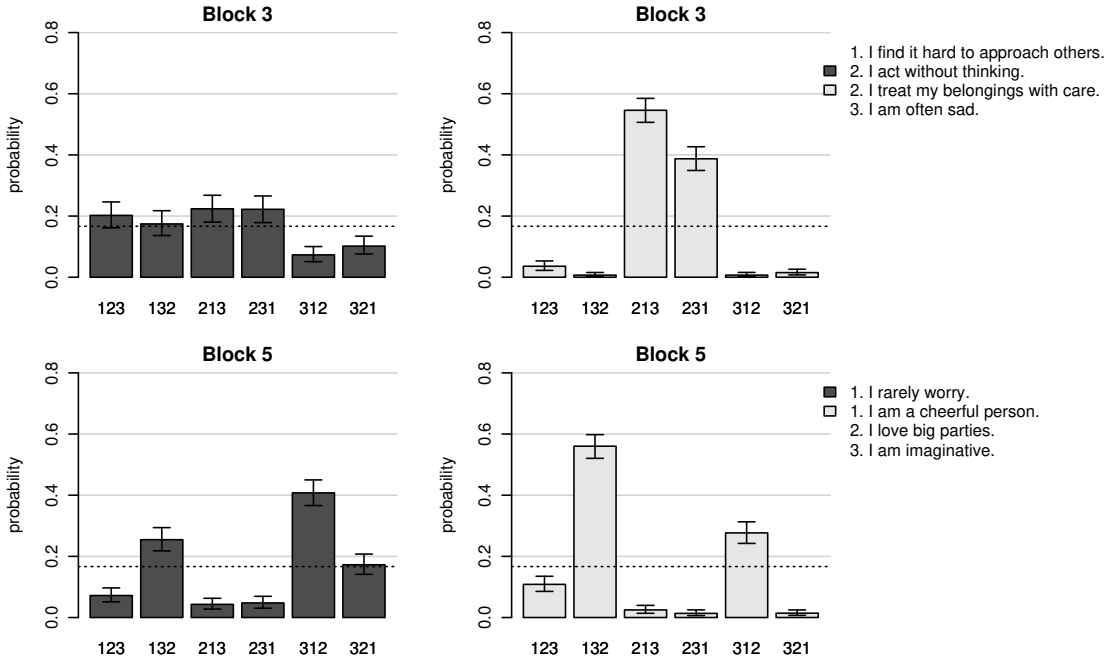


FIGURE 9: Probabilities for rank orders when faking in MFC-matched versus MFC-mixedfor two selected blocks. The dotted line indicates where all rank orders are equally probable. Results for MFC-matched are depicted in dark-grey on the left side, for MFC-mixed in light-grey on the right side.

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

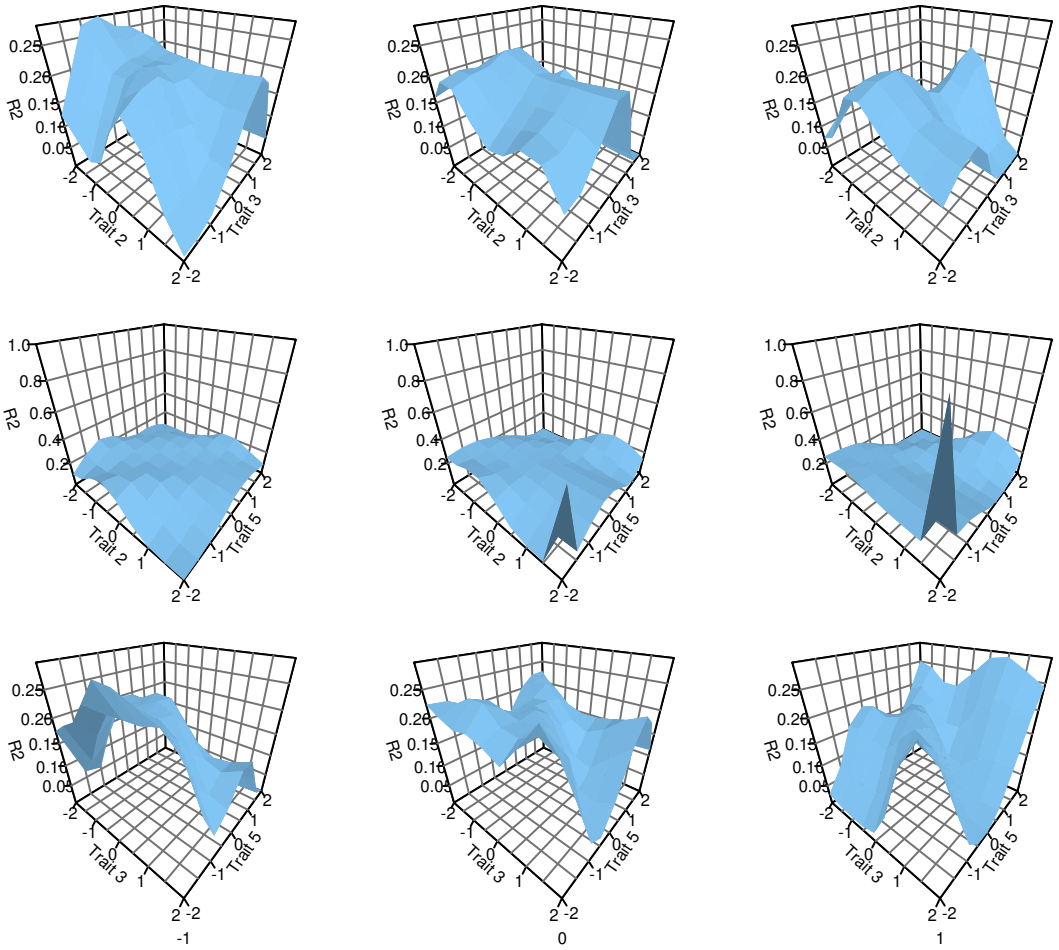


FIGURE 10: Block R^2 for Trait 5 from a simulated test block. Items 1-3 measured traits 2, 3, and 5, respectively. The simulated item parameters were $\mu_1 \approx 0.73$, $\mu_2 \approx -0.89$, $\mu_3 \approx -0.62$, $\lambda_1 \approx 0.92$, $\lambda_2 \approx -0.90$, and $\lambda_3 \approx 0.94$.

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.

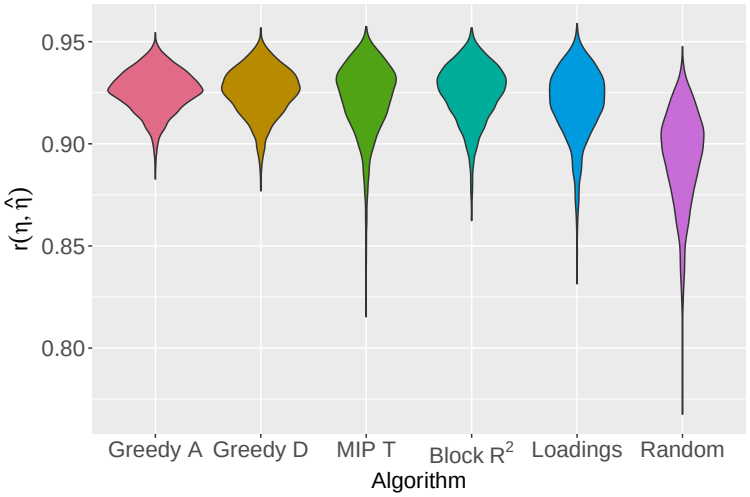


FIGURE 11: Correlation between true and estimated traits ($r(\eta, \hat{\eta})$) by algorithm in the simulation study on test extension for target information proportional to the block pool. A = A-optimality, D = D-optimality, T = T-optimality, MIP = Mixed Integer Programming.

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

A Statement of Originality

Eidesstattliche Versicherung gemäß § 9 Absatz 1 Buchstabe e) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Sozialwissenschaften:

1. Bei der eingereichten Dissertation mit dem Titel *Item Response Theory for the Analysis and Construction of Multidimensional Forced-Choice Tests* handelt es sich um mein eigenständig erstelltes eigenes Werk.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bisher nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Susanne Frick	Ort, Datum

B Co-Authors' Statements

Anna Brown

It is hereby confirmed that the following manuscript included in the present thesis was primarily conceived and written by its first and main author Susanne Frick.

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

Susanne Frick, Anna Brown and Eunike Wetzel jointly developed the idea and design of the simulation study. Susanne Frick implemented all simulation procedures and analyses. She planned and carried out the data collection for the empirical study and most parts of the analyses. She was solely responsible for writing the first draft as well as for the finalization and submission of the paper.

Anna Brown wrote the initial code for trait estimation in the simulation study. She developed the idea for the empirical study and helped with the initial model fitting. Furthermore, she contributed to refining specific parts of the manuscript. Apart from that, Anna Brown helped in numerous discussions to refine the design and interpretation of both studies.

Anna Brown	Place, Date

Eunike Wetzel

It is hereby confirmed that the following manuscript included in the present thesis was primarily conceived and written by its first and main author Susanne Frick.

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

Susanne Frick, Anna Brown and Eunike Wetzel jointly developed the idea and design of the simulation study. Susanne Frick implemented all simulation procedures and analyses. She planned and carried out the data collection for the empirical study and most parts of the analyses. She was solely responsible for writing the first draft as well as for the finalization and submission of the paper.

Eunike Wetzel contributed to the hypotheses and analysis plan for the empirical study. She contributed to refining specific parts of the manuscript. Apart from that, Eunike Wetzel helped in numerous discussions to refine the design and interpretation of both studies.

Eunike Wetzel	Place, Date

C Copies of Articles

Investigating the Normativity of Trait Estimates From Multidimensional Forced-Choice Data

Susanne Frick[1], Anna Brown[2], Eunike Wetzel[3,4] *

[1]Department of Psychology, School of Social Sciences, University of Mannheim

[2]Department of Psychology, University of Kent

[3]Department of Psychology, University of Wien

[4] Department of Psychology, Otto-von-Guericke University Magdeburg

This is the accepted version of the manuscript which was published in Multivariate Behavioral Research.

*This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group „Statistical Modeling in Psychology“ (SMiP). The author acknowledges support by the state of Baden-Württemberg through bwHPC. Eunike Wetzel is now at the Department of Psychology, University of Koblenz-Landau, Landau.

Abstract

The Thurstonian item response model (Thurstonian IRT model) allows deriving normative trait estimates from multidimensional forced-choice (MFC) data. In the MFC format, persons must rank-order items that measure different attributes according to how well the items describe them. This study evaluated the normativity of Thurstonian IRT trait estimates both in a simulation and empirically. The simulation investigated normativity and compared Thurstonian IRT trait estimates to those using classical partially ipsative scoring, from dichotomous true-false (TF) data and rating scale data. The results showed that, with blocks of opposite-keyed items, Thurstonian IRT trait estimates were normative in contrast to classical partially ipsative estimates. Unbalanced numbers of items per trait, few opposite-keyed items, traits correlated positively or assessing fewer traits did not decrease measurement precision markedly. Measurement precision was lower than that of rating scale data. The empirical study investigated whether relative MFC responses provide a better differentiation of behaviors within persons than absolute TF responses. However, criterion validity was equal and construct validity (with constructs measured by rating scales) lower in MFC. Thus, Thurstonian IRT modeling of MFC data overcomes the drawbacks of classical scoring, but gains in validity may depend on eliminating common method biases from the comparison.

Keywords: forced-choice format, Thurstonian IRT model, ipsative data, true-false, rating scale

Investigating the Normativity of Trait Estimates From Multidimensional Forced-Choice Data

In many assessment contexts, it is important to be able to compare persons on certain attributes or traits. For example, an employer might want to compare the conscientiousness levels of applicants. The multidimensional forced-choice (MFC) format[1] has become increasingly popular for such purposes, as evidenced by work-related personality questionnaires such as TAPAS (Drasgow et al., 2012), OPQ (Brown & Bartram, 2009), and the personality questionnaire by TalentQ (Holdsworth, 2006). In the MFC format, several items measuring different attributes are combined into blocks. One type of instruction for an MFC format is to ask respondents to rank all statements within a block. Panel A of Figure 1 shows an example of an MFC block with three statements measuring personality traits.

The MFC format overcomes some of the biases associated with rating scale (RS) items (for an overview, see Brown & Maydeu-Olivares, 2018a). For example, faking can be reduced (Cao & Drasgow, 2019; Pavlov et al., 2019; Wetzel et al., 2021) and halo effects avoided (Brown et al., 2017). Further, construct validity is mostly similar to rating scales (Brown & Maydeu-Olivares, 2013; Lee et al., 2018; Walton et al., 2019; Wetzel & Frick, 2020; Zhang et al., 2019). For an overview on the current state of research on MFC versus rating scales see Wetzel et al. (2020).

However, trait estimates derived from MFC data with classical test theory (CTT) are not normative, but rather ipsative. Trait estimates are termed fully ipsative when the total score is constant across persons (Clemans, 1966). Most authors agree that ipsative trait estimates do not allow inter-individual comparisons (e.g. Closs, 1996; Johnson et al., 1988). Furthermore, correlations based on fully ipsative trait estimates are mathematically constrained (Clemans, 1966). Consequently, correlation-based analyses such as reliability and factor structures are biased (Brown & Maydeu-Olivares, 2013; Clemans, 1966; Hicks, 1970). Several procedures have been developed within CTT that allow the total score to differ between respondents while still retaining some dependency between scale

[1]The MFC format is both an item format and a response format. For simplicity in comparing it with true/false and rating scale formats, we refer to it as a response format in the following.

scores (Hicks, 1970), thereby yielding partially ipsative trait estimates. Partially ipsative trait estimates can prove useful for the prediction of criteria (Salgado & Táuriz, 2014). Nevertheless, they are said to retain characteristics of ipsative trait estimates (Brown & Maydeu-Olivares, 2018b). Several item response theory (IRT) models have been developed for MFC data; most with the aim to provide normative trait estimates (see Brown, 2016a; Brown & Maydeu-Olivares, 2018b for an overview and classification).

The purpose of this study was to evaluate the normativity of IRT trait estimates both in a simulation and empirically in the framework of the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011). So far, the Thurstonian IRT model is the most widely applicable IRT model for MFC data (Brown & Maydeu-Olivares, 2018b): First, it can accommodate MFC formats with varying block sizes and ranking instructions, such as ranking all items within a block or selecting the most and/or least preferred item, in contrast to some other IRT models for MFC data (e.g. Morillo et al., 2016; Stark et al., 2005) . Second, it assumes dominance response process items which are most common in personality psychology (Hontangas et al., 2016). With a dominance response process, the preference for an item increases (or decreases) monotonically with increasing trait levels. In contrast, with an ideal-point response process, the preference for an item is highest at one point of the trait continuum and decreases with increasing distance from this point. Third, item parameters can be estimated directly from MFC responses, whereas some other IRT models for MFC data rely on item parameters obtained from single-stimulus data (e.g. McCloy et al., 2005). Further, we focused on a full ranking instruction, because full ranking provides the most information and therefore the highest reliability (Brown, 2016b; Brown & Maydeu-Olivares, 2018a).

Thurstonian Item Response Model

According to the Thurstonian IRT model, ranking patterns can be encoded with binary variables representing outcomes of the pairwise comparisons. For example, ranking three items involves three pairwise comparisons: between items 1 and 2, items 1 and 3, and items 2 and 3, respectively. The response probability for the outcomes may be calculated depending on the two latent traits η_a and η_b:

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

where γ_l denotes the threshold of outcome l, λ_i and λ_k denote the loadings of items i and k, respectively, and ψ_i^2 and ψ_k^2 denote their uniquenesses. $\Phi\left(x\right)$ denotes the cumulative standard normal distribution function evaluated at x. Equation 1 shows that relative differences between traits impact responses, in contrast to models for single-stimulus data

(e.g. rating scale or true-false data), in which absolute trait levels impact responses. The Thurstonian IRT model's item parameters and trait correlations can be estimated from thresholds and tetrachoric correlations of the binary outcome variables (i.e., using limited information methods). Trait estimates can be estimated with previously obtained item parameters and trait correlations using maximum a posteriori (MAP) or expected a posteriori (EAP) methods. Brown and Maydeu-Olivares (2011, 2012) present details on model restrictions, identification, and estimation.

In IRT, the precision of trait estimation is captured by the item information and depends on the level of the latent trait. MFC questionnaires have an inseparable design, meaning estimation of one trait is dependent on all other traits in the questionnaire (Brown & Maydeu-Olivares, 2018b). With inseparable designs, item information can be described by the Fisher information matrix, which is an $f \times f$ matrix showing information about all possible pairs of f traits. Assuming that items i and k measure traits 1 and 2, respectively, the Fisher information matrix for outcome l is (Brown & Maydeu-Olivares, 2018b):

$$I_l\left(\eta_1, \eta_2\right) = \frac{1}{\psi_i^2 + \psi_k^2} \begin{pmatrix} \lambda_i^2 & -\lambda_i\lambda_k & \cdots & 0 \\ -\lambda_i\lambda_k & \lambda_k^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \frac{\left[\varphi\left(\frac{-\gamma_l + \lambda_i\eta_1 - \lambda_k\eta_2}{\sqrt{\psi_i^2 + \psi_k^2}}\right)\right]^2}{P_l\left(\eta_1, \eta_2\right)\left[1 - P_l\left(\eta_1, \eta_2\right)\right]}, \tag{2}$$

where $\varphi\left(x\right)$ denotes the standard normal density function evaluated at x. Two points are noteworthy in comparison to single-stimulus data with a simple structure: First, the matrix has entries for the two measured traits. Thus, the outcome is informative about those two traits and their combination. In separable designs, each item contributes information only to the trait it measures. In MFC questionnaires however, each item contributes information to several traits compared in the same block, errors of measurement are correlated, and therefore measurement precision is generally lower than in separable designs (Brown & Maydeu-Olivares, 2018b). Second, information is provided per binary outcome. For example, a block of $n = 3$ items provides $n(n-1)/2 = 3$ bits of information.

Previous research on the normativity of trait estimates in the MFC format

The key to deriving normative trait estimates is identifying the scale origin for the latent traits. This depends on the questionnaire design: For the Thurstonian IRT model, Brown (2016a) showed that the scale origin is identified when there are no linear dependencies

between item loadings within blocks and between item loadings within traits[2]. This is in contrast to ideal-point models, where differences between item locations are necessary to identify the scale origin (Brown, 2016a).

In simulation studies, any remaining ipsativity will result in bad recovery of the true parameters. In their simulation studies, Brown and Maydeu-Olivares (2011) found that recovery of item parameters was worse with all positively keyed items. Similarly, Bürkner, et al. (2019) and Schulte et al. (2020) found that trait estimates were ipsative when all items had very similar factor loadings (all positive). Simulation studies employing other IRT models to generate MFC data found similar results, even with partially ipsative CTT scoring (Hontangas et al., 2015, 2016; Morillo et al., 2016). Brown and Maydeu-Olivares pointed out that the low recovery achieved with all positively keyed items and all positively correlated traits is not a limitation of Thurstonian IRT scoring but applies to MFC questionnaires more generally. These empirical results comply with the theoretical rules of identifying the scale origin (Brown, 2016a). Recovery improved when there were more items, when the trait correlations decreased from positive to negative (for mixed item keys), and with larger blocks (Brown & Maydeu-Olivares, 2011).

Mixed-keyed item blocks may have empirical implications: First, Bürkner et al. (2019) argued that only MFC questionnaires with all positively keyed items can be fake-proof. However, on the group level, MFC questionnaires were found to be less fakable than rating scale questionnaires even when blocks contained mixed-keyed items (Heggestad et al., 2006; Wetzel et al., 2021). Second, negatively keyed items and items containing negation must be distinguished. Whereas negations should be avoided in any questionnaire format, negatively keyed items might increase cognitive load in MFC questionnaires. In one study examining the response process to MFC items, participants sometimes reported difficulties in responding to blocks of mixed keyed items (Sass et al., 2020).

In empirical research, normativity is evaluated by comparing MFC trait estimates to single-stimulus trait estimates, such as those from RS data. In addition, Thurstonian IRT trait estimates are compared to (partially) ipsative CTT trait estimates to examine whether the technically demanding IRT scoring provides an advantage over simple CTT scoring. For example, Brown and Maydeu-Olivares (2013) investigated how closely MFC trait estimates approximate normative trait estimates. They compared IRT and CTT scoring of MFC and RS data from a questionnaire that employs both response formats. They found Thurstonian IRT trait estimates to be more similar to RS trait estimates, both from IRT and CTT scoring, than to the CTT ipsative MFC trait estimates. Similarly, Lee et al. (2018) found Thurstonian IRT trait estimates corresponded slightly better to RS trait estimates than trait estimates derived from two partially ipsative scoring methods. Hontangas et al. (2015) transferred this to a simulation study and generated MFC data assuming an

[2]Linear dependencies occur when all item loadings within a block are equal or multiples of each other, or when all loadings for one trait are equal or multiples of each other, see also Brown (2016a).

ideal-point process and analyzed it with an ideal-point IRT model and with CTT scoring, which assumes a dominance response process. Hontangas et al. (2016) repeated the same analyses with data generated under a dominance response process. Thus, in Hontangas et al.'s 2015 simulation, data generation and analyses mismatched for CTT, whereas in his 2016 simulation they mismatched for IRT.

The present research

In this article, we address research questions on the normativity of Thurstonian trait estimates using a simulation study and using an empirical study. The key difference of the simulation study in the present paper with previously published studies is that we investigate the role of various factors in suboptimal questionnaire designs systematically, and evaluate quantitatively their contribution to the normativity of resulting person scores. While previous studies identified the key factors that influence the trait estimates, they did not provide the size of impact depending on the levels in these factors, nor were the levels investigated always representative of questionnaires that are commonly applied. To our knowledge, all previous simulations varied item keying with the levels of 1) all positively keyed items and 2) half of the outcomes involving comparisons between opposite-keyed items. Because mixed-keyed blocks are needed to identify the Thurstonian IRT model parameters, it is especially important to examine levels beyond the optimal balance. Further, the number of items per trait was balanced in previous research. This balance of items per trait and of same and mixed keyed comparisons might be difficult to achieve when constructing an MFC questionnaire – especially when items are matched for their social desirability (Wetzel & Frick, 2020). Indeed, several studies employed questionnaires where the number of items per trait was not balanced (Brown & Maydeu-Olivares, 2013; Heggestad et al., 2006; Ng et al., 2020). In general, traits measured with fewer items will have lower reliability. In the MFC format, this may have unknown consequences for person score because estimation of one trait depends on all other traits. Further, the effects of predominantly positive correlations, which characterize many questionnaires, have not been examined thoroughly. Previous simulation studies using empirical correlation matrices did not investigate the different levels of positive trait correlations (Bürkner et al., 2019; Schulte et al., 2020). Simulation studies investigating different levels of correlations used identical correlations for all traits (Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015, 2016; Morillo et al., 2016). Moreover, we investigate two factors that have not been examined thoroughly: block size and number of traits. Previous simulation studies on trait recovery from MFC questionnaires almost exclusively simulated one fixed block size (Bürkner et al., 2019; Hontangas et al., 2015, 2016; Schulte et al., 2020). The only study that varied block size investigated trait recovery only in one replication (Brown &

Maydeu-Olivares, 2011). Further, it has been long known that the number of traits affects trait recovery for ipsative scoring (Baron, 1996). However, the effect of number of traits on Thurstonian IRT trait recovery was examined in a limited number of studies (Brown & Maydeu-Olivares, 2011; Schulte et al., 2020). The first aim of the simulation study in this paper was to examine Thurstonian IRT trait estimates systematically, under questionnaire design conditions that occur in real-world applications.

The second aim of this simulation was to examine Thurstonian IRT trait estimates using true model parameters. Previous simulation studies confounded the estimation of item/model parameters and person parameters. This is because empirical underidentification might occur in designs with all positively keyed items, (Brown & Maydeu-Olivares, 2012; Bürkner et al., 2019), leading to bias in item parameters and trait correlations. Bias in item parameters then propagates to trait estimates because in the Thurstonian IRT model, traits are estimated with previously obtained item parameters and trait correlations using maximum a posteriori (MAP) or expected a posteriori (EAP) methods. To overcome this confounding effect, we fixed item parameters and trait correlations to their true values to examine trait score estimation in isolation[3]. This procedure is similar to operational assessment settings, in which item parameters and trait correlations are obtained a priori, often from single-stimulus data.

The third aim of this simulation study was to compare Thurstonian IRT trait estimates to those derived from CTT as well as from RS and true-false (TF) data. To our knowledge, there has been no simulation study investigating the comparison to single-stimulus data in detail. However, this is a vital complement to the various validity studies (e.g. Guenole et al., 2018; Lee et al., 2018). We chose the RS format as a comparison because it is the standard in self-report questionnaires. However, the same number of items usually provide more information in the RS than in the MFC format. Therefore, we additionally included the TF format, because, theoretically, the MFC format with three-item blocks provides three bits of binary information, one per each pairwise comparison, which is the same as these three items would provide in the TF format (Brown & Maydeu-Olivares, 2018). To our knowledge, a simulation study comparing IRT and CTT scoring of MFC responses when a dominance model underlies both data generation and analysis is missing. However, as most current MFC questionnaires employ dominance items (Brown & Bartram, 2009; Maydeu-Olivares & Brown, 2010), this comparison is especially important.

The goal of our empirical study was to examine the differentiation of judgments in the MFC and the true-false format by evaluating reliability and validity of person scores. The MFC format elicits relative judgments, as incorporated in choice models for ranking tasks (Brown, 2016a) and indicated by a think-aloud study (Sass et al., 2020). In contrast, single-stimulus formats should elicit absolute judgments. The two types of judgments might

[3]We also carried out a similar simulation in which item and trait parameters were estimated from the data, see https://osf.io/whv9k/?view_only=1e1fde593a424d13a7bac442017a13ae.

correspond to different levels of differentiation (Kahnemann, 2011). The MFC format requires participants to weigh different behaviors against each other, providing potentially more information about the differences between traits, whereas an absolute response format might elicit fast and heuristic response processes. If this is true, it should translate to differences in validity between relative and absolute response formats when the amount of information is held constant. Therefore, we compared latent traits from the MFC and the TF format with regard to reliability, construct validity, and criterion-related validity to gain insight into the differentiation of judgments. To our knowledge, there has been no empirical study comparing validity between the MFC and the TF format in a within-subject design.

Simulation Study

The hypotheses and design of this simulation study were preregistered on the Open Science Framework (https://osf.io/exqb2/?view_only=7692f926a8a34e9f930f75ef02fd0ed0, https://osf.io/uh4t9/?view_only=9e85a4e733fa49f4be2e3dc4aaf8f423). To investigate the role of the factors noted above, data were simulated under different conditions, namely, varying the number of traits, trait correlations, the proportion of comparisons involving opposite-keyed items, the number of items per trait, and block size. MFC data were analyzed with the Thurstonian IRT model and with CTT. TF and RS data were analyzed with appropriate IRT models. The aims above translate to the following research questions:

Research Questions (RQ)

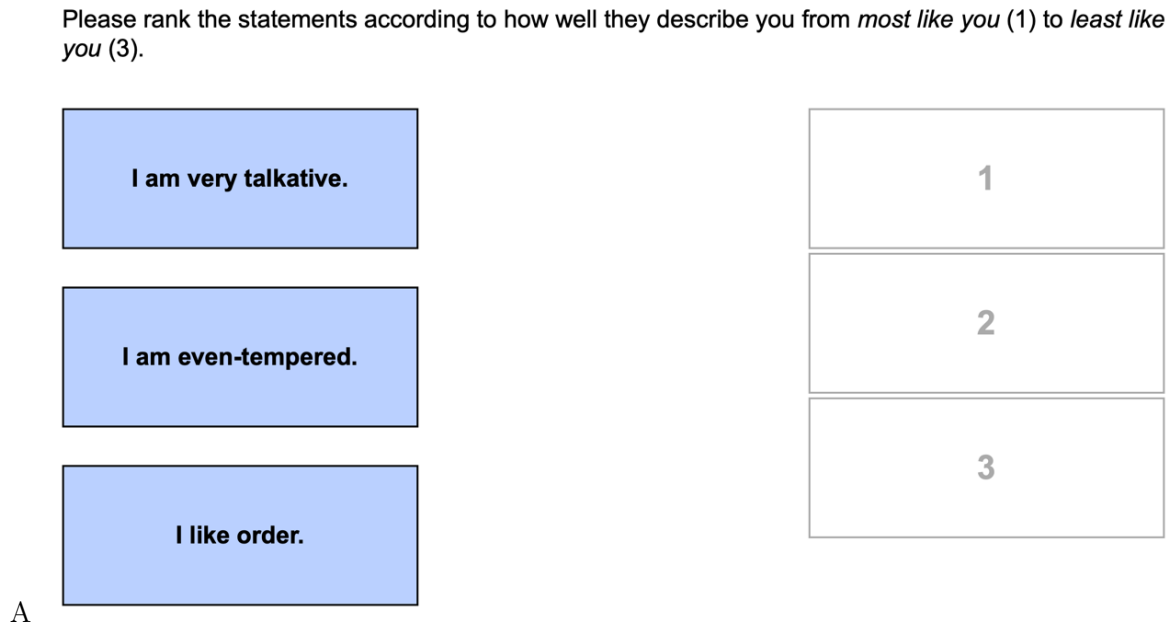
1. How do questionnaire design factors (number of traits, trait correlations, item keying, unequal numbers of items per trait, block size) impact Thurstonian IRT trait recovery?

2. How normative are Thurstonian IRT traits estimated from true item parameters and trait correlations?

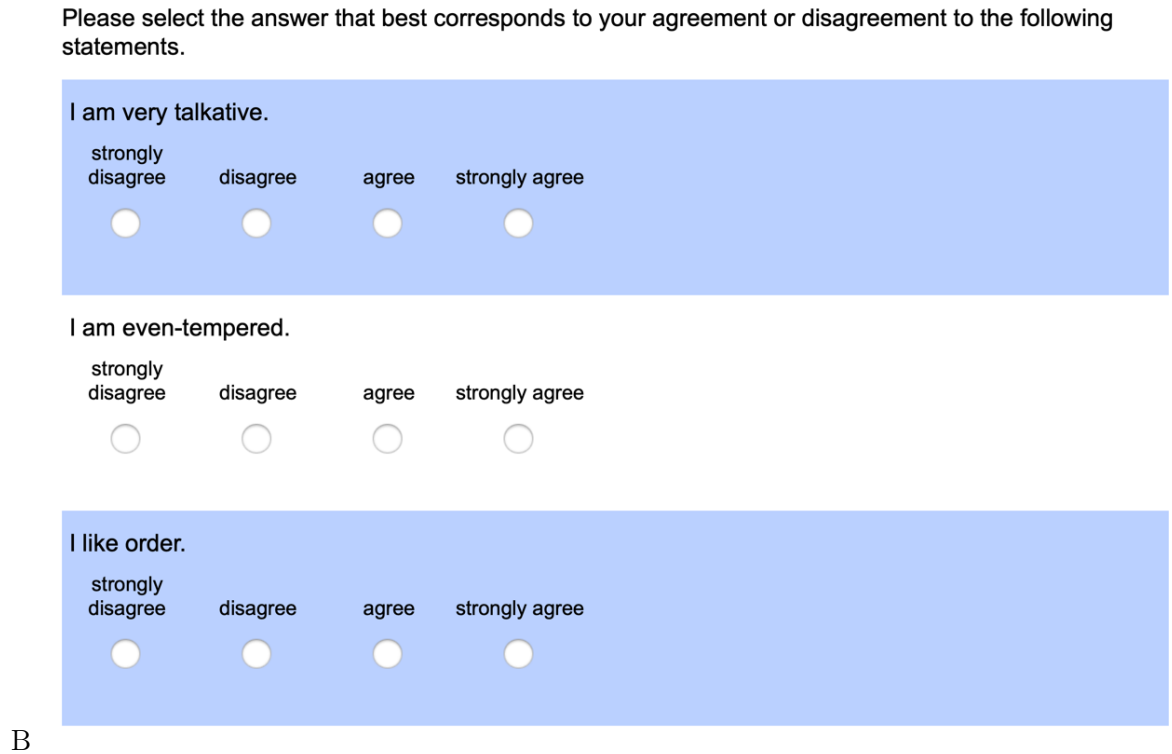
3. How do Thurstonian IRT-estimated traits compare to a) classical (partially) ipsative scores, b) TF scores, and c) RS scores?

4. Which of the factors influencing Thurstonian IRT trait estimation also impact the classical ipsative scoring method?

To investigate these questions, we set up the following simulation design.



A



B

Figure 1: Panel A shows an example for a multidimensional forced-choice format. Panel B shows an example for a rating scale format. In both examples, the first item assesses extraversion, the second neuroticism, and the third conscientiousness.

Table 1: Simulation Design.

Data structure		Factor	Levels			
Level 3	Same trait levels	Number of traits	5, 15			
		Correlations	Mixed, all positive, uncorrelated			
Level 2	Same item parameters	Number of items per trait	Equal, Unequal 1, Unequal 2			
Level 1		Item keying	0, 1/2, 2/3 mixed comparisons			
		Block size	2, 3, 4			
		Score type				
		Response format (model)	TF (normal ogive)	RS (GRM)	MFC (Thurstonian factor model)	
		Analysis model	Normal ogive	GRM	Mean scores	Thurstonian IRT model
			TF	RS	MFC-CTT	MFC-IRT

Note. All factors were completely crossed. TF = true-false, RS = rating scale, MFC = multidimensional forced-choice, GRM = graded response model, Thurstonian IRT model = Thurstonian item response model, IRT = item response theory scoring, CTT = classical test theory scoring.

Simulation Design

Six factors were manipulated and completely crossed: number of traits, trait correlations, block size, number of items per trait, item keying, and score type, as depicted in Table 1[4]. The factor *number of traits* had two levels: five and 15. Five traits are representative of constructs like the Big Five. Fifteen traits are representative of work-related personality constructs such as those assessed in TAPAS (Drasgow et al., 2012), O*NET (Peterson et al., 1999), or Talent-Q (Holdsworth, 2006). The second factor *trait correlations* had three levels: uncorrelated, mixed, and all positive. All uncorrelated traits were included as a neutral benchmark. To increase ecological validity, correlations were based on meta-analytic correlations of the Big Five (neuroticism, extraversion, openness, agreeableness, and conscientiousness), as reported by van der Linden et al. (2010): –.36, –.17, –.36, –.43, .43, .26, .29, .21, .20, .43 for correlations between neuroticism and extraversion, neuroticism and openness, and so forth. For 15 traits, this means that three traits were negatively correlated with the 12 other traits, 59% of the correlations were small, 40% were medium and 1% were negligible (according to Cohen, 1988). To achieve this, absolute values for correlations were drawn randomly from an inverse Wishart distribution with 100 degrees of freedom and covariances set to .3. Then, traits 1, 6, and 11 were reversed. For the mixed correlation condition, the correlations described above were used directly (resulting in Mean correlation .05 for 5 traits and .08 for 15 traits). For the all positive correlation condition, for five traits, the correlations with neuroticism (Trait 1) were reversed, turning neuroticism into its positive counterpart emotional stability (resulting in Mean correlation .31). For 15 traits, Traits 1, 6, and 11 were reversed (Mean = .29).

The third factor, *block size*, had three levels: two (pairs), three (triplets), and four (quads).

The fourth factor, *number of items per trait*, had three levels: Equal, Unequal 1, and Unequal 2 (see Tables 2 and 3). For five traits, in Unequal 1, Traits 1 and 4 were measured with half the number of items than the rest of traits. To obtain Unequal 2, Traits 1 and 2 were switched such that Traits 2 and 4 were measured with fewer items. For 15 traits, in Unequal 1, Traits 1, 4, 6, 9, 11, and 14 were measured with fewer items and in Unequal 2, Traits 2, 4, 7, 9, 12, and 14 were measured with fewer items. Thus, the Unequal 1 and Unequal 2 conditions were created to vary the less reliably measured traits, so that no confounding with the trait correlation factor could occur. The result of some traits having fewer items had an impact on the balance of pairwise comparisons in the MFC version. For example, in the unequal versions, some pairwise comparisons were missing (see Table S1). The full design matrices for all conditions are available from

[4]Following the suggestion of two reviewers, we extended the simulation design by two factors (number of traits and block size). In the previous version, we also investigated questionnaire length and 1/3 mixed comparisons. Results of these analyses can be found on https://osf.io/kpumb/?view_only=3d058747724e4c66999f3d97c376d448.

https://osf.io/pcnwv/?view_only=35fae1b0ec474d768bf7688a17d16208.

The fifth factor was *item keying*. Specifically, the proportion of pairwise comparisons between opposite-keyed items in the MFC format, termed mixed comparisons in the following, was varied. The proportion of mixed comparisons was held constant across all pairwise trait comparisons. The factor item keying had three levels: 0 (i.e., all items positively keyed), 1/2, and 2/3 mixed comparisons. Numbers of items were chosen such that all mixed comparison levels could be constructed[5].

The sixth factor, *score type*, refers to the four response format × scoring method combinations: MFC-CTT, MFC-IRT, TF and RS.

For each research question (RQ), we formulated several hypotheses based on the theory of comparative judgements and previous simulation studies. If not otherwise stated, hypotheses concern the recovery of true scores across traits or pairs of traits. The central hypotheses are listed below, a few subordinate hypotheses can be found in the supplemental online material (SOM) as well as the preregistration.

[5]Under these premises, it was not possible to construct questionnaires with equal and unequal numbers of items per trait and the same total number of items. For this reason, total numbers of items were selected with a minimal difference between those questionnaire versions. Furthermore, they were selected to be representative of the typical lengths of questionnaires.

Table 2: Number of positively and negatively keyed items per trait in the simulated questionnaires with 5 traits.

Block-size	Mixed comparisons	Item keying	Equal					Unequal 1(2)				
			Trait					Trait				
			1	2	3	4	5	1(2)	2(1)	3	4	5
2	1/2	-	6	6	6	6	6	4	24	20	4	20
		+	18	18	18	18	18	14	12	16	14	16
	2/3	-	8	8	8	8	8	6	24	14	6	22
		+	16	16	16	16	16	12	12	22	12	14
		Total	24	24	24	24	24	18	36	36	18	36
							120					144
3	1/2	-	3	3	3	3	3	3	4	4	3	4
		+	9	9	9	9	9	6	14	14	6	14
	2/3	-	4	4	4	4	4	3	6	6	3	6
		+	8	8	8	8	8	6	12	12	6	12
		Total	12	12	12	12	12	9	18	18	9	18
							20					24
4	1/2	-	2	2	2	2	2	2	3	4	2	3
		+	6	6	6	6	6	4	9	8	4	9
	2/3	-	4	4	4	4	4	3	6	6	3	6
		+	4	4	4	4	4	3	6	6	3	6
		Total	8	8	8	8	8	6	12	12	6	12
							40					48

Note. Versions 1 and 2 of unequal numbers of items per trait differed in that traits 1 and 2 were switched. Number of items per trait are displayed for the short questionnaires For all positively keyed items, the total number of items for each trait was positively keyed.

Table 3: Number of positively and negatively keyed items per trait in the simulated questionnaires with 15 traits.

Block-size	Mixed com-parisons	Item keying	Equal	Unequal 1(2)														
			Trait 1–15	Trait 1(2)	2(1)	3	4	5	6(7)	7(6)	8	9	10	11(12)	12(11)	13	14	15
2	1/2	-	6	5	8	10	5	10	5	8	10	5	8	5	8	8	5	8
		+	18	13	28	26	13	26	13	28	26	13	28	13	28	28	13	28
	2/3	-	8	6	12	12	6	12	6	12	12	6	12	6	12	12	6	12
		+	16	12	24	24	12	24	12	24	24	12	24	12	24	24	12	24
		Total	24	18	36	36	18	36	18	36	36	18	36	18	36	36	18	36
			360															432
3	1/2	-	3	3	4	4	3	4	3	4	4	3	4	3	4	4	3	4
		+	9	6	14	14	6	14	6	14	14	6	14	6	14	14	6	14
	2/3	-	4	3	6	6	3	6	3	6	6	3	6	3	6	6	3	6
		+	8	6	12	12	6	12	6	12	12	6	12	6	12	12	6	12
		Total	12	9	18	18	9	18	9	18	18	9	18	9	18	18	9	18
			60															72
4	1/2	-	2	2	3	4	2	4	2	3	4	2	3	2	3	3	2	3
		+	6	4	9	8	4	8	4	9	8	4	9	4	9	9	4	9
	2/3	-	4	3	6	6	3	6	3	6	6	3	6	3	6	6	3	6
		+	4	3	6	6	3	6	3	6	6	3	6	3	6	6	3	6
		Total	8	6	12	12	6	12	6	12	12	6	12	6	12	12	6	12
			120															144

Note. Versions 1 and 2 of unequal numbers of items per trait differed in that traits 1 and 2, 6 and 7, and 11 and 12 were switched. For all positively keyed items, the total number of items for each trait was positively keyed.

RQ 1. Questionnaire Design and MFC-IRT Scoring

In RQ 1, we investigated whether questionnaire design factors impact trait recovery. These hypotheses only concern MFC-IRT. They are seen as supported when they hold true for correlations between true and estimated scores ($r(\theta, \widehat{\theta})$), mean absolute bias (MAB), and mean squared error (MSE).

Item Keying

Identification of the scale origin relies on differences in factor loadings (see Brown, 2016a). Those differences are much smaller than when loadings are allowed to differ in sign, leading to worse trait recovery. Therefore, we expected that:

H1a: Recovery will be worse with 0 than with 1/2, and 2/3 mixed comparisons.

Trait Correlations

Previous simulations showed that trait recovery improved when trait correlations decreased from positive to negative (Brown & Maydeu-Olivares, 2011).

H1b: Trait recovery quality will be ordered as follows: mixed > all uncorrelated > all positive correlations.

Number of Items per Trait

We ran a previous, unpublished simulation with a similar design and with item parameters estimated from the data3. In this simulation, trait recovery did not differ between questionnaires with equal and unequal numbers of items per trait.

H1c: We do not expect recovery to differ between questionnaires with equal and unequal numbers of items per trait

Number of Traits

Previous simulations (Bürkner et al., 2019; Schulte et al., 2020) and empirical studies (Baron, 1996) found recovery to improve with more traits even for ipsative scoring methods. This is because the more traits there are, the less likely the person true scores will be all high or all low – thus reducing the distortion to most scores by the ipsative centering on the person mean (Baron, 1996). Therefore, we expect that:

H1g: Trait recovery will be better with 15 than with 5 traits.

Block Size

In blocks of three or more items, there are local dependencies among the pairwise comparisons that are ignored in the person score estimation (Brown & Maydeu-Olivares, 2011).

Thus, it is assumed that each pairwise comparison contributes unique information when in fact they do not. Therefore, given the same number of pairwise comparisons, smaller blocks provide more information.

H1h: Trait recovery will be better for smaller blocks, i.e. it will be better for block size two than three, and better for block size three than four, holding the number of pairwise comparisons equal.

Item Keying × Trait Correlations

Item loadings interact with trait correlations (Brown & Maydeu-Olivares, 2011). The more positively the traits correlate, the smaller the variance in trait differences and the higher the variance in trait sums. Thus, with strongly positively correlated traits, differentiation between persons is better when comparing opposite-keyed items whereas with strongly negatively correlated traits, equally-keyed items provide a better differentiation.

H1d.1: The effect in H1a will be larger for all positive than for 0 and mixed correlations.

For each trait correlation level, there will be an optimum (or best performing) level of item keying (H1d.2-5, see SOM).

Number of Items per Trait × Trait Correlations

The effect of trait correlations should be more pronounced when the negatively correlated traits are measured with more items (Unequal 2) than with less (Unequal 1) because trait recovery improves with negatively correlated traits (e.g. Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; H1e.1-3, see SOM).

Item Keying × Number of Traits

Ipsativity effects are larger with fewer traits (see Number of traits).

H1f.1: The effect in H1a will be larger for 5 than for 15 traits.

Empirical Reliability

H1i: We expect empirical reliability to overestimate true reliability due to local dependencies in blocks of size > 2. Overestimation will be larger for block size 4 than 3.

RQ 2. Normativity of MFC-IRT Scores

In RQ2, we investigated how normative Thurstonian IRT traits estimated from true item parameters and trait correlations are. These hypotheses only concern MFC-IRT. We used two indicators to quantify normativity. 1) For fully ipsative trait estimates, the mean intercorrelation of k traits is constrained to $-1/(k-1)$(Clemans, 1966). Therefore, according to Hicks (1970), the mean trait intercorrelation can be used to quantify the degree of

normativity in the trait estimates. 2) MFC trait estimates are fully ipsative when the sum of all trait scores is constant for everyone (Hicks, 1970). In this case, it is impossible to distinguish between two persons who have the same shape of the trait profile, or in other words, equal differences between the trait scores, but differ on the absolute location of the trait profile, i.e. the sum of the trait scores. Therefore, we used the recovery of sums (absolute trait levels) and differences (relative trait levels) of traits as a second indicator of normativity.

Mean Correlation

H2a: The mean trait correlation will be unbiased for all item keying levels.

This is because item loadings are drawn such that the scale origin should be identified for all item keying levels. Empirical underidentification should not occur due to fixed item parameters and trait correlations.

Sums and Differences

As Equation 2 shows, item information is dependent on the intercept and the difference between the two traits times their loadings. It follows that item loadings play a crucial role in measuring sums and differences of traits: Brown and Maydeu-Olivares (2011) showed that comparisons between items with loadings of the same sign (for example, positive) contribute to the measurement of differences between traits within a person. In contrast, comparing items with loadings of opposite signs contributes to the measurement of the sum of the two traits. The following hypotheses are seen as supported when they hold true for MAB and MSE.

H2b: Trait recovery for sums of traits and the total score will be better for 1/2 and 2/3 than for 0 mixed comparisons.

H2c: Trait recovery for differences of traits will be better for 0 and 1/2 than for 2/3.

Number of Traits

H2d: The effects in H2b and H2c will be larger for 5 than for 15 traits.

RQ 3. Comparison Between Formats and Scoring Methods

In RQ3, we compared Thurstonian IRT-estimated traits to classical (partially) ipsative scores, and IRT-estimated TF and RS scores. We expected trait recovery to be ordered as follows: RS-IRT $>>$ TF-IRT $>$ MFC-IRT $>>$ MFC-CTT, where $>>$ signifies larger differences than $>$. This is because 5-point rating scales provide more bits of information than TF or MFC with three-item blocks. In our design, MFC and TF provide the same number of bits of information, but for MFC, estimation for all traits is interdependent,

which should lower information slightly. Recovery for MFC-CTT should be worse because of ipsativity. This translates to the following hypotheses:

H3a: Trait recovery will be better in RS than in TF and MFC-IRT.

H3b: Trait recovery will be better in TF than in MFC-IRT.

H3c: Trait recovery will be worse in MFC-CTT than in the other score types.

H3d: The differences in H3a and H3c will be larger than those in H3b.

RQ 4. Questionnaire Design and CTT Scoring

In RQ4, we investigated which questionnaire design factors impact scores from MFC-CTT scoring. Most of these hypotheses are based on the same reasoning as for MFC-IRT (see above).

Item Keying

Previous research has shown trait recovery to be best with completely balanced number of comparisons between items keyed in the same direction and between items keyed in the opposite directions. With this design, both sums and differences of traits are measured equally well. Therefore, we expected the 1/2 mixed comparisons level to show best trait recovery.

H4a.1: Trait recovery will be better with 1/2 than with 2/3 mixed comparisons.

For 0 mixed comparisons, MFC-CTT scoring yields fully ipsative trait estimates.

H4a.2: Trait recovery will be worse with 0 than with 1/2 and 2/3 mixed comparisons.

H4a.3: The difference in H4a.2 will be larger than that in H4a.1.

Trait Correlations

H4b: Trait recovery will be ordered as follows: mixed $>$ all uncorrelated $>$ all positive correlations.

Item Keying $\times$ Trait Correlations

H4c: The effect in H4a.2 will be larger for all positive than for 0 and mixed correlations.

Items per Trait $\times$ Trait Correlations

Trait correlations play a more important role in Unequal 2 than in Unequal 1 and for traits that correlate negatively with the rest (H4e.1-3, see SOM).

Normativity

Mean Trait Correlation

H4f.1: We expect the mean trait correlation to be biased in all designs.

Sums and Differences of Traits

H4f.2: Trait recovery for differences of traits will be better than for sums of traits.

Number of Traits

H4f.3: The bias in the mean trait intercorrelation will be larger for 5 than for 15 traits.

H4f.4: The effect in H4f.2 will be larger for 5 than for 15 traits.

Methods

Data were generated for a sample size of 1000 persons. Samples as large as this allow more outliers and therefore allow examining cases of unusual score combinations thus providing less favorable conditions. RS responses were simulated for a five-point scale and TF responses were simulated as binary in all 162 conditions.

The basic simulation procedure was as follows: First, trait levels and item parameters were generated for all conditions. Second, MFC, RS, and TF data were simulated with the generated trait levels and item parameters. IRT trait estimates were estimated based on the item parameters and trait correlations by maximizing the mode of the posterior likelihood distribution (maximum a posteriori, or MAP). CTT trait estimates were computed as mean scores and subsequently z-standardized. Third, indices for trait estimation quality were computed in each condition. There were 1000 replications per condition. The software R (R Core Team, 2017) was used for data generation and analysis. In addition, we used the R packages mvtnorm (Genz et al., 2020), car (Fox & Weisberg, 2019), and psych (Revelle, 2019).

As much as the design allowed, common random numbers were used to reduce overall variance (Skrondal, 2000), resulting in a three-level hierarchical data structure as depicted in the left column of Table 1. First, the same trait levels were used for one replication within one number of traits $\times$ trait correlation combination. Second, the same item parameters were used for one replication within one block size $\times$ number of items per trait combination.

Data Generation

Trait levels were drawn from a multivariate normal distribution with means of zero and standard deviations of one for each trait and the trait correlation levels as appropriate for the condition (i.e. mixed, all positive, uncorrelated). Following the suggestion of an anonymous reviewer, we conducted an additional simulation on the size of standard errors for the IRT-based scoring methods. Here, Trait 2 was fixed to 0, 2 and -2, while the other traits were drawn from the same multivariate normal distribution, for 300 persons each. Standard errors were averaged across persons with the same level on Trait 2 and across the 1000 replications per condition.

Item loadings were drawn from U(.65,.95) and item means (i.e. item intercepts in item

factor analysis) were drawn from U(–1,1). These are typical values for standardized continuous item utilities (Brown & Maydeu-Olivares, 2011). The loadings were redrawn until there were no linear dependencies2 between item loadings within blocks and between item loadings within traits (for reasons of identification, see Brown, 2016a). This was ensured for all item keying levels. For the RS format, deviation factors were sampled from U(–1.8, –0.9), U(–0.6, –0.15), U(0.15, 0.6), and U(0.9, 1.8) for the first, second, third, and fourth threshold, respectively. Sampling distributions for deviation factors were chosen to be similar to empirical datasets and to be symmetrical. RS-thresholds can be calculated with the item mean as location: item mean + deviation. Uniquenesses were specified as 1–loading2.

Errors were sampled for each person on each item from N(0, uniqueness). Then, continuous item utilities were generated with the loadings, errors, and item means, according to a respective factor model. The same utilities were used to generate the data for MFC, RS, and TF. MFC data were generated under the Thurstonian factor model by computing pairwise differences of item utilities within each block and dichotomizing them using the threshold of 0, so that the outcome was 1 if the first utility was greater than the second and it was 0 otherwise, as the Thurstonian models suggest. TF data were generated under the normal ogive model (Tucker, 1946) by dichotomizing the item utilities using the threshold of 0, so that the outcome was 1 if the utility was greater than the item mean and it was 0 otherwise. RS data were generated under the graded response model (Samejima, 1969) by categorizing the item utilities by the deviation factors.

For MFC-CTT, ranks were transformed to scores using the following procedure: For positively keyed items, for block size n ranks 1 to n were recoded to n to 0. For negatively keyed items, ranks 1 to n were recoded to 0 to n, as shown in Table 4 for block size 3. In this example, the sum score across three items can assume the values of 1, 3, and 5 as opposed to only 3 with all positively keyed items. However, different ranking patterns can still lead to the same sum score. Then, mean scores were computed for each trait and z-standardized for comparability with the IRT-based trait estimates.

Data Analysis

Summary measures were computed in each replication for each condition, including ordering, bias measures, the Mahalanobis distance, empirical reliability and bias of mean correlation. Ordering was defined as the correlation between true and estimated trait levels. As bias measures, mean absolute bias (MAB) and mean square error (MSE) were computed, adapting formulas from Feinberg and Rubright (2016) to the case of multiple parameters per replication. For $d = 1 \ldots D$ person parameters with true and estimated values of η_d and $\widehat{\eta}_d$, respectively, MAB and MSE were defined as follows:

Table 4: Ipsative and partially ipsative scoring for block size 3.

Item content	Trait	Keying	Respondent A		Respondent B	
			Rank	Score	Rank	Score
Fully ipsative						
I get stressed out easily.	Neuroticism	+	1	2	3	0
I love big parties.	Extraversion	+	3	0	2	1
I am imaginative.	Openness	+	2	1	1	2
Sum				3		3
Partially ipsative						
I rarely worry.	Neuroticism	–	3	2	1	0
I love big parties.	Extraversion	+	2	1	3	0
I am imaginative.	Openness	+	1	2	2	1
Sum				5		1

$$MAB = \frac{\sum_{d=1}^{D} |\widehat{\eta}_d - \eta_d|}{D - 1}, \tag{3}$$

$$MSE = \frac{\sum_{d=1}^{D} (\widehat{\eta}_d - \eta_d)^2}{D - 1}. \tag{4}$$

Both MAB and MSE are measures of accuracy because they combine systematic and random error, also known as bias and variance. MSE weights extreme values more strongly than MAB (Feinberg & Rubright, 2016). Bias measures were computed for single traits, for the total score (i.e. the sum of all five or 15 traits), and for the sums and differences of two traits (for all 10 or 105 combinations of two traits).

Analogous to Brown and Maydeu-Olivares (2013), the Mahalanobis distance was used as a multivariate distance measure between trait profiles that accounts for correlated traits (Cronbach & Gleser, 1953). The Mahalanobis distance between true and estimated trait profiles was computed for each simulated person with the true trait correlations as correlations between the axes. To summarize Mahalanobis distances across persons, the mean and the squared mean (analogous to MAB and MSE), the median, and the standard deviation were computed.

The correlation between the true and estimated trait score (ordering) was squared to obtain true reliability. In addition, empirical reliability was computed from the SEs of

factor scores estimated by the model, using the formula:

$$r_{empirical} = \frac{Var\left(\widehat{\eta}\right) - Mean(SE^2)}{Var\left(\widehat{\eta}\right)}. \quad (5)$$

Reliabilities above .80 were regarded as acceptable, and above .90 as good (Evers et al., 2013). Raw bias for the mean correlation was calculated by subtracting the mean correlation of estimated factor scores from the true mean correlation. The R-script including all simulation procedures can be found on the Open Science Framework (https://osf.io/pcnwv/?view_only=35fae1b0ec474d768bf7688a17d16208).

Summary measures were analyzed across traits or pairs of traits, except for H1c.3. For statistical analysis, the hypotheses were transformed into planned contrasts. Variance explanation within an ANOVA framework was then calculated for each contrast. This allowed us to evaluate relative effect sizes within the studied conditions. Further, we examined the absolute levels of the summary measures descriptively. For ANOVAs, we considered effects with an associated variance explanation of at least 1% to be meaningful. In contrast to inferential tests, variance explanation is insensitive to heterogeneous variances, which occurred in some conditions as indicated by Levene's test. Moreover, it is insensitive to sample size, which could be arbitrarily increased in simulation studies. For RQs 1 and 2, ANOVAs were restricted to MFC-IRT, for RQ 3, the ANOVA was run across all four score types. For RQ 4, it was restricted to MFC-CTT.

Results

Convergence

In total, scores were estimated for 162,000 Thurstonian IRT, normal ogive and graded response models. For the Thurstonian IRT model, there were 14 runs in which scores for one person could not be estimated. For the graded response model, 14% of models failed to estimate scores of up to 13 persons with 7% being only one person. For the binary normal ogive model, 136 models failed to estimate scores of up to 2 persons. We considered the estimation problems as minor enough to not warrant any further treatment.

RQ 1. Questionnaire Design and MFC-IRT Scoring

In the following, findings from a preregistered hypothesis are marked with their hypothesis number; the other reported findings are exploratory. Overall, item keying showed the largest effect (43% to 47% of total variance, Table 5), followed by the interaction of trait correlations with item keying (16% to 20%). Residual variances were moderate, namely between 14% and 19%.

Item Keying and Trait Correlations. Recovery was worse with 0 mixed comparisons (e.g. mean MAB = .39) as compared to the other levels (mean MAB = .28; in favor of H1a, see also Tables 5 and 6). Only for 0 mixed comparisons, recovery for all positively correlated traits was worse (e.g. mean MAB = .47) than for uncorrelated or mixed trait correlations (mean MAB = .35; in favor of H1d.1), accounting for the whole interaction effect. Differences between the other levels of trait correlations and item keying were negligible (contradicting H1b; see Table S2). Similarly, mean standard errors were larger with 0 mixed comparisons and more so with positively correlated traits (Table 7).

Number of Traits. Standard errors were larger and recovery was worse for 5 (e.g. mean MAB = .34) than for 15 traits (mean MAB = .31), in favor of H1g, but only for 0 mixed comparisons (e.g. mean MAB 15 traits = .37; mean MAB 5 traits = .45, in favor of H1f.1), accounting for the whole interaction effect.

Block Size. Recovery decreased with increasing block size, explaining 3% of variance (in favor of H1h). However, the effect of block size was rather small, for example the mean MAB was .30 for block size two and .32 for block size three (Table 6). Mean standard errors did not vary by block size (Table 7).

Number of Items per Trait. Recovery was almost identical between equal (e.g. mean MAB = .33) and unequal numbers of items per trait (mean MAB = .31; Table 6; in favor of H1c). The effect of trait correlations was equal across the levels of numbers of items per trait, both overall and for single traits (contradicting H1e.1-3, see Tables 5 and S3). Mean standard errors for Trait 2 were largest in Unequal 2, followed by Equal and Unequal 1 (Table 7), reflecting the number of items, 9, 12, and 18, respectively.

To summarize, if the questionnaire included both positively and negatively keyed items, recovery did not vary substantially across different questionnaire designs.

RQ 2. Normativity and MFC-IRT Scoring

Across item keying levels, the mean correlation was negatively biased, as evidenced by a significant intercept, reflecting the grand mean, of –0.05 ($t(161{,}838) = -2{,}328.76$, $p < .001$, 95% Cl [–0.04855; –0.04847], contradicting H2a). Bias for sums of traits and the total score was smaller for 1/2 and 2/3 (mean MAB = .40; mean MSE = .26) compared to 0 mixed comparisons (mean MAB = .64; mean MSE = .69; in favor of H2b; Tables 8 and 9; see also Figure 2). For differences between traits, bias was larger for 2/3 (mean MAB = .4, mean MSE = .26) compared to 0 and 1/2 mixed comparisons (mean MAB = .39, mean MSE = .24), however this effect was rather small (Table 9; in favor of H2c). This effect was larger for 5 than for 15 traits, but only for sums of traits (Tables 8, 9, S4 and

Table 5: Contrasts and % of variance in summary measures explained by questionnaire design within Thurstonian IRT scoring.

Hyp.	Factor / Contrast	$r(\theta, \widehat{\theta})$	MAB	MSE
H1g	Number of Traits	4	3	3
	Trait correlations	9	7	9
	Block size	3	3	3
	Number of items per trait	0	1	0
	Item keying	43	47	44
	Number of Traits × Item keying	7	5	6
	Trait correlations × Item keying	20	16	19
	Residuals	14	19	15
	Planned Contrasts			
H1a	1/2, 2/3 vs. 0	43	47	44
H1d.1	in mixed, uncorrelated vs. in all positive	20	16	19
H1f.1	many vs. few traits	7	5	6
H1h	2 vs. 3	1	1	1
H1h	3 vs. 4	2	2	2
H1b	Mixed vs. uncorrelated	0	0	0
H1b	Uncorrelated vs. all positive	9	7	9
H1e.1	in Unequal 1	0	0	0
		0	0	0
H1e.1	in Unequal 2	0	0	0
		0	0	0
H1e.2	Unequal 2 vs. Unequal 1 in mixed	0	0	0

Note. Hyp. = Hypothesis, MAB = mean absolute bias, MSE = mean squared error. Main effects are based on the saturated model and are only shown when the associated variance explanation was above 1%. Horizontal lines separate non-orthogonal contrasts.

Table 6: Means and standard deviations for relevant conditions of questionnaire design within Thurstonian IRT scoring.

Factor 1	Factor 2	$r(\theta, \widehat{\theta})$	MAB	MSE
Block size				
2		0.92 (0.05)	0.30 (0.07)	0.15 (0.08)
3		0.91 (0.05)	0.32 (0.07)	0.17 (0.08)
4		0.90 (0.05)	0.33 (0.07)	0.18 (0.09)
Trait correlations	Item keying			
mixed	0	0.90 (0.03)	0.34 (0.04)	0.18 (0.05)
	1/2	0.94 (0.02)	0.28 (0.04)	0.12 (0.03)
	2/3	0.93 (0.02)	0.28 (0.04)	0.13 (0.03)
positive	0	0.80 (0.05)	0.47 (0.05)	0.36 (0.07)
	1/2	0.94 (0.02)	0.28 (0.04)	0.12 (0.03)
	2/3	0.93 (0.02)	0.28 (0.04)	0.13 (0.03)
uncorrelated	0	0.89 (0.03)	0.35 (0.05)	0.20 (0.06)
	1/2	0.93 (0.02)	0.29 (0.04)	0.13 (0.04)
	2/3	0.93 (0.02)	0.29 (0.04)	0.13 (0.04)
Number of Traits	Item keying			
5	0	0.82 (0.06)	0.45 (0.07)	0.33 (0.10)
	1/2	0.93 (0.02)	0.28 (0.04)	0.13 (0.04)
	2/3	0.93 (0.02)	0.28 (0.04)	0.13 (0.04)
15	0	0.88 (0.05)	0.37 (0.07)	0.22 (0.08)
	1/2	0.93 (0.02)	0.28 (0.04)	0.13 (0.04)
	2/3	0.93 (0.02)	0.28 (0.04)	0.13 (0.03)
Number of items per trait	Trait correlations			
Equal	mixed	0.92 (0.02)	0.31 (0.04)	0.15 (0.04)
	positive	0.89 (0.07)	0.35 (0.10)	0.21 (0.12)
	uncorrelated	0.92 (0.03)	0.32 (0.04)	0.16 (0.04)
Unequal 1	mixed	0.93 (0.03)	0.29 (0.05)	0.14 (0.05)
	positive	0.89 (0.07)	0.34 (0.10)	0.20 (0.12)
	uncorrelated	0.92 (0.03)	0.30 (0.06)	0.15 (0.06)
Unequal 2	mixed	0.93 (0.03)	0.29 (0.05)	0.14 (0.05)
	positive	0.89 (0.07)	0.34 (0.10)	0.20 (0.12)
	uncorrelated	0.92 (0.03)	0.30 (0.06)	0.15 (0.06)

Note. MAB = mean absolute bias, MSE = mean squared error. Standard deviations are given in parentheses.

Table 7: Means standard errors for relevant conditions of questionnaire design within Thurstonian IRT scoring.

Factor 1	Factor 2	Low	Medium	High
Block size				
2		0.44	0.37	0.44
3		0.44	0.37	0.44
4		0.44	0.37	0.44
Trait correlations	Item keying			
mixed	0	0.48	0.43	0.48
	1/2	0.39	0.31	0.39
	2/3	0.39	0.31	0.39
positive	0	0.63	0.60	0.63
	1/2	0.39	0.31	0.39
	2/3	0.39	0.31	0.39
uncorrelated	0	0.50	0.44	0.50
	1/2	0.40	0.32	0.40
	2/3	0.40	0.32	0.40
Number of Traits	Item keying			
5	0	0.58	0.54	0.58
	1/2	0.39	0.31	0.39
	2/3	0.40	0.31	0.40
15	0	0.49	0.44	0.49
	1/2	0.39	0.32	0.39
	2/3	0.39	0.32	0.39
Number of items per trait				
Equal (12 items)		0.45	0.38	0.45
Unequal 1 (18 items)		0.40	0.33	0.40
Unequal 2 (9 items)		0.48	0.41	0.48

Note. Low = –2, medium = 0, high = 2, Mean standard errors are given for Trait 2 which was measured with 12, 18, and 9 items in Equal, Unequal 1, and Unequal 2, respectively.

Table 8: Contrasts and % of variance in of sums and differences of traits explained by questionnaire design within Thurstonian IRT scoring

Hyp.	Factor	Sums		Differences	
		MAB	MSE	MAB	MSE
	Number of Traits	1	2	0	0
	Trait correlations	10	13	3	4
	Blocksize	2	1	19	19
	Number of items per trait	0	0	3	2
	Item keying	58	50	6	5
	Number of Traits × Item keying	3	4	0	0
	Trait correlations × Item keying	20	26	0	0
	Residuals	5	3	68	69
H2b	1/2, 2/3 vs. 0	58	50		
	15 vs. 5	3	4		
H2c	0, 1/2 vs. 2/3			3	3
	15 vs. 5			0	0

Note. MAB = mean absolute bias, MSE = mean squared error.

S5, contradicting H2d). To summarize, we found evidence for ipsativity and bias of trait sums and showed that this pertained only to the condition with all positively keyed items. Ipsativity effects were smaller with more traits.

RQ 3. Comparison Between Formats and Scoring Methods

For illustration, Figure 3 depicts the correlation between true and estimated scores for all score types and questionnaire design factors. The score types were ordered as predicted: Recovery was highest in RS (e.g. mean MAB = .17), followed by TF (mean MAB = .26), MFC-IRT (mean MAB = .32) and MFC-CTT (mean MAB = .39; confirming H3a-c; Tables 10 and 11). The difference between TF and MFC-IRT was smaller than the other differences (in favor of H3d). The difference between MFC-CTT and the other score types showed the largest effect.

True reliability was good for RS and acceptable for TF (Table 12). For MFC-IRT and MFC-CTT, with 1/2 and 2/3 mixed comparisons, it was acceptable, but below acceptable with 0 mixed comparisons (Table 12). Reliability varied with an SD of .03 to .05, comparable to TF, for 1/2 and 2/3 mixed comparisons, but with an SD of .08 to .11 with 0 mixed

Table 9: Means and standard deviations for relevant conditions of questionnaire design for sums and differences of traits within Thurstonian IRT scoring.

Factor 1	Factor 2	Sums		Differences	
Number of Traits	Item keying	MAB	MSE	MAB	MSE
5	0	0.82 (0.14)	1.09 (0.39)	0.36 (0.05)	0.21 (0.06)
	1/2	0.40 (0.04)	0.26 (0.05)	0.39 (0.04)	0.25 (0.05)
	2/3	0.40 (0.04)	0.25 (0.05)	0.41 (0.04)	0.27 (0.05)
15	0	0.63 (0.15)	0.65 (0.32)	0.38 (0.05)	0.23 (0.05)
	1/2	0.40 (0.04)	0.26 (0.06)	0.40 (0.04)	0.25 (0.06)
	2/3	0.40 (0.04)	0.25 (0.05)	0.40 (0.04)	0.26 (0.05)

Note. MAB = mean absolute bias, MSE = mean squared error. Standard deviations are given in parentheses.

comparisons. In general, empirical reliability overestimated true reliability, both for MFC and single-stimulus formats. To gain insight into the size of the overestimation, we Fisher Z-transformed true and estimated reliability and classified their difference according to Cohen's (1988) criteria. On average, for MFC-IRT with 0 mixed comparisons, there was a small to medium overestimation. As expected, the overestimation was larger for block size 4 (mean difference in Fisher Z = –.16) than for block size 3 (mean difference in Fisher Z = –.10; Table S6; in favor of H1i). For MFC-CTT with 0 mixed comparisons there was a medium to large overestimation.

For RS-IRT, mean standard errors were .20 for the Trait 2 level of 0 and .25 for Trait 2 levels of ±2. For TF-IRT, they were .28 for 0 and .44 for ±2. For MFC-IRT, with 1/2 and 2/3 mixed comparisons, they were .31 for 0 and .39 for ±2, comparable to TF-IRT. They were higher with all positively keyed items, with .49 for 0 and .54 for ±2.

To summarize, reliability for MFC-IRT with both positively and negatively keyed items was good and close to TF, but lower than for RS. It was clearly lower for MFC-CTT. Empirical reliability overestimated true reliability in conditions with ipsativity and more so with increasing block size.

RQ 4. Questionnaire Design and (Partially) Ipsative Scoring

Trait recovery was worse with 0 (e.g. mean MAB = .49) than with 1/2, and 2/3 mixed comparisons (mean MAB = .34; explaining 48% to 50% of variance, in favor of H4a.2, Table S7). The difference between 1/2 and 2/3 mixed comparisons was negligible (contradicting H4a.1, in favor of H4a.3, Table S8). Some effects only occurred with 0 mixed comparisons:

Table 10: Contrasts and % of variance in summary measures explained by score type and questionnaire design.

Hyp.	Factor	$r(\theta, \widehat{\theta})$	MAB	MSE
	Trait correlations	2	1	2
	Block size	4	6	4
	Item keying	12	8	11
	Score type	41	57	44
	Number of Traits × Item keying	1	1	1
	Trait correlations × Item keying	5	2	5
	Trait correlations × Score type	3	1	3
	Block size × Score type	1	2	1
	Item keying × Score type	12	8	12
	Number of Traits × Item keying × Score type	1	1	1
	Trait correlations × Item keying × Score type	5	3	5
	Residuals	11	10	10
H3a	RS vs. MFC-IRT, TF	13	21	12
H3b	TF vs. MFC-IRT	3	4	2
H3c	RS, TF, MFC-IRT vs. MFC-CTT	26	32	29

Note. Hyp. = Hypothesis. MAB = mean absolute bias, MSE = mean squared error, RS = rating scale format, MFC = multidimensional forced-choice format, TF = true-false format, IRT = item response theory scoring, CTT = classical test theory scoring. Main effects are based on the saturated model and are only shown when the associated variance explanation was above 1%.

Table 11: Means and standard deviations for relevant conditions of score type and questionnaire design.

Factor 1	Factor 2	$r(\theta, \widehat{\theta})$	MAB	MSE
0	MFC-IRT	0.87 (0.06)	0.39 (0.08)	0.25 (0.10)
	MFC-CTT	0.81 (0.08)	0.49 (0.10)	0.38 (0.15)
	RS-IRT	0.98 (0.01)	0.17 (0.04)	0.05 (0.03)
	TF-IRT	0.94 (0.03)	0.26 (0.06)	0.12 (0.05)
1/2	MFC-IRT	0.93 (0.02)	0.28 (0.04)	0.13 (0.04)
	MFC-CTT	0.91 (0.03)	0.34 (0.05)	0.18 (0.05)
	RS-IRT	0.98 (0.01)	0.17 (0.04)	0.05 (0.03)
	TF-IRT	0.94 (0.03)	0.26 (0.06)	0.12 (0.05)
2/3	MFC-IRT	0.93 (0.02)	0.28 (0.04)	0.13 (0.03)
	MFC-CTT	0.91 (0.02)	0.34 (0.04)	0.19 (0.05)
	RS-IRT	0.98 (0.01)	0.17 (0.04)	0.05 (0.03)
	TF-IRT	0.94 (0.03)	0.26 (0.06)	0.12 (0.05)

Note. MAB = mean absolute bias, MSE = mean squared error. MFC = multidimensional forced-choice format, TF = true-false format, IRT = item response theory scoring, CTT = classical test theory scoring. Standard deviations are given in parentheses.

Table 12: True and estimated reliability for relevant conditions of score type and questionnaire design.

Scoring	Number of Traits	Item keying	True Reliability		Estimated Reliability		Difference in Fisher Z	
			Mean	SD	Mean	SD	Mean	SD
MFC-IRT	5	0	0.67	(0.10)	0.79	(0.10)	-0.29	(0.14)
	5	12	0.87	(0.04)	0.88	(0.04)	-0.03	(0.07)
	5	23	0.87	(0.04)	0.88	(0.04)	-0.06	(0.08)
	15	0	0.78	(0.08)	0.83	(0.07)	-0.16	(0.12)
	15	12	0.87	(0.04)	0.88	(0.04)	-0.03	(0.07)
	15	23	0.87	(0.03)	0.88	(0.04)	-0.06	(0.09)
MFC-CTT	5	0	0.59	(0.11)	0.86	(0.04)	-0.63	(0.15)
	5	12	0.84	(0.04)	0.82	(0.05)	0.06	(0.05)
	5	23	0.83	(0.04)	0.84	(0.05)	-0.05	(0.07)
	15	0	0.68	(0.11)	0.83	(0.05)	-0.35	(0.18)
	15	12	0.82	(0.05)	0.83	(0.05)	-0.02	(0.05)
	15	23	0.82	(0.04)	0.84	(0.05)	-0.09	(0.07)
RS-IRT	5	0	0.95	(0.03)	0.95	(0.03)	-0.05	(0.09)
	5	12	0.95	(0.03)	0.96	(0.03)	-0.08	(0.12)
	5	23	0.95	(0.03)	0.96	(0.03)	-0.08	(0.12)
	15	0	0.95	(0.03)	0.95	(0.03)	-0.05	(0.09)
	15	12	0.95	(0.03)	0.96	(0.03)	-0.08	(0.12)
	15	23	0.95	(0.03)	0.96	(0.03)	-0.08	(0.12)
TF-IRT	5	0	0.88	(0.05)	0.88	(0.05)	-0.02	(0.04)
	5	12	0.88	(0.05)	0.88	(0.05)	-0.02	(0.04)
	5	23	0.88	(0.05)	0.88	(0.05)	-0.02	(0.04)
	15	0	0.88	(0.05)	0.89	(0.05)	-0.02	(0.04)
	15	12	0.88	(0.05)	0.89	(0.05)	-0.02	(0.04)
	15	23	0.88	(0.05)	0.89	(0.05)	-0.02	(0.04)

Note. MFC = multidimensional forced-choice format, TF = true-false format, IRT = item response theory scoring, CTT = classical test theory scoring. Standard deviations are given in parentheses.

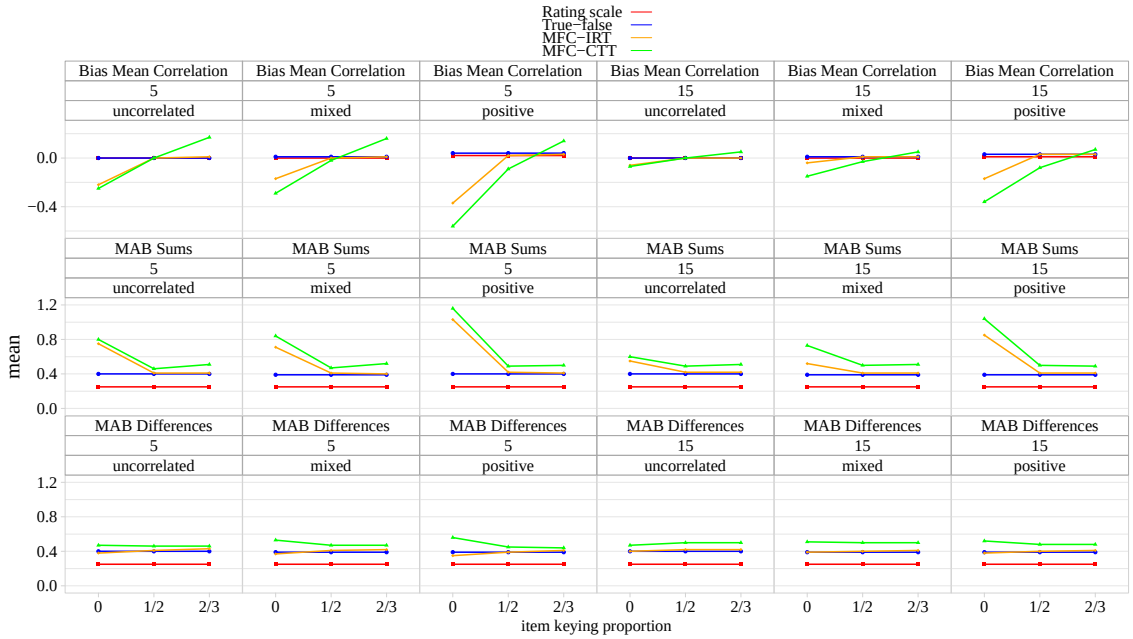


Figure 2: Means of mean trait correlation and of mean absolute bias for sums and differences of two traits. For sums and differences, the results were averaged across the 10 trait pairs. MFC = multidimensional forced choice format; IRT = item response theory scoring; CTT = classical test theory scoring, mixed = mixed positive and negative trait correlations, positive = all positive trait correlations, MAB = mean absolute bias.

First, trait recovery was lower for 5 (e.g. mean MAB = .54) than for 15 traits (mean MAB = .47; in favor of H4.f3). Second, it was lower for all positive trait correlations (e.g. mean MAB = .59) than for mixed correlations or uncorrelated traits (mean MAB = .43; in favor of H4c). Third, with mixed correlations bias was smaller in Unequal 2 than in Unequal 1 (see Tables S7-S9; contradicting H4e.1-3). Overall, trait recovery was higher with uncorrelated traits (e.g. mean MAB = .36) than with all positively correlated traits (mean MAB = .42) or with mixed trait correlations (mean MAB = .38; contradicting H4b). The mean trait correlation was biased as evidenced by a significant intercept, reflecting the grand mean, of –0.07 ($t(161838) = -2406.95$; $p < .001$; 95% CI [–0.658; –0.657]; in favor of H4f.1). Bias in the mean trait correlation was descriptively larger for 5 than for 15 traits, but only for 0 (mean bias for 5 traits = –.35, mean bias for 15 traits = –.19) and 2/3 (mean bias for 5 traits = .16, mean bias for 15 traits = .06) mixed comparisons (contradicting H4f.3). Trait recovery for differences of traits was better than for sums of traits (13% to 14% of total variance, see Table S10; in favor of H4f.2), but this difference was not larger for 5 than for 15 traits (contradicting H4f.4, Tables S10 and S11).

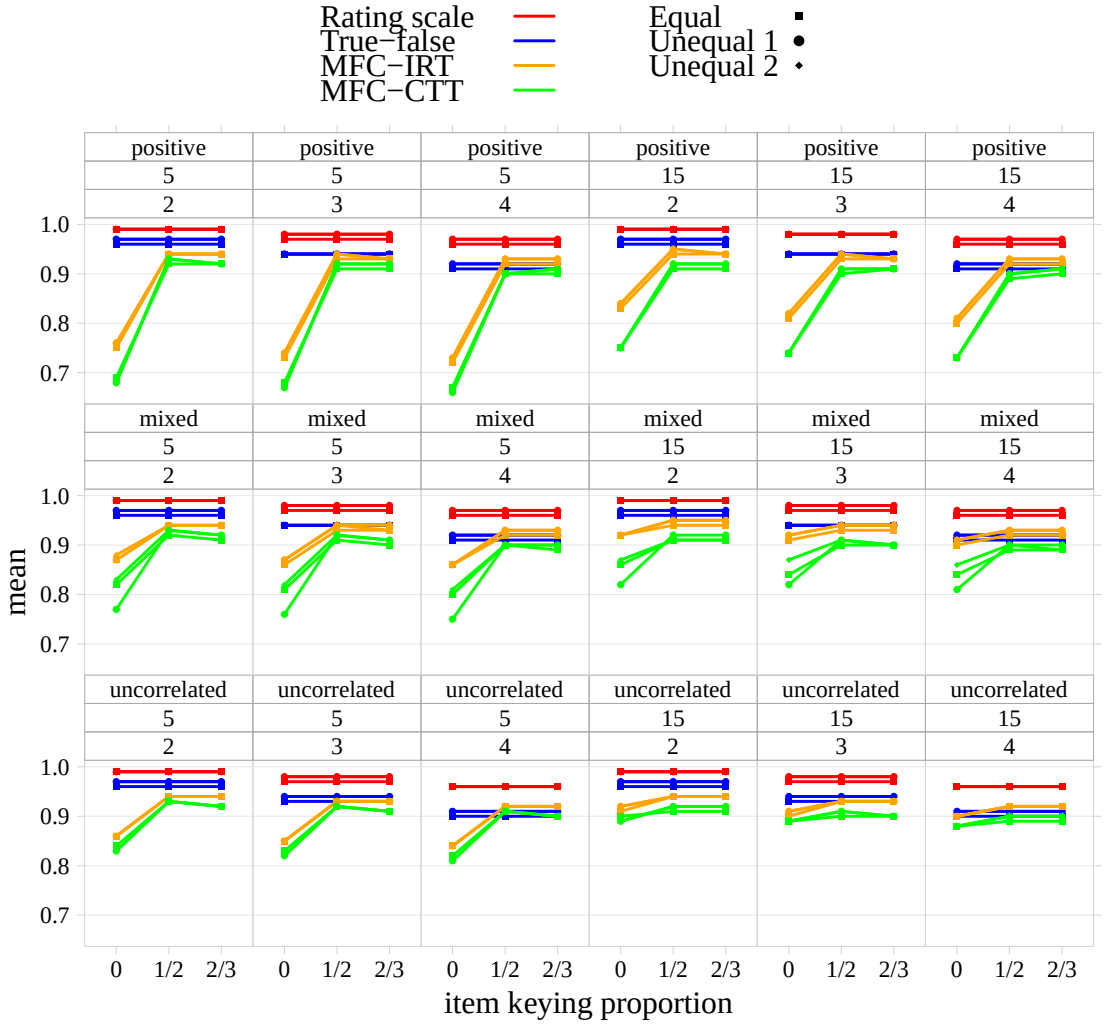


Figure 3: Mean correlation between true and estimated traits (i.e. $r(\theta, \hat{\theta})$) by condition. The results were averaged across the five traits. MFC = multidimensional forced-choice format; IRT = item response theory scoring; CTT = classical test theory scoring, Equal = equal number of items per trait, Unequal 1 (2) = version 1 (2) of unequal numbers of items per trait, mixed = mixed positive and negative trait correlations, positive = all positive trait correlations, 5 = 5 traits, 15 = 15 traits, 2(3,4) = block sizes.

Discussion

In sum, our simulation study showed that Thurstonian IRT trait recovery was acceptable across various questionnaire designs as long as mixed keyed items were used. Thurstonian IRT scoring achieved similar trait recovery as TF, but substantially less effective trait recovery than RS. MFC-CTT trait recovery was clearly worse than the other three and varied more across factors. In the following, we will first discuss the different factors of questionnaire design and then the degree of normativity and the comparison to other response formats and scoring methods. Last, we will discuss the effects of questionnaire design with partially ipsative scoring.

Questionnaire Design and MFC-IRT Scoring

Item Keying. Concerning the effects of questionnaire design on Thurstonian IRT trait estimation, item keying was clearly the most relevant factor, explaining about 40% to 50% of the total variance. Across our analyses, we saw that this was driven by the effect of all positively keyed items. We remind the reader that in our simulation, the positive factor loadings were highly similar, varying in the rather small range of 0.65 to 0.95.

Number of Traits. We found trait recovery to be better with more traits. However, this only pertained to the conditions with all positively keyed items. Trait recovery was acceptable even with as few as five traits as long as mixed keyed items were used, which is in line with previous studies (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2020).

Trait Correlations. Apparently, our conditions of mixed and zero correlations between traits did not differ enough to impact trait recovery differentially. This might have been because our mixed correlations had a mean of approximately zero. In contrast, all positive correlations decreased overall recovery and more so with all positively keyed items. Contrary to our expectations, there was no optimal item keying level depending on trait correlations. This is in contrast to CAT simulations with the Thurstonian IRT model (Brown, 2012), where optimally selected questionnaires for Big Five correlations contained about one third mixed comparisons.

Block size. As expected, we found MFC-IRT trait recovery to slightly decrease with increasing block size, holding the number of pairwise comparisons equal. However, this effect was rather small. Empirical reliability overestimated true reliability and more so with increasing block size, but the overestimation was substantial only with all positively keyed items.

Number of Items per Trait. As expected, a questionnaire design with unequal numbers of items per trait was not detrimental to overall trait estimation. However, we also did not find differential effects of trait correlations depending on how many comparisons with negatively correlated traits the questionnaire included. Apparently, if it exists, this effect was too small to impact recovery within our questionnaire designs and/or to show up in our analyses.

Normativity and MFC-IRT Scoring

With all positively keyed items, the mean trait correlation was biased towards the negative as would be expected from ipsative data (Clemans, 1966; Hicks, 1970). This indicates that the lower recovery with all positively keyed items was a sign of ipsativity. In contrast, with mixed keyed items, bias for the mean trait correlation was small and close to that in TF and RS. Similarly, recovery was comparable to the TF format in these conditions. This illustrates that with mixed keyed items, trait estimates from the Thurstonian IRT model are indeed normative, to at least the same extent as trait estimates from single-stimulus formats. In this study, item keying had only a minor impact on the measurement of trait differences. Thus, trait profiles are generally captured well with comparative data. In contrast, the measurement of sums of traits was clearly impacted by item keying levels such that they were measured worse with all positively keyed items (with loadings of similar magnitudes). Among the different proportions of negatively keyed items, the measurement of sums and of differences of traits was interdependent. However, those differences were small compared to the bias in conditions with all positively keyed items.

Comparison Between Formats and Scoring Methods

Reliability in the RS format was almost perfect. This is in accordance with previous simulation studies without response distortions (e.g. Macdonald & Paunonen, 2002). For MFC, overall reliability levels mirrored ipsativity issues: They were acceptable to good except with all positively keyed items. Recovery levels found in this study for IRT scoring of MFC data are comparable to those found in previous studies (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Hontangas et al., 2015; Morillo et al., 2016). For example, Brown and Maydeu-Olivares (2011) reported a reliability of .86 for a questionnaire with five traits, 20 blocks of three items, and 1/2 mixed comparisons, which is similar to the mean of .86 found in this study in the same condition.

Questionnaire Design and (Partially) Ipsative Scoring

With classically scored MFC responses, there were clearer differences between the item keying levels than with Thurstonian IRT scoring. This probably reflects an interaction

between the scoring procedure and the response process for CTT scoring: From the side of the scoring procedure, the degree of normativity should be strongest when score variability is largest. In CTT scoring, this is achieved when all blocks contain opposite-keyed items (corresponding to 2/3 mixed comparisons). However, the Thurstonian IRT model, which was used to generate the data, favors having both same and mixed keyed comparisons to measure both sums and differences of traits well. The response process is also reflected in the interaction between item keying and trait correlations for CTT scoring: With positively correlated traits, there is more variability in sums than in differences, and this variability is not captured well by mostly equally-keyed item blocks (1/3 mixed comparisons). Regarding normativity, with CTT scoring, the mean trait correlation deviated from the true one across all item keying levels, except for 1/2 of all triplets containing a negatively keyed item – the completely balanced design of comparisons with equally-keyed and opposite-keyed items. In addition, differences (trait profiles) were measured better than sums (absolute trait levels) with CTT scoring.

Empirical Study: Differentiation of Judgments

To complement our simulation study, we conducted an empirical study that investigated how the relative nature of MFC responses contributes to the measurement of individual differences. Following Kahnemann (2011), we assume that comparative judgments as elicited in the MFC format provide more information on the differentiation between behaviors within a person than absolute judgments as elicited in the TF format. Because the two formats are comparable in terms of information with three-item blocks, this should translate to differences in validity. The hypotheses and the design of this study were preregistered on the Open Science Framework (https://osf.io/2673z/?view_only=05ae155a7a5c41f48d2bb4a7a2069c5c).

H1: Big Five latent traits in the MFC format and Big Five latent traits in the TF format will correlate strongly ($r > .50$), but not perfectly ($r <$ reliability level[6]).

H2: Big Five latent traits in the MFC format will show higher convergent validities than Big Five latent traits in the TF format.

H3: Big Five latent traits in the MFC format will show higher criterion-related validities than Big Five latent traits in the TF format.

Instead of exploring all possible correlations for differences between MFC and TF, we tested H2 and H3 with specific relationships between the Big Five and constructs and

[6]Later, we realized that latent correlations should be compared to 1 not to the reliability level, because they are not attenuated by reliability.

Table 13: Study design of the empirical study on differentiation of judgments.

Time 1	MFC first	TF first
	Big Five Triplets MFC	Big Five Triplets TF
	Self-report criteria: employment	
	WHOQOL-BREF	
	2 – 4 weeks	
Time 2		
	Big Five Triplets TF	Big Five Triplets MFC
	SWLS	
	Self-report criteria: social, health, relationships, other	
	CES-D short form	

Note. MFC = multidimensional forced-choice, TF = true-false, WHOQOL-BREF = World Health Organization Quality of Life BREF, SWLS = Satisfaction with Life Scale, CES-D short form = Center for Epidemiologic Studies–Depression Scale.

criteria relevant to personality, which are depicted in Table S12. For example, for number of Facebook friends, we expected a correlation with extraversion but not with neuroticism. Our expectations were based on meta-analyses or studies with large samples. We expected all correlations to be small (.10 to .20) to typical (.20 to .30; Gignac & Szodorai, 2016).

Theoretically, reliability in the MFC format is slightly lower than in the TF format, because latent traits cannot be estimated separately (see Introduction). The comparison of the reliability of MFC trait estimates and TF trait estimates was exploratory.

Methods

Study Design

The data were collected in a within-subject design. We applied the original MFC version of the Big Five Triplets (Wetzel & Frick, 2020) and another version in which the items were presented separately with the response options true and false. Participants filled out the two versions with an interval of at least two weeks between measurement occasions (maximum: 31 days, with 70% at 14 days). They were randomly assigned to begin either with the MFC or the TF version. The criteria and other questionnaires were distributed across the two measurement occasions (see Table 13).

Sample

The data were collected with an online access panel (Prolific Academic; https://www.prolific.co/). Participants were rewarded 0.84 British pounds for each part. We recruited participants from the United States, United Kingdom, and Canada to ensure sufficient language proficiency in English. We recruited 1000 participants to ensure stable model estimation. To achieve a balanced age distribution, we recruited 300 participants between the ages of 18 and 29 and 700 participants between 30 and 65. An additional 18 participants, who had been dropped via Prolific's payment regulations at T1, were mistakenly re-invited to T2. Nine cases (of 1025) and seven cases (of 993) were removed from T1 and T2, respectively, because their response time was less than –2 SD below the mean of their questionnaire group. Due to technical issues, five participants restarted the questionnaire in either T1 or T2. For those, the runs with more complete data were kept. One participant was removed from T1 on request via email. Nineteen participants (of 1018) were removed because they failed either one or both instructed response items, resulting in a final N of 999. Out of those, 491 participants began with the MFC version. Thirty-six participants provided only data for T1 and three only for T2.

Sixty percent were female, 39% male and 1% transgender. The mean age was 37 years (SD = 12 years). As their highest level of education, 13% had completed a high-school diploma, 29% some college, 35% a Bachelor's degree, and 17% some graduate school or higher.

Measures

Big Five Triplets. We used the Big Five Triplets (BFT; Wetzel & Frick, 2020) to assess the Big Five domains neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. This MFC questionnaire consists of 20 blocks of three items (triplets) that are matched for their social desirability. Due to the desirability matching, the number of items per trait is not balanced with the number of items ranging from seven for agreeableness to 16 for neuroticism. To construct the TF questionnaire, we used the same items with the response options true and false, presenting three items per webpage.

Questionnaires. Quality of life was assessed with the World Health Organization Quality of Life BREF (WHOQOL group, 1996), which contains 26 items rated on a five-point scale with varying category labels. A sample item reads: "To what extent do you feel that physical pain prevents you from doing what you need to do?" with scale categories: not at all, a little, a moderate amount, very much, an extreme amount. We excluded the first two items from our analysis, because they represent overall ratings of quality of life and health. Life satisfaction was assessed with the Satisfaction with Life Scale (Diener

et al., 1985), comprising five items rated on a seven-point scale with category labels 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree. A sample item reads: "In most ways my life is close to my ideal." Mental health was assessed with the Center for Epidemiologic Studies–Depression Scale (Cole et al., 2004), comprising ten items. Participants are asked to indicate how often they have felt a certain way during the past week on a four-point scale with scale categories rarely or some of the time (less than 1 day), some or little of the time (1-2 days), occasionally or a moderate amount of time (1-4 days), most or all of the time (5-7 days). A sample item reads: "I was bothered by things that usually don't bother me."

Criteria. The criteria can be grouped into five areas: social, health, relationships, work, and other. Social criteria included Facebook (yes/no) and number of Facebook friends. Health criteria included body mass index (BMI), exercise regularly (at least once a week; yes/no), frequency of drinking (never/$\leq$ once a month/2-4 times a month/2-3 times a week/$\geq$ 4 times a week), and smoking (yes/no). Relationship criteria included duration/begin of relationship (year, month), marriage (yes/no), duration of marriage (marriage date: year, month), divorce (yes/no), time since divorce (divorce date: year, month), and having broken up with a romantic partner within the past 10 years (yes/no). Work criteria included supervising people directly (yes/no), number of supervised people, ability to hire employees (yes/no), ability to fire employees (yes/no), responsibility for a budget (yes/no), and having changed place of employment within the past 10 years (yes/no). Other/uncategorized criteria included charity work (yes/no). Table 14 shows descriptive statistics on the criterion variables.

Analyses

Latent variable models were fit in Mplus (8.2; Muthén & Muthén, 1998-2017). MFC data were modeled with the Thurstonian IRT model and TF data with the two-parameter normal ogive model. Rating scale data (from WHOQOL-BREF, CES-D short form, and SWLS) were modeled with the probit version of the graded response model.

For each construct (life satisfaction, quality of life, depression/mental health), a GRM fitted to the respective questionnaire was combined with either the Thurstonian IRT for MFC or the binary normal ogive model for TF[7]. Similarly, each criterion was regressed on the Big Five from either the Thurstonian IRT for MFC or the binary normal ogive model for TF. Regression coefficients were converted to correlations, i.e. we used regression

[7]Our preregistration indicated that we should fit a joint Thurstonian IRT – normal ogive model. However, we realized that this would be mis-specified (e.g. uncorrelated errors for the same items) and model complexity would bear the danger of estimation problems. Therefore, we decided to estimate separate normal ogive and Thurstonian IRT models.

Table 14: Descriptive Statistics for the Criterion Variables.

Criterion	**Mean**	***SD***	**Min**	**Max**	***N***
Social					
Number of Facebook friends	281.32	392.49	0.00	5000.00	739
Health					
Body mass index	26.48	6.61	13.85	59.17	773
Frequency of drinking alcohol	2.88	1.28	1.00	6.00	868
Frequency of smoking	1.79	1.66	1.00	6.00	868
Relationships					
Months in serious relationship	155.76	128.57	1.00	585.00	583
Months in marriage	163.94	127.97	1.00	533.00	318
Months since divorce	160.07	101.95	2.00	421.00	82
Work					
Number of people supervised	14.19	28.26	1.00	250.00	224

Dichotomous variables	**% No**	**% Yes**	***N***
Social			
Facebook account	19	81	961
Health			
Exercise regularly	35	65	962
Relationships			
Married	63	37	867
Divorced	90	10	864
Broke up with a romantic partner within the past 10 years	58	42	866
Work			
Supervises people	64	36	663
Ability to hire employees	78	22	663
Ability to fire employees	82	18	660
In charge of a budget	70	30	663
Changed place of employment within the past 10 years	34	66	996
Uncategorized/other			
Charity	74	26	962

Note. Number of Facebook friends, body mass index and variables measuring time were log-transformed prior to analysis. Health and relationship criteria were erroneously not assessed for participants who received the true-false version at T1 and reported having no Facebook account.

coefficients standardized for both variables involved. The difference between Fisher Z-transformed correlations of MFC versus TF latent traits with the construct or criterion was tested in R.

Heteromethod correlations were estimated in a Thurstonian IRT model for MFC where a normal ogive model for TF for one trait at a time was added. Error variances involving the same item were allowed to covary. Empirical reliability was calculated from separate Thurstonian IRT and normal ogive models with standard errors of MAP trait estimates obtained from Mplus.

Results

We allowed two openness items to cross-load on neuroticism to improve model fit for the normal ogive TF model. Those items had a strong content overlap with neuroticism and high modification indices in the original model. The final model fit well according to the RMSEA (RMSEA = .043), though other fit indices indicated a less than acceptable fit (SRMR = .112, CFI = .801). However, note the general limitations of applying arbitrary model fit cut-off criteria to models of personality data (Hopwood & Donnellan, 2010). For the Thurstonian IRT model, we started with the same factor structure (i.e. including the two cross-loadings). Although the Thurstonian IRT model should generally be identified with mixed keyed comparisons, in our questionnaire, comparisons including opposite-keyed items almost exclusively involved neuroticism. If this trait is defined in the opposite direction (i.e. emotional stability), there are only 8/60 (13%) mixed keyed comparisons and all traits are positively correlated. This might be the reason why the Thurstonian IRT model produced a Heywood case. We fixed an additional factor loading for agreeableness and two instead of one residual variance for the first item block. This resulted in a reasonable model fit: RMSEA = .036, SRMR = .081. (We do not report CFI because cutoffs for CFI are not appropriate for MFC because the estimation is based on pairwise outcomes which do not correlate as highly as individual items.) Table S13 displays the standardized factor loadings for both the Thurstonian IRT and the normal ogive model.

Our first analysis investigated the correlations between the Big Five in the MFC format and the Big Five in the TF format. Monotrait-heteromethod correlations ranged from .70 for conscientiousness to .93 for neuroticism, confirming H1. The pattern of intercorrelations between the Big Five within each method, i.e. heterotrait-monomethod correlations was mostly quite similar between the two versions, although some correlations indicated that the measured constructs differed slightly. For example, the correlation between neuroticism and conscientiousness was .28 in the MFC version and –.35 in the TF version (see Figure S1 for the full multitrait-multimethod matrix). The mean intercorrelation within MFC (.07) differed slightly from that in TF (.00), but did not indicate ipsativity.

Next, we added one construct or criterion a time to the Thurstonian IRT or normal

ogive model to compare validity between MFC and TF. Twelve percent of the estimated correlations went in the direction opposite to our prediction for both MFC and TF or were around zero. We excluded these from the data analysis because investigating whether the correlation is larger for MFC or TF is not sensible when either correlation goes in the wrong direction. For example, the frequency of drinking alcohol correlated negatively with neuroticism in both formats. As literature predicts a positive correlation, it is unclear whether a higher or smaller negative correlation would be a sign of higher criterion validity in this case. Table 15 displays correlations for the constructs and criteria that went in the predicted direction together with their differences and test statistics[8]. Table S14 displays the full correlation table. Correlations with constructs ranged from –.74 for neuroticism with quality of life to .81 for neuroticism with depression (both in the TF format). For the constructs, five differences were small and two medium: agreeableness with depression (rMFC = –.08, rTF = –.41, difference in Fisher Z = 0.33) and agreeableness with quality of life (rMFC = .11, rTF = .42, difference in Fisher Z = .30). All indicated a higher correlation for TF, contradicting H2. Correlations with criteria ranged from –.22 for conscientiousness with BMI to .33 for extraversion with number of Facebook friends (both in the MFC format). For the criteria, differences between MFC and TF correlations were negligible except for openness with the ability to fire employees, which correlated higher in MFC than in TF (rMFC= .14; rTF = .04, difference in Fisher Z = .10), though this difference was not significant. Thus, H3 predicting higher criterion validity for MFC was not confirmed. For each construct and criterion, we examined the mean correlation across the Big Five within each version for ipsativity. For fully ipsative trait estimates, the mean correlation with an external criterion is constrained to zero (Clemans, 1966). Overall, the mean correlations did not tend more towards zero in the MFC than in the TF version, indicating no notable ipsativity.

Empirical reliabilities ranged from .67 for agreeableness to .89 for neuroticism (both in the TF format). Differences between empirical reliabilities were mostly small except for neuroticism, for which reliability was higher in the TF format (RelMFC = .83, RelTF = .89, difference in Fisher Z = .23).

[8] Due to a programming mistake, health and relationship criteria were skipped for those who filled out the TF version at T1 and indicated having no Facebook account. For those criteria, we report analyses from the subgroup who indicated having a Facebook account. Analyses with the subgroup who filled out the MFC version at T1 led to the same conclusions regarding the differences between MFC and TF, although there were two small differences in favour of RS.

Table 15: Model-based convergent validity coefficients for MFC and TF with significance tests and effect sizes of the differences.

Criterion	Trait	r MFC	r TF	R^2 MFC	R^2 TF	Z MFC	Z TF	Difference	Size	Est/SE	p-value
CES-D short form	N	0.74	0.81	0.55	0.66	0.96	1.13	-0.07	negligible	-1.50	0.13
	E	-0.22	-0.37	0.05	0.14	-0.22	-0.39	-0.15	small	3.34	$\leq$ .001
	A	-0.08	-0.41	0.01	0.17	-0.08	-0.44	-0.33	medium	7.36	$\leq$ .001
	C	-0.21	-0.30	0.04	0.09	-0.21	-0.31	-0.09	negligible	1.97	0.06
SWLS	N	-0.53	-0.60	0.28	0.37	-0.59	-0.70	-0.08	negligible	1.68	0.10
	E	0.18	0.37	0.03	0.14	0.18	0.39	-0.19	small	-4.18	$\leq$ .001
	O	0.05	0.10	0.00	0.01	0.05	0.10	-0.05	negligible	-1.02	0.24
	A	0.11	0.38	0.01	0.14	0.11	0.40	-0.27	small	-5.97	$\leq$ .001
	C	0.21	0.35	0.04	0.12	0.21	0.36	-0.14	small	-3.03	$\leq$ .001
WHO-QoL BREF	N	-0.65	-0.74	0.43	0.55	-0.78	-0.95	-0.09	negligible	1.90	0.07
	E	0.22	0.40	0.05	0.16	0.22	0.42	-0.18	small	-3.97	$\leq$ .001
	A	0.11	0.42	0.01	0.17	0.11	0.44	-0.30	medium	-6.78	$\leq$ .001
	C	0.35	0.40	0.12	0.16	0.36	0.42	-0.05	negligible	-1.16	0.20

—continued —

Criterion	Trait	r MFC	r TF	R2 MFC	R2 TF	Z MFC	Z TF	Difference	Size	Est/SE	p-value
Frequency of drinking alcohol	C	-0.08	-0.08	0.01	0.01	-0.08	-0.08	0.01	negligible	-0.16	0.39
Body mass index	C	-0.22	-0.16	0.05	0.03	-0.22	-0.16	0.06	negligible	-1.02	0.24
Broke up with a romantic partner within the past 10 years	N	0.08	0.05	0.01	0.00	0.08	0.05	0.03	negligible	0.53	0.35
	C	-0.08	-0.07	0.01	0.01	-0.08	-0.07	0.01	negligible	-0.26	0.39
Divorced	N	0.02	0.03	0.00	0.00	0.02	0.03	0.00	negligible	-0.08	0.40
	C	-0.04	-0.06	0.00	0.00	-0.04	-0.06	-0.02	negligible	0.43	0.36
Exercise regularly	E	0.07	0.12	0.00	0.01	0.07	0.12	-0.05	negligible	-0.99	0.25
	C	0.05	0.13	0.00	0.02	0.05	0.13	-0.08	negligible	-1.66	0.10
Time since divorce	A	-0.04	-0.07	0.00	0.00	-0.04	-0.07	-0.02	negligible	0.13	0.40
Time in serious relationship	N	-0.18	-0.19	0.03	0.04	-0.18	-0.19	0.00	negligible	0.07	0.40

—continued —

Criterion	Trait	r MFC	r TF	R2 MFC	R2 TF	Z MFC	Z TF	Difference	Size	Est/SE	p-value
	C	0.12	0.14	0.01	0.02	0.12	0.14	-0.03	negligible	-0.42	0.37
Married	N	-0.17	-0.16	0.03	0.03	-0.17	-0.16	0.02	negligible	-0.30	0.38
	C	0.16	0.13	0.03	0.02	0.16	0.13	0.03	negligible	0.59	0.34
Responsible for a budget	N	-0.08	-0.12	0.01	0.02	-0.08	-0.13	-0.04	negligible	0.76	0.30
	E	0.06	0.08	0.00	0.01	0.06	0.08	-0.02	negligible	-0.38	0.37
	O	0.07	0.08	0.01	0.01	0.07	0.08	0.00	negligible	-0.07	0.40
	A	0.01	0.05	0.00	0.00	0.01	0.05	-0.03	negligible	-0.61	0.33
	C	0.04	0.08	0.00	0.01	0.04	0.08	-0.04	negligible	-0.68	0.32
Charity work	C	0.04	0.05	0.00	0.00	0.04	0.05	-0.01	negligible	-0.22	0.39
Facebook account	E	0.14	0.13	0.02	0.02	0.14	0.13	0.01	negligible	0.18	0.39
Ability to fire employees	N	-0.12	-0.10	0.02	0.01	-0.12	-0.10	0.03	negligible	-0.47	0.36
	E	0.08	0.08	0.01	0.01	0.08	0.08	0.00	negligible	-0.04	0.40
	O	0.14	0.04	0.02	0.00	0.14	0.04	0.10	small	1.85	0.07

—continued —

Criterion	Trait	r MFC	r TF	R2 MFC	R2 TF	Z MFC	Z TF	Difference	Size	Est/SE	p-value
	C	0.08	0.02	0.01	0.00	0.08	0.02	0.07	negligible	1.17	0.20
Ability to hire employees	N	-0.14	-0.17	0.02	0.03	-0.14	-0.18	-0.03	negligible	0.54	0.35
	E	0.06	0.08	0.00	0.01	0.06	0.08	-0.02	negligible	-0.34	0.38
	O	0.10	0.05	0.01	0.00	0.10	0.05	0.04	negligible	0.77	0.30
	C	0.04	0.02	0.00	0.00	0.04	0.02	0.02	negligible	0.40	0.37
Number of facebook friends	E	0.33	0.32	0.11	0.10	0.34	0.33	0.01	negligible	0.15	0.39
Changed place of employment within the past 10 years	N	-0.02	-0.02	0.00	0.00	-0.02	-0.02	0.01	negligible	-0.13	0.40
	O	0.14	0.10	0.02	0.01	0.14	0.10	0.04	negligible	0.88	0.27
	C	-0.04	-0.04	0.00	0.00	-0.04	-0.04	0.00	negligible	0.00	0.40
Ability to supervise people at work	N	-0.17	-0.18	0.03	0.03	-0.17	-0.18	-0.01	negligible	0.25	0.39

—continued —

Criterion	Trait	r MFC	r TF	R2 MFC	R2 TF	Z MFC	Z TF	Difference	Size	Est/SE	p-value
	E	0.13	0.16	0.02	0.03	0.13	0.16	-0.03	negligible	-0.59	0.34
	O	0.08	0.05	0.01	0.00	0.08	0.05	0.04	negligible	0.63	0.33
	A	0.01	0.03	0.00	0.00	0.01	0.03	-0.02	negligible	-0.32	0.38
	C	0.08	0.05	0.01	0.00	0.08	0.05	0.04	negligible	0.63	0.33
Number of people supervised at work	N	-0.10	-0.05	0.01	0.00	-0.10	-0.05	0.05	negligible	-0.56	0.34
	E	0.08	0.09	0.01	0.01	0.08	0.09	-0.01	negligible	-0.07	0.40
	C	0.04	0.01	0.00	0.00	0.04	0.01	0.02	negligible	0.22	0.39

Note. MFC = multidimensional forced-choice format. TF = true-false format, 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. Only correlations that went in the predicted direction for both MFC and TF are shown.

Discussion

In sum, the empirical study showed that for the constructs, validities were slightly higher for TF than for MFC whereas for the criteria, there were mostly no differences. Thus, contrary to our expectations, we did not observe higher validity in the MFC than in the TF version. There are some possible explanations for this. First, correlations between constructs assessed with RS and the TF format might be increased by method biases common to absolute responses such as acquiescence or social desirability. Some correlations with constructs, especially for neuroticism, were even higher than might be expected. If TF correlations were inflated by common method bias, the MFC method with smaller but meaningful and still significant correlations actually indicated good validity. Second, we tried to select criteria that could be evaluated more or less objectively and that were predicted by differences between traits, i.e. a combination of high levels on one and low on another trait would be predictive (e.g. high conscientiousness and low neuroticism predicting relationship/marriage duration). However, from previous research it is unclear whether the criteria we selected truly value differentiation, or whether high levels on one trait can be compensated for by low levels on another trait. In the latter case, sums would actually be predictive. Third, Baron (1996) argued that the MFC format should result in greater differentiation between traits because they facilitate direct comparisons between indicator behaviors. However, this might not happen in all cases. Participants sometimes report that multiple items describe them equally well or badly, i.e., their utility is subjectively identical (Bartram & Brown, 2003; Sass et al., 2020). This could either foster deeper retrieval or facilitate random responding, thereby diminishing validity. Moreover, according to a recent study on the response process in the MFC format (Sass et al., 2020), sometimes participants first evaluate the items in a block in absolute terms without proceeding to more differentiated comparisons.

Besides the comparison of MFC and true-false, we observed some of correlations that went into directions opposite to what would be expected from the literature. For example, the frequency of smoking correlated positively with agreeableness in the true-false version. This might have been due to specifics of our questionnaire or the sample. For example, in another study using the same questionnaire and a younger sample, the frequency of smoking did not correlate significantly with agreeableness, both in a rating scale and the MFC version (Wetzel & Frick, 2020).

Empirical reliabilities were smaller than would be expected from the simulation study. However, they were mostly similar between MFC and TF, indicating that the amount of systematic or unsystematic error might be comparable.

General Discussion

There is increasing interest in applying the MFC format as an alternative to the rating scale format. The Thurstonian IRT model has emerged as the most popular choice for its analysis. However, previous simulation studies investigating Thurstonian IRT trait recovery (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019) were limited in terms of conditions and replications. A key challenge in analyzing MFC data is to derive normative trait estimates that are comparable between persons. The aim of our simulation study was to investigate important aspects of normativity under realistic conditions. We found that Thurstonian IRT model scoring resulted in normative trait estimates with mixed keyed items, and was only marginally affected by the exact proportion of mixed keyed items, unbalanced numbers of items per trait, positive trait correlations, number of traits and block size. With all positively keyed items, Thurstonian IRT trait estimates showed some properties of ipsative data. For normative trait estimates, recovery was similar to TF, but lower than RS, which can be improved with longer MFC questionnaires. Bias of trait correlations indicated that partially ipsative CTT trait estimates retained ipsative properties in contrast to Thurstonian IRT trait estimates.

To gain insight into whether the relative judgment process underlying MFC responses provides a higher level of differentiation, we conducted an empirical study, which compared construct and criterion validity between the MFC and the TF format. Convergent validity coefficients (with external constructs measured by RS) were generally lower in MFC than TF, and criterion validities were generally the same. Moreover, we observed slight changes of constructs. In the following, we discuss the effects of item keying on normativity, the role of trait correlations, to what extent the MFC format facilitates deeper differentiation between attributes, and the level of reliability compared to other response formats.

Normativity and Effects of Item Keying

In our simulation study, we observed ipsativity for questionnaires with all positively keyed items. In previous simulation studies (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019) this could have been attributed to empirically non-identified models, manifesting in biased item parameters and trait correlations (Brown & Maydeu-Olivares, 2012). However, in this study, we fixed item parameters and trait correlations to their true values, mimicking the use of values obtained from single-stimulus data in operational assessment. Our results show that this procedure was not sufficient to overcome ipsativity with all positively keyed items. Apparently, the differences in loadings, which are required for identification of the scale origin (Brown, 2016a), were not sufficiently pronounced. To illustrate, we simulated response probabilities for two persons with identical profile shapes (differences between traits) but different total scores (sums of traits). Response probabilities for those two

persons differed by less than .10 for all 60 simulated item pairs in the MFC format when all items were positively keyed. Thus, persons with similar profile shapes can hardly be distinguished by their response probabilities in an MFC questionnaire measuring five traits with all positively keyed items. Possible solutions include increasing the range of loadings (though this would also affect item information) or introducing distractor items that have zero loadings on the measured traits but match in terms of social desirability. For instance, mixed keyed comparisons or different trait profiles led to more pronounced differences in response probabilities. Future studies may investigate where the best balance lies. Further, as model identification issues in the empirical application showed, applied researchers should consider item keying under all possible trait directions.

From an empirical perspective, fixing parameters to values obtained from single-stimulus response formats bears the danger of masking effects of the MFC format, such as changes in item parameters depending on how items are assembled to blocks (Lin & Brown, 2017). Luckily, for applied researchers, our results show that this procedure yields no benefit in terms of normativity. An MFC questionnaire measuring a few traits with all positively keyed items is just not recommended – regardless how it is scored. Moreover, classical scoring is not recommended, because those trait estimates remain ipsative, regardless of item keying. The only questionnaire design allowing non-biased results with CTT scoring – all uncorrelated traits and half of comparisons between opposite-keyed items – is difficult to realize in practice. In contrast, with mixed keyed items, the Thurstonian IRT model allows deriving trait estimates that are normative, to at least the same extent as trait estimates from single-stimulus formats.

Trait Correlations, Number of Traits and Number of Items per Trait

Our simulation showed that designing an MFC questionnaire in which all traits correlate positively and/or measuring few traits can decrease the quality of recovery of true scores with all positively keyed items, but only slightly with mixed keyed items. To our knowledge, this simulation study was the first to investigate the effect of designing MFC questionnaires with unequal numbers of items per trait. This was not detrimental to person score recovery. Presumably, if the questionnaire includes mixed keyed comparisons, trait estimation might be relatively insensitive to other questionnaire design factors. Thus, according to our simulation results, researchers and practitioners designing MFC questionnaires should ensure that at least some item blocks include both positively and negatively keyed items. As long as this condition is met, unequal numbers of items per trait, positive trait correlations and few traits will probably not be detrimental to trait recovery.

Block size

Our simulation was one of the first to vary block size for the Thurstonian IRT model systematically (see also Brown & Maydeu-Olivares, 2011). The results showed that, holding the number of pairwise comparisons constant, trait recovery decreased with larger blocks, but only to a small extent. However, in comparison to presenting the same number of items in a true-false format, the amount of information was still larger for block size four. Moreover, we found that empirical reliability overestimated true reliability and more so with increasing block sizes. This is in accordance with previous simulations varying block size, though trait recovery was only examined in one replication there (Brown & Maydeu-Olivares, 2011). When the Thurstonian IRT model is applied to empirical data, researchers and practitioners should bear in mind that true reliability is probably slightly lower than the estimate for block sizes > 2. However, the overestimation was especially pronounced for all positively keyed items. With mixed keyed items it is probably negligible for practical purposes.

MFC Responses and Differentiation Between Stimuli

In model-based scoring of MFC data, absolute trait standings are derived from relative item comparisons (Brown & Maydeu-Olivares, 2018a). Specifically, all response process models proposed so far, dominance or ideal point alike, can be expressed in terms of pairwise item utility differences (Brown, 2016a). Predictably, simulation studies implementing other response process and analysis models (Hontangas et al., 2015, 2016; Morillo et al., 2016) showed the same results for item keying as the Thurstonian IRT model supporting the notion that this is a fundamental property of comparative data, not a property of the Thurstonian IRT model (Brown, 2016a). To gain detailed insight into this issue, in this study, item keying was varied and bias for sums and differences of traits was computed. Our results showed that differences were captured well with the MFC response format across all conditions, but sums only with mixed keyed comparisons included.

We also looked at the relative nature of responses within an empirical study. We found no evidence of higher predictive validity for the relative MFC responses as compared to the absolute TF responses. It is likely that the MFC format, with its better measurement of trait differences, shows the largest advantage for criteria that are predicted by differences between traits. Previous studies on validity using normative scoring usually observed similar criterion validity for MFC as for RS (Wetzel, Roberts, et al., 2016; Wetzel & Frick, 2020; Zhang et al., 2019) or TF data (Wetzel, Roberts, et al., 2016). Results are mixed for ipsative scoring with a meta-analysis showing higher criterion validities for partially ipsative trait estimates than for RS data (Salgado & Táuriz, 2014). As our simulation shows, trait differences are measured better than trait sums with such scoring. This suggests that there might be contexts in which trait differences are more predictive than trait sums, for

example, when criteria are more specific to individual traits. A study of organizational 360-degree appraisals (Brown et al., 2017) found that the MFC format consistently increased validity, as measured by inter-rater agreement between self- and others' appraisals. Future research could investigate whether the benefits of normative scoring of MFC data emerge more clearly in high-stakes contexts and/or where social desirability is a concern (Guenole et al., 2018), or when response biases are more present such as in cross-cultural research.

Further, we observed lower convergent validity with similar constructs (all measured by RS) for MFC than for TF. Previous studies observed lower convergent validity for MFC compared to RS when there was common method bias on the side of RS (Lee et al., 2018; Wetzel, Roberts, et al., 2016; Wetzel & Frick, 2020). The same might be true for our study. Higher convergent validities were observed for MFC as compared to RS in studies when the constructs were measured with the same format (i.e. MFC-MFC vs. RS-RS; Brown et al., 2017; Wetzel & Frick, 2020). Some authors concluded that the formats measure slightly different constructs (Guenole et al., 2018; Wetzel, Roberts, et al., 2016; Wetzel & Frick, 2020). Our predictions were based on relationships established with absolute judgment data. If the measured constructs change their meaning with response format, which is likely given the prevalence of format-specific response biases (Wetzel, Böhnke, et al., 2016), we do not know what relationships to expect and might have missed out on some.

Comparing Recovery of True Scores Across Response Formats and Scoring Methods

For normative questionnaire designs, recovery of true scores in MFC-IRT was clearly lower than in RS, but only slightly lower than in TF. This is attributed to the amount of information: We kept the number of pairwise comparisons constant across block sizes so that it was equal to the true-false format for block size three: With the Thurstonian IRT model, three items (per block) provide three bits of binary information (because each pairwise comparison has one threshold). The same items presented separately with a five-point RS yield 12 bits of binary information (four thresholds per item). For block size two, the amount of information was higher and for block size four lower in the TF format than in MFC. When the number of items was duplicated, reliability was good with MFC-IRT scoring (see Table S2 and Footnote 4). Similarly, other studies found trait estimation to improve with longer questionnaires, larger blocks, and more informative ranking instructions (Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015, 2016; Morillo et al., 2016). Thus, when applying MFC questionnaires, researchers should bear in mind that precision is generally lower than with RS questionnaires. When constructing new MFC questionnaires, precision can be increased through more binary comparisons as with longer questionnaires or larger blocks, though the expected increase is not linear because dependencies between

pairwise comparisons also increase, as can be seen from our simulation results (see also Yousfi, 2019). However, more comparisons might go along with higher cognitive load and decreased test motivation – though no support was found for the latter in one study (Sass et al., 2020). Alternatively, one can combine absolute and relative processes with using graded comparisons (Brown & Maydeu-Olivares, 2018b) or a percentage-of-total format (Brown, 2016b).

Limitations and Future Research Directions

We analyzed our simulation results with the condition yielding ipsative estimates always included, conforming to our preregistration. Future research using statistical analyses of simulation results could examine Thurstonian IRT trait estimation only including normative questionnaire designs (Frick, 2017).

In our empirical study, we encountered issues with Thurstonian IRT model identification, similar to reports from other authors (Bürkner et al., 2019; Guenole et al., 2018). When researchers wish to estimate all MFC parameters freely or single-stimulus data are not available, guidelines on how to cope with model identification issues in Thurstonian IRT would be helpful. Using Bayesian estimation procedures might help to identify otherwise problematic models (Bürkner et al., 2019). Part of our model identification problems might be due to only 13% of comparisons between opposite-keyed items in our questionnaire when the direction of neuroticism was reversed. Thus, future questionnaire construction should consider item keying under different definitions of trait direction.

This simulation study aimed at discovering maximal precision of trait estimation when no response distortion was present. Thus far, research comparing the MFC and the RS format based on normative trait estimates under conditions that elicit response distortions is scarce. Any absolute judgements are open to systematic biases influencing all responses, and these general factors are often difficult to separate from the true scores, thus artificially inflating reliability and potentially validity. However, more research is needed to illuminate this question. Moreover, we simulated optimal questionnaires, i.e. with high factor loadings,– in contrast to other recent simulation studies (Bürkner et al., 2019; Schulte et al., 2020). Future research could examine how wider ranges of factor loadings and varying sample sizes might interact with ipsativity and the questionnaire design factors specific to our simulation study, namely number of items per trait and block size.

In this simulation, we compared trait recovery across different block sizes holding the number of pairwise comparisons constant. This allowed us to gain insight into the effect of local dependencies. However, the number of items changed between the block sizes. To examine the effect of designing MFC questionnaires with different block sizes from the same item pool, future research could hold the number of items instead of pairwise comparisons constant, though this changes the amount of information. Moreover, there

is little empirical research on the effect of different block sizes on participants' response processes (Sass et al., 2020) and on the extent of item context effects.

Both in our simulation and in the empirical study, we compared pure MFC with pure single-stimulus designs. Future research could include comparisons with a graded-response format (Brown & Maydeu-Olivares, 2018b) or percentage-of-total formats (Brown, 2016b). Further, the effect of different ranking instructions on trait recovery and on validity, fakability, response processes and item context effects have not been examined thoroughly.

Conclusion

In general, trait estimates from the Thurstonian IRT model were normative in contrast to trait estimates from CTT scoring. Precision was comparable to the true-false but lower than the rating scale format. With all positively keyed items and positively correlated traits, Thurstonian IRT trait estimates displayed ipsative properties despite using true item parameters and trait correlations for their estimation. Nevertheless, as long as item keys were mixed, normative trait estimates could be derived and other questionnaire design factors were less important. Comparing construct and criterion validities between the multidimensional forced-choice and the true-false format showed that direction and size of validity coefficients to expect may depend on the response format. It is possible that criteria that value differentiation or contexts where biases are more pronounced would be needed for the MFC format to show its advantages.

References

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*(1), 49–56. https://doi.org/10.1111/j.2044-8325.1996.tb00599.x

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

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of marketing research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Brown, A. (2012). *Multidimensional CAT in non-cognitive assessments.* Conference of the International Test Comission, Amsterdam.

Brown, A. (2016a). 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. (2016b). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research, 51*(2–3), 345–356. https://doi.org/10.1080/00273171.2016.1150152

Brown, A., & Bartram, D. (2009). *OPQ32r Technical Manual.* SHL group.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-Degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121–148. https://doi.org/10.1177/1094428116668036

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

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (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 (Hrsg.), *The Wiley Handbook of Psychometric Testing (S. 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 Monograph No. 14). Psychometric Society.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

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

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456–473. https://doi.org/10.1037/h0057173

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology, 24*(4), 349–354. https://doi.org/10.1037/h0047358

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

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*(1), 20–30. https://doi.org/10.1027//1015-5759.16.1.20

Evers, A., Hagemeister, C., Høstm, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests—Test review form and notes for reviewers (version 4.2.6).*

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

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression, Third Edition* (Version 3.0-3) [Computer software]. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Frick, S. (2017). *Deriving normative trait estimates from multidimentional forced-choice data—A simulation study.* Unpublished Bachelor Thesis.

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *The Journal of Marketing Management, 9*(3), 114–123. https://doi.org/10.1007/s11336-009-9141-0

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2019). *mvtnorm: Multivariate Normal and t Distributions* (Version 1.0-11) [Computer software]. http://CRAN.R-project.org/package=mvtnorm

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of thurstonian item response modeling. *Assessment, 25*(4), 513–526. https://doi.org/10.1177/1073191116641181

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

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

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

Hopwood, C. J., & Donnellan, M. B. (2010). How Should the Internal Structure of Personality Inventories Be Evaluated? *Personality and Social Psychology Review, 14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*(2), 153–162. https://doi.org/10.1111/j.2044-8325.1988.tb00279.x

Kahnemann, D. (2011). *Thinking fast and slow.* Farrar, Straus and Giroux.

King, M. F., & Bruner, G., C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing, 17*(2), 79–103. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0

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

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

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921–943. https://doi.org/10.1177/0013164402238082

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

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

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*, 1–14. https://doi.org/10.1080/00223891.2020.1739056

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Hrsg.), *Measures of Personality and Social Psychological Attitudes* (S. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

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

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (1999). *An occupational information system for the 21st century: The development of O*NET* (S. xii, 336). American Psychological Association. https://doi.org/10.1037/10313-000

Revelle, W. (2019). *psych: Procedures for Personality and Psychological Research* (Version 1.8.12) [Computer software]. https://CRAN.R-project.org/package=psych

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

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

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*(1), 1–13. https://doi.org/10.1007/BF02288894

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*, 107319111984358. https://doi.org/10.1177/1073191119843585

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong & D. Iliescu (Hrsg.), *The ITC international handbook of testing and assessment* (S. 349–363). Oxford University Press.

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.

Yousfi, S. (2019). *Person parameter estimation for IRT models of forced-choice data—Mertis and perils of pseudo-likelihood approaches* [Manuscript submitted for publication].

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*, 109442811983648. https://doi.org/10.1177/1094428119836486

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

Susanne Frick[1] *

[1]Department of Psychology, School of Social Sciences, University of Mannheim

This is the accepted version of the manuscript which was published in Psychometrika.

*This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group „Statistical Modeling in Psychology“ (SMiP). The author acknowledges support by the state of Baden-Württemberg through bwHPC. The author would like to thank Eunike Wetzel, Thorsten Meiser and three anonymous reviewers for their helpful comments and discussions on previous versions of this manuscript, Safir Yousfi for sharing his R code and Jane Zagorski for proofreading.

Abstract

The multidimensional forced-choice (MFC) format has been proposed to reduce faking because items within blocks can be matched on desirability. However, the desirability of individual items might not transfer to the item blocks. The aim of this paper is to propose a mixture item response theory model for faking in the MFC format that allows to estimate the fakability of MFC blocks, termed the Faking Mixture model. Given current computing capabilities, within-subject data from both high- and low-stakes contexts are needed to estimate the model. A simulation showed good parameter recovery under various conditions. An empirical validation showed that matching was necessary but not sufficient to create an MFC questionnaire that can reduce faking. The Faking Mixture model can be used to reduce fakability during test construction.

Keywords: multidimensional forced-choice, faking, item response theory, mixture model

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

In many personality assessment contexts, individuals are motivated to obtain certain results. For example, when personality assessment is used in personnel selection, individuals may be motivated to distort their responses to make a favorable impression and get a job offer. In clinical assessment, individuals may be motivated to distort their responses to obtain a certain diagnosis or compensation. Such distortions are called *faking*. Although there is no agreed-upon definition of faking, the consensus is that it is a motivated behavior that occurs as an interaction between a person and a situation, and results in leaving an inaccurate or enhanced impression (MacCann et al., 2011). Thus, faking must be distinguished from unintentional response distortions such as self-deceptive enhancement (Paulhus, 2002), careless responding (Curran, 2016; Meade & Craig, 2012) or response styles (Paulhus, 1991; Wetzel et al., 2016). Most assessments are currently conducted with rating scale questionnaires. Faking seems to be quite prevalent on rating scales resulting in increases of .1 to .6 SD in trait estimates (Birkeland et al., 2006; Viswesvaran & Ones, 1999) in real or simulated high-stakes situations.

The Multidimensional Forced-Choice Format as a Remedy

The multidimensional forced-choice (MFC) format has been proposed as an alternative to rating scales in order to prevent faking. In the MFC format, several items measuring different attributes are combined into blocks. People have to rank order the items within a block according to how well the items describe them. There are other variants, for example, selecting the items that describe oneself most and/or least (for an overview of different variants of the MFC format and how to model them, see Brown & Maydeu-Olivares, 2018a). In order to prevent faking in the MFC format, items with equal desirability are combined into blocks, such that all rank orders of items are equally desirable. In this case, items cannot be ranked by their desirabilities. This implies that estimates of item desirability are accurate at the group level and individuals will not differ in their evaluations of item desirability. By contrast, for rating scales, it is often theoretically obvious that higher (or lower) response options are indicative of desirable behaviors or traits.

Overall, using the MFC format has resulted in less evidence of faking than when rating scales are used (Cao & Drasgow, 2019). This has been particularly apparent when faking was evaluated on the group level via mean differences between groups that were instructed to respond honestly compared with groups that were instructed to fake (e.g., Christiansen et al., 2005; Jackson et al., 2000). So far, only two studies have examined faking on the individual level by correlating rank orders of individuals between honest and faking conditions (Heggestad et al., 2006; Wetzel et al., 2021). Heggestad et al. (2006) found similar faking for the MFC format and rating scales. Wetzel et al. (2021) found that rank orders changed when faking, for both the MFC format and the rating scales. The finding that MFC questionnaires are still fakable to some extent raises the question of whether item matching can be improved.

Matching and Item Interactions

Indeed, there is evidence that poorly matched MFC blocks elicit higher agreement as to the optimal rank order than closely matched blocks (Hughes et al., 2021). Further, item desirability might differ across contexts. For example, in an instructed faking study, participants reported higher agreeableness when applying for a job as a nurse compared to as a manager (Pauls & Crost, 2005). Thus, if desirability values were obtained with a manager sample, items that are positively keyed towards agreeableness will increase in desirability and items that are negatively keyed will decrease in desirability when evaluated by nurse applicants. A questionnaire that is fake-proof for one sample or assessment context might not necessarily be so for another one.

Current procedures for matching items to MFC blocks are based on the assumption that item desirability is roughly the same when items are answered separately (i.e., in single-stimulus formats) or relative to each other (i.e., in an MFC format). However, item desirability might change or be evaluated in a more differentiated manner within the context of item blocks (Feldman & Corah, 1960; Hofstee, 1970). In line with this idea, some researchers have argued that desirability should be viewed as a property of response options rather than of items (Kuncel & Tellegen, 2009). More generally, several researchers have observed changes in item parameters or slight changes in constructs between single-stimulus and MFC formats (e.g., Guenole et al., 2018; Wetzel & Frick, 2020) and even changes in item parameters within the MFC format, depending on which items were combined into blocks (Lin & Brown, 2017). To improve the construction of fake-proof MFC questionnaires, a method is needed to estimate the *fakability* of each MFC block (i.e., the extent to which it can be faked). However, current methods only allow the fakability of the whole questionnaire to be assessed (i.e., for all blocks simultaneously). Moreover, they make assumptions that might be wrong.

Assessing Fakability

The fakability of a questionnaire is most often evaluated by comparing responses and trait estimates from a real or simulated high-stakes situation with those from a low-stakes situation. Usually, traits in a high-stakes situation are estimated by fixing the item parameters to the estimates obtained in a low-stakes situation (e.g., Wetzel et al., 2021). This approach, termed the *trait shift approach* in the following, can be applied to both rating scales and MFC questionnaires. It shows how estimated traits would change in practice, when the true model was naively applied. From a modeling perspective, the trait shift approach assumes a shift in the response process, similar to many models formulated for response styles (Henninger & Meiser, 2020; Plieninger & Heck, 2018). As a shift model, the trait shift approach assumes that both the content trait and faking influence all responses. Specifically, the shift is weighted by the item loading, and, across all items measuring the same trait, fakers are assumed to increase their trait level to the same extent. These assumptions might be wrong.

Building on the trait shift approach, Pavlov et al. (2019) proposed that faking scores should be regressed on honest scores to model both a tendency to fake and moderation effects on it. One drawback of their approach is that it relies on observed scores, which are inherently less reliable for MFC than for rating scales (Brown & Maydeu-Olivares, 2018b; Frick et al., 2021). Further, none of the current modeling approaches for faking in MFC questionnaires allows the fakability of individual items or blocks to vary. Thus, they allow faking to vary by person, but fakability is the same for all items or blocks. From a basic research perspective, this assumption might be wrong. From a practical perspective, these approaches allow the fakability of the questionnaire to be examined as a whole, but they do not show how to modify the questionnaire to reduce its fakability.

Within item response theory, Böckenholt (2014) proposed the Retrieve-Edit-Select Framework in which responses based on initial retrieval are edited in a certain direction. Leng et al. (2019) proposed the retreive-deceive-transfer model, which incorporates both the retrieval of socially desirable information and the editing of responses. However, these models are formulated for single-stimulus items, for which a desirable direction can clearly be identified. For MFC blocks, a desirable rank-order cannot easily be identified a priori.

To improve the construction of fake-proof MFC questionnaires, an item response model for faking tailored to the MFC format is needed. Such a model would allow researchers to evaluate fakability on the block level and discard or modify blocks accordingly. More generally, it would show which properties describe less fakable blocks. By focusing on the block level, it would reflect the MFC response process better than the current approaches.

Aim

The aim of this paper is to propose a model for faking in the MFC format that allows the fakability of individual MFC blocks to be evaluated considering that the tendency to fake varies between individuals. Hence, the main purpose of the model is to help construct more fake-proof questionnaires.

Effects of Faking on MFC Rank Orders

In a high-stakes situation, respondents do not necessarily fake all items or all traits (Ziegler, 2011). Estimates of how many people fake their answers range from 14% to 40%, with the consensus being that around 25% fake (MacCann et al., 2011). Therefore, it is appropriate to conceptualize the response process in a high-stakes situation as a mixture of two processes: an honest responding process and a faking process. Both processes have implications for the rank orders selected on MFC blocks.

In a high-stakes situation respondents might select the same rank order as in a low-stakes situation for several reasons: (a) they are not motivated to fake, (b) they do not need to fake, because their responses are already desirable, or (c) the MFC block is closely matched (i.e., all items are equally desirable). When facing a closely matched block, respondents might react in several ways: First, respondents might decide to give an honest response (Berkshire, 1958), or second, they might perceive their honest response as more desirable (Gordon, 1951).

Alternatively, in a high-stakes situation, respondents might select rank orders that do not reflect their content trait levels but might instead reflect what they perceive as desirable. In this case, the distribution of rank orders can show how well the items were matched. For a closely matched block, respondents who are motivated to fake might either respond randomly or retrieve more information about the desirabilities of the items. For example, they might evaluate the item desirabilities in a more differentiated manner (Feldman & Corah, 1960; Hofstee, 1970) in order to be able to rank the items on the basis of their desirabilities. To the observer, both choices result in a uniform distribution of rank orders across respondents. In line with this idea, Kuncel and Borneman (2007) used rating scale items that showed bivariate or trivariate response distributions under faking instructions as indicators of faking.

By contrast, respondents facing a poorly matched block will agree on which rank orders are preferable. To the observer, the distribution of rank orders will be skewed. For example, there might be one rank order that is clearly most desirable and therefore has the highest frequency. Or, there might be one item in a block that is most (or least) desirable, leading to higher frequencies of rank orders that favor (or disfavor) this item. Indeed, in one study, agreement about which rank order should be preferred was higher the worse the matching was (Hughes et al., 2021).

The Faking Mixture model

The proposed model for analyzing faking in the MFC format is a mixture model that allows the occurrence of faking to vary by block and by person. Therefore, I will call it the *Faking Mixture model* in the following. The Faking Mixture model can be used to estimate the fakability of blocks, the probabilities of different rank orders when faking and individuals' faking tendencies. Given current computing capabilities and programs, within-subject data from both a low- and a high-stakes situation are needed for model estimation. The current purpose of the model is thus to improve MFC test construction, but not to correct content trait estimates for faking.

Model Formulation

The proposed model is a mixture model: In a high-stakes situation, individuals base their responses *either* on the desirability *or* on the content trait. By distinguishing between two response processes, it makes explicit that not all individuals fake all blocks when in a high-stakes situation. This is in contrast to a shift model, in which both the content trait and faking influence all responses. For simplicity, in the following, I will use the terms *honest* responding for responses that are based on the content trait and *faking* for responses that are based on desirability. However, the assessment situation and the sample both shape, for example, whether honest responding is influenced by socially desirable responding. The model properties are defined by the following model equations.

The Faking Mixture model models the probability of selecting a certain rank order. This is in contrast to most IRT models for MFC data that can be expressed in terms of pairwise preferences (Brown, 2016), for an exception, see Joo et al. (2018). Let k index the item block within the questionnaire and $r = 1 \dots R$ the possible rank orders (i.e., the permutations). For a block of size B, there are $R = B!$ possible rank orders (permutations). For example, if $B = 3$ the possible rank orders are 1-2-3, 1-3-2, 2-1-3, 2-3-1, 3-1-2, and 3-2-1. Further, let F be an indicator variable that takes on a value of 1 if a person fakes and 0 otherwise. Then, the probability that the observed rank order X for person j on block k takes on a value of r can be described as follows:

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

The probability of selecting rank order r for block k when faking, termed the *rank order probability* $P(X_k = r|F_{jk} = 1)$, varies only by block but not by person. To reflect this, the person subscript j is dropped. The rank-order probabilities are linked to continuous,

unconstrained rank-order parameters β_{kr}:

$$P(X_k = r|F_{jk} = 1) = \frac{\exp(\beta_{kr})}{\sum_{u=1}^{R} \exp(\beta_{ku})} \tag{2}$$

For identification, and without loss of generality, the parameter for the first rank order in each block is fixed to zero: $\beta_{k1} \equiv 0$. As a consequence of this indeterminacy, the rank-order parameters β_{kr} cannot be interpreted in absolute terms but only relative to each other. For example, a relatively high rank-order parameter β_{kr} translates into a high probability of selecting this rank order when faking. It is convenient to focus on the rank-order probabilities, because they can be interpreted in absolute terms. If the block is closely matched (i.e., the differences between the item desirabilities are small), the rank-order probabilities will be approximately equal. If matching is poor, the rank-order probabilities will be relatively higher for one or several rank orders. The $R = B!$ rank-order parameters β_{kr} are not linked to the B items; hence, they cannot be expressed in terms of individual item desirabilities. However, in this way, the rank-order probabilities reflect differences in item desirabilities within a block and hence capture item interactions and the relative nature of the MFC responses.

The probability of faking a block depends on the block fakability α_k and a faking trait θ_j modeled via a probit link:

$$P(F_{jk} = 1) = \Phi\left(\theta_j + \alpha_k\right) \tag{3}$$

The block fakability α_k is not an independent parameter but is obtained from the rank-order parameters β_{kr}. It is the quantile of the standard normal distribution at the sum of squares of the rank-order probabilities. The quantile is used to transform the sum of squares from the probability scale into a continuous scale. If $\Phi(x)$ denotes the cumulative standard normal distribution function, evaluated at x, and Φ^{-1} its inverse, then:

$$\alpha_k = \Phi^{-1}\left(\sum_{r=1}^{R} \left(P(X_k = r|F_{jk} = 1) - M\left[\boldsymbol{P(X_k|F_{jk} = 1)}\right]\right)^2\right) \tag{4}$$

Thus, the block fakability α_k, and with it the probability of faking a block, increases with greater variance in the rank-order probabilities. If a preferable ranking can be clearly identified for a block, the block fakability α_k is higher, and respondents are more likely to fake this block. If items within a block cannot be ranked by desirability, the block fakability α_k is lower, and respondents are more likely to respond honestly (i.e., based on the content traits).

The faking trait θ_j captures the propensity to fake, both in terms of the sample mean and interindividual variation around it. Even for closely matched blocks, a high faking

trait leads to high faking probabilities. This can capture the fact that a situation might strongly motivate respondents to fake, even when they differ in which rank orders they perceive as desirable.

The response probabilities when responding honestly $P(X_{jk} = r|F_{jk} = 0)$ follow an IRT model for MFC data. For my applications, I chose the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011), because it is the most broadly applicable model and is based on a linear factor structure, which is commonly assumed for personality questionnaires. In the Thurstonian IRT model, it is assumed that a latent, continuous value called *utility* underlies the responses. For personality questionnaires, the utility reflects how useful the item is for describing the person. The utility t of person j on item i is a linear function of a latent content trait η_j, weighted with an item loading λ_i and having an intercept μ_i and an error term ε_{ji}:

$$t_{ji} = \mu_i + \lambda_i \eta_j + \varepsilon_{ji} \tag{5}$$

The errors of item i are normally distributed with $\boldsymbol{\varepsilon_i} \sim N(0, \psi_i)$.

According to Thurstone's law of comparative judgment (Thurstone, 1927), items within blocks are ranked according to the magnitude of their utilities. Response probabilities for rank orders can be calculated by following the formulation by Yousfi (2018):

$$P(X_{jk} = r|F_{jk} = 0) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty MVN\left(\boldsymbol{A t_{jk}}(r), \boldsymbol{A \psi_k^2}(r)\right) d\boldsymbol{A t}_{jk}(r) \tag{6}$$

Vectors of utilities $\boldsymbol{t_{jk}}$ and of error variances $\boldsymbol{\psi_k^2}$ are sorted in descending order, according to the selected rank order r. Then, the response probability is the area under the multivariate density where the first utility is larger than the second, the second is larger than the third and so forth. This order is ensured by the limits of the integral and the comparison matrix A. For example, for a block of size $B = 3$:

$$\boldsymbol{A}_{B=3} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{7}$$

Within a structural equation framework, Thurstonian IRT models can easily be estimated via limited information methods (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares, 1999; Maydeu-Olivares & Brown, 2010). However, estimating item parameters on the basis of response probabilities for rank orders is hardly feasible with the current computing capabilities and programs that are available to the usual researcher. The multiple integral in Equation (6) can be solved by numerical integration (Genz & Bretz, 2009), which is so far implemented only in R, Matlab, and Fortran. However, this numerical integration is apparently too complex to be used for item parameter estimation as implemented in mirt (Chalmers, 2012), resulting in unreasonable computing times and estimates. Bayesian estimation programs, such as JAGS (Plummer, 2017) or Stan (Carpenter et al.,

2017), do not include such a function.

As a solution, response probabilities when responding honestly are estimated with data that can be assumed to primarily reflect the content traits. For this, item and trait parameters are estimated in a structural equation framework in a first step and the response probabilities are calculated with these parameters. The Faking Mixture model is then estimated in a second step, by using data from the same respondents in a high-stakes situation, with the response probabilities that occur when the respondents answer honestly $P(X_{jk} = r|F_{jk} = 0)$ fixed to those obtained in the first step.

Filling in the mixture Equation (1) with the above specifications gives the response probability under the Faking Mixture model, where the $P(X_{jk} = r|F_{jk} = 0)$ are calculated a priori:

$$P(X_{jk} = r|\theta_j) = \Phi\left(\theta_j + \alpha_k\right) \frac{\exp(\beta_{kr})}{\sum_{u=1}^{R} \exp(\beta_{ku})} + \left(1 - \Phi\left(\theta_j + \alpha_k\right)\right) P(X_{jk} = r|F_{jk} = 0) \quad (8)$$

Implementation

The Faking Mixture model is implemented in a Bayesian framework with the following priors and hyperpriors:

$$\begin{aligned} \boldsymbol{\beta} &\sim N\left(0, SD(\boldsymbol{\beta})\right) \\ \boldsymbol{\theta} &\sim N\left(M(\boldsymbol{\theta}), SD(\boldsymbol{\theta})\right) \\ M(\boldsymbol{\theta}) &\sim N(1, 2) \\ Var(\boldsymbol{\theta}) &\sim Inverse\ Gamma(1.5, 1) \\ Var(\boldsymbol{\beta}) &\sim truncated\ N(5, 10, 0, 15) \end{aligned} \quad (9)$$

Thus, these hyperpriors allow the mean block fakability (determined by Var($\boldsymbol{\beta}$)) and the mean and the variance of the faking trait to be estimated from the data, thereby reducing prior sensitivity (Fox, 2010). In preliminary simulations, the priors and hyperpriors were fine-tuned to ensure model convergence and good recovery under various conditions. The parameters can be sampled from the posterior (e.g., via Hamiltonian Monte Carlo sampling as implemented in Stan; Stan Development Team, 2020b). The Stan Code for estimating the Faking Mixture model can be found at https://osf.io/wfhz4/.

Simulation Study: Parameter Recovery

A small simulation study was conducted to evaluate parameter recovery for the Faking Mixture model. Response probabilities when responding honestly ($P(X_{jk} = r|F_{jk} = 0)$) were based on the true item and trait parameters. 1,000 replications were conducted. The R code used to run and analyze the simulation, along with the simulation materials and

results, can be found at https://osf.io/wfhz4/.

Simulation Design

Responses were simulated for 500 respondents, 20 blocks of three items each, and five traits. The five traits were drawn from a multivariate normal distribution with a mean vector of 0, variances of 1, and correlations set to meta-analytic estimates for the Big Five (neuroticism, extraversion, openness, agreeableness, and conscientiousness) as reported by van der Linden et al. (2010), see Table 1. The faking trait $\boldsymbol{\theta}$ was drawn from a

Table 1: Correlations used in the simulation study

Trait	E	O	A	C
N	−.36	−.17	−.36	−.43
E		.43	.26	.29
O			.21	.20
A				.43

Note. N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness. These are meta-analytic correlations between the Big Five as reported by van der Linden et al., 2010.

normal distribution, independent of the content traits. Note that although this might be unrealistic, it represents a correctly specified model as long as the faking trait and the content traits cannot be estimated at the same time.

Three factors were varied and completely crossed, that is, all possible combinations of the factor levels were realized: fakability, faking trait mean $M(\boldsymbol{\theta})$, and faking trait variance $Var(\boldsymbol{\theta})$. The *fakability* factor was varied with the levels high and low. For high and low fakability, the rank-order parameters β_{kr} were drawn from $U(-4, 4)$ and $U(-2, 2)$, respectively. For low fakability, 90% of the highest rank-order probabilities per block are between .3 and .6 (M = .4, SD = .1). For high fakability, 90% are between .4 and .9 (M = .6, SD = .2). The *faking trait mean* factor was varied with the levels 0, 1, and 2. With these levels, across blocks, the mean probability of faking ranges from .15 for low fakability and $M(\boldsymbol{\theta}) = 0$, mimicking the lower estimate of the propensity to fake from MacCann et al. (2011), to .90 for high fakability and $M(\boldsymbol{\theta}) = 2$, representing the extreme end of faking. The *faking trait variance* factor was varied with the levels 0.2, 0.5, and 1. In relation to the content trait variance of 1, 0.2 is a typical variance for a trait that captures response biases (Billiet & McClendon, 2000; Plieninger & Heck, 2018).

Data Generation

To simulate the data for honest responding, item parameters were drawn from the following distributions: $\boldsymbol{\mu} \sim U(-1, 1)$, $\boldsymbol{\lambda} \sim U(.65, .95)$. These are typical values for standardized item utilities with good measurement properties (Brown & Maydeu-Olivares, 2011). To

ensure that the loadings allow for the recovery of normative trait levels, they were redrawn until there were no linear dependencies between loadings within a block. If there are linear dependencies within a block, for example, because all loadings have approximately the same size or are multiples of each other, the Thurstonian IRT model is not identified (see Brown, 2016). In addition, the direction of factor loadings was set such that half of the pairwise item comparisons within blocks were between differently keyed items (i.e., one negative, one positive factor loading). This has been shown to aid the recovery of normative trait levels in simulation studies (e.g., Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Frick et al., 2021). To simulate standardized utilities, the error variances were set to $\psi_i^2 = 1 - \lambda_i^2$. Errors were drawn from $\boldsymbol{\varepsilon_i} \sim N(0, \psi_i)$. Using Equation (5), utilities were calculated given the traits, item parameters and errors. To obtain the responses when responding honestly, the rank order of utilities within each block was determined for each person.

To simulate the data in a high-stakes situation, the parameters for the Faking Mixture model were drawn according to the respective condition of the simulation design. Next, the probabilities of faking the blocks were calculated using Equation (3), for each person and each block. To determine whether a block was faked by a person, a dichotomous outcome was drawn from a binomial distribution with the respective probability. To obtain the response for a faked block, a categorical outcome was drawn from a multinomial distribution with the rank-order probabilities $P(X_k = r|F_{jk} = 1)$ for this block. To obtain the responses in a high-stakes situation, for each faked block, the response when responding honestly was replaced by the respective categorical outcome.

Response probabilities when responding honestly were calculated with Equation (6) with the true item and trait parameters. Then, the Faking Mixture model was fit to the faking data, given the response probabilities when responding honestly calculated in the previous step. R (R Core Team, 2020) was used for data generation and analysis along with the packages rstan (Stan Development Team, 2020a), MASS (Venables & Ripley, 2002), psych (Revelle, 2019), and mvtnorm (Genz & Bretz, 2009; Genz et al., 2020). Stan (Carpenter et al., 2017) was used for model estimation. Three chains were run with 3,500 iterations out of which the first 750 were discarded. In preliminary simulations, these values were selected to ensure convergence and sufficient parameter recovery.

Data Analysis

The coverage of 95% posterior intervals was computed along with the correlation between true and estimated values, mean bias and variance in bias for posterior medians of (a) the main parameters: rank-order parameters β_{kr} and the faking trait $\boldsymbol{\theta}$, (b) the hyperparameters: faking trait mean $M(\boldsymbol{\theta})$, faking trait variance $Var(\boldsymbol{\theta})$, and rank order parameter variance $Var(\boldsymbol{\beta})$, and (c) the derived parameters: block fakabilities α_k, and rank-order

probabilities $P(X_k = r|F_{jk} = 1)$. Within each condition, I evaluated whether the absolute levels of parameter recovery were satisfactory. Furthermore, for each dependent variable, I calculated the explained variance within an ANOVA framework to see whether recovery varied by the manipulated simulation factors. In contrast to the F-test, explained variance is descriptive and insensitive to heterogeneous variances, which occurred in some of the simulation conditions.

Results and Discussion

For all parameters, the coverage of the 95% posterior intervals was almost perfect across conditions ($\geq .95$) and did not vary systematically with the manipulated factors (Table 2). Mean bias was generally low in relation to the scale of the respective parameter (Table 3). The hyperparameters, faking trait mean $M(\boldsymbol{\theta})$ and variance $Var(\boldsymbol{\theta})$, and the variance of the rank-order parameters $Var(\boldsymbol{\beta})$ were recovered well across the simulated conditions.

Regarding the main parameters, the manipulated factors showed some effects (Table 2). For the faking trait $\boldsymbol{\theta}$, as would be expected for any latent trait, the faking trait variance factor $Var(\boldsymbol{\theta})$ explained the largest amount of variance in the correlations (67%, Table 2) and SDs of the bias (72%): The correlations (see Figure 1) and SDs of the bias increased (with means of 0.34, 0.44, and 0.54) as the faking trait variance $Var(\boldsymbol{\theta})$ increased. Moreover, 18% of the variance in the correlations and 14% in the SDs of the bias were explained by the faking trait mean factor $M(\boldsymbol{\theta})$ (Table 2): The correlations (Figure 1) were highest and the SDs of the bias were lowest for a faking trait mean $M(\boldsymbol{\theta})$ of 1, followed by a means of 0 and 2 (the mean SDs of the bias were: 0.40, 0.43, and 0.48, for $M(\boldsymbol{\theta}) = 1$, 0, and 2, respectively). Such values are to be expected because differences in $\boldsymbol{\theta}$ are harder to detect the more extreme the mean is, because, for a high (low) faking trait mean, most (almost no) individuals are predicted to fake.

For the rank-order parameters β_{kr}, the faking trait mean factor $M(\boldsymbol{\theta})$ explained the largest amount of variance in the correlations (63%, Table 2) and the second largest amount of variance in the SDs of the bias (24%). The correlations (Figure 1) increased and the SDs of the bias decreased (with means of 1.07, 0.78, and 0.63) with higher faking trait means $M(\boldsymbol{\theta})$. Thus, the rank-order parameters are better recovered the more the sample is inclined to fake. There was a systematic negative bias for the rank-order parameters β_{kr}: 94% of the mean biases were < 0. This negative bias emerged by design because the only effect of the β_{kr} is through their transformation into rank-order probabilities $P(X_k = r|F_{jk} = 1)$, which add up to 1. Consequently, there is some dependency even among $k - 1$ free rank-order parameters β_{kr}. If any of the β_{kr} within a block have a positive bias, all others must have a negative bias, leading to a negative mean bias. Bias in absolute terms increased with fakability (with means of -0.13 for low and -0.36 for high

Table 2: Variance explained in % by the manipulated factors in the simulation study

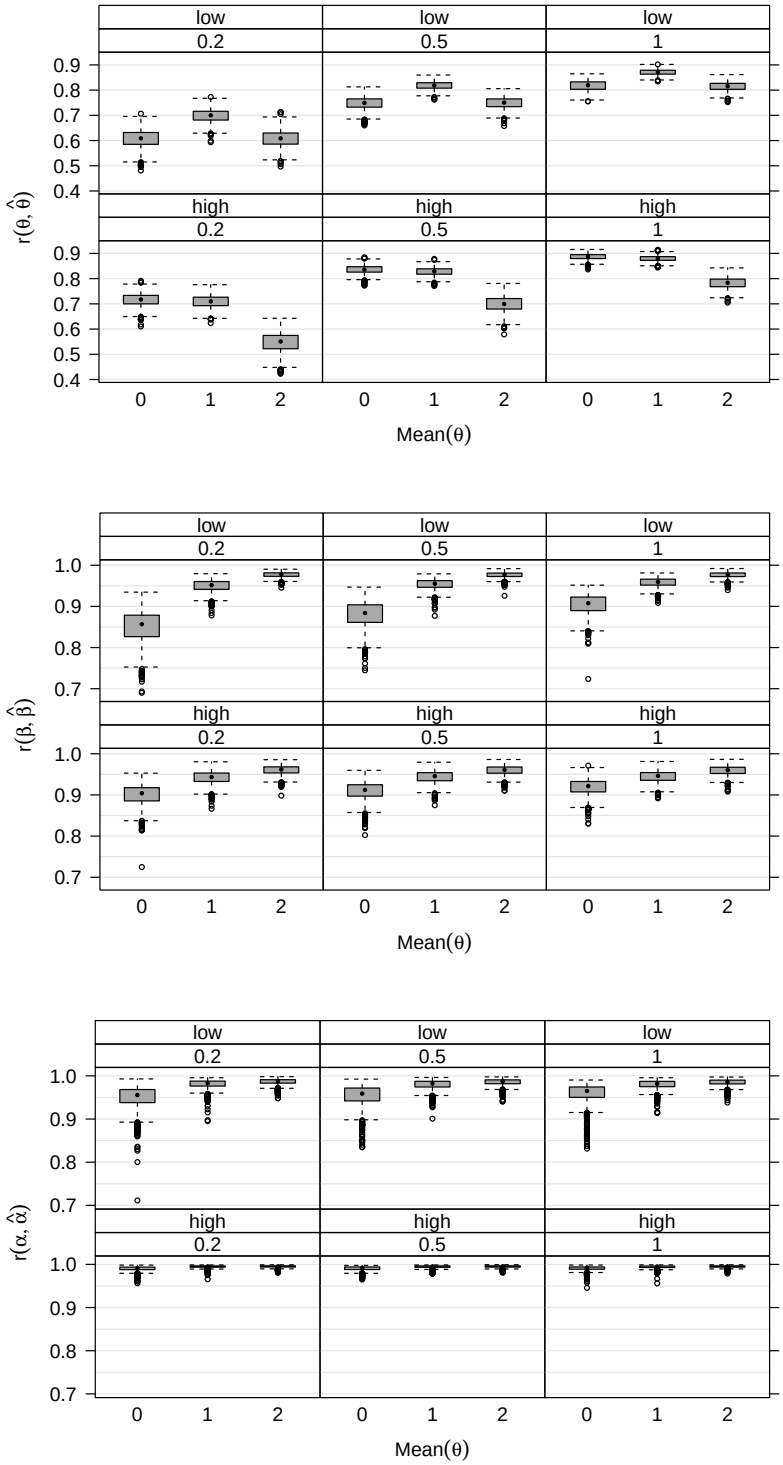
Factor	Main parameters							
	θ_{fj}				β_{kr}			
	r	MB	SDB	95%	r	MB	SDB	95%
$Var(\boldsymbol{\beta})$	1	0	1	0	0	31	63	0
$Mean(\boldsymbol{\theta})$	18	8	14	0	63	4	24	6
$Var(\boldsymbol{\theta})$	67	1	72	3	2	0	0	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})$	8	0	6	0	7	0	0	1
$Var(\boldsymbol{\beta})\times Var(\boldsymbol{\theta})$	0	0	0	0	1	0	0	0
$Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	1	5	2	0	3	0	1	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	0	0	1	0	1	0	0	0
Residuals	6	85	5	96	23	65	13	93

	Hyperparameters					
	$Mean(\boldsymbol{\theta})$		$Var(\boldsymbol{\theta})$		$Var(\boldsymbol{\beta})$	
	B	95%	B	95%	B	95%
$Var(\boldsymbol{\beta})$	0	0	0	0	1	0
$Mean(\boldsymbol{\theta})$	0	0	0	0	4	1
$Var(\boldsymbol{\theta})$	0	1	1	1	0	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})$	0	0	0	0	1	0
$Var(\boldsymbol{\beta})\times Var(\boldsymbol{\theta})$	0	0	0	0	0	0
$Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	0	0	0	0	0	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	0	0	0	0	0	0
Residuals	99	99	99	99	94	99

	Derived parameters							
	α_k				$P(X_k = r \mid F_{jk} = 1)$			
	r	MB	SDB	95%	r	MB	SDB	95%
$Var(\boldsymbol{\beta})$	30	0	19	0	27	5	18	1
$Mean(\boldsymbol{\theta})$	18	0	37	0	32	76	62	1
$Var(\boldsymbol{\theta})$	0	0	0	0	2	2	2	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})$	10	0	6	0	18	7	6	0
$Var(\boldsymbol{\beta})\times Var(\boldsymbol{\theta})$	0	0	0	0	1	1	0	0
$Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	0	0	1	0	3	4	3	0
$Var(\boldsymbol{\beta})\times Mean(\boldsymbol{\theta})\times Var(\boldsymbol{\theta})$	0	0	0	0	2	1	0	0
Residuals	42	99	38	100	16	4	9	98

Note. r = correlation, MB = Mean bias, SDB = SD bias, 95% = coverage of 95% posterior intervals, $Var(\beta)$ = fakability.

Figure 1: Correlations between true and estimated parameters in the simulation study by condition



Note. low = low fakability, high = high fakability; 0.2, 0.5, 1.0 = faking trait variance

Table 3: Mean bias across conditions in the simulation study

Variable	Mean	SD	lower	upper
θ_{fj}	0.00	0.05	-0.09	0.08
β_{kr}	-0.24	0.20	-0.63	0.01
$Mean(\boldsymbol{\theta})$	0.00	0.05	-0.08	0.09
$Var(\boldsymbol{\theta})$	0.01	0.07	-0.10	0.12
$Var(\boldsymbol{\beta})$	0.05	1.02	-1.66	1.84
α_k	0.00	0.03	-0.05	0.04
$P(X_k = r \mid F_{jk} = 1)$	0.00	0.00	0.00	0.00

Note. lower = 5% quantile, upper = 95% quantile, $Var(\beta)$ = fakability.

fakability), explaining the largest amount of variance (31%, Table 2), because fakability was operationalized as variance in the rank-order parameters. Similarly, the SDs of the bias increased with higher fakability (with means of 0.54 and 1.12 for low and high fakability, respectively), explaining 63% of the variance (Table 2). Still, in absolute terms, the bias across conditions was small (Table 3) and the correlations between the true and estimated rank-order parameters were almost perfect (mean of .94, $SD = .04$, Figure 1). The effects for the block fakability parameters α_k mirrored those found for the rank-order parameters β_{kr}. For the rank-order probabilities, differences between the conditions were negligible in size. Additional information on parameter recovery can be found in Figures S1-S16[1].

To sum up, the simulation showed that the parameters of the Faking Mixture model could be recovered well across the simulated conditions. As to be expected, the reliability of the faking trait increased as the variance increased. There was a slight dependency between the recovery of the faking trait and of the rank-order parameters, such that recovery for the latter was better with a higher faking trait mean and vice versa. However, this dependency was small enough to be negligible.

Empirical Validation

To validate the Faking Mixture model and to illustrate its application, I used a dataset that was already analyzed by Wetzel et al. (2021) with the trait shift approach. Participants in this sample filled out an MFC personality questionnaire under both instructions to be honest and instructions to fake good. They were randomly assigned to either a version of the questionnaire in which all items within blocks were matched for desirability or a version in which some blocks were not matched (termed mixed blocks in the following). This dataset allowed the Faking Mixture model to be validated: If the Faking Mixture model works, mixed blocks should have higher fakability parameters than matched blocks.

[1]The supplemental online material is available at https://osf.io/wfhz4

Further, rank-order probabilities for mixed blocks should favor rank orders in which the most desirable item is ranked highest.

The R code and data used in this empirical validation can be found at https://osf.io/wfhz4/.

Method

Sample and Procedure

The sample consisted of two subsamples: one laboratory sample and one sample from an online access panel. In both subsamples, participants were remunerated for their participation and some participants were excluded due to data quality checks (for details see Wetzel & Frick, 2020). The final sample consisted of 1,244 participants. There were $N = 592$ participants in the group with the matched version of the questionnaire, called MFC-matched (M(age) = 23.40, SD(age) = 4.10, 63% female, 0.3% transgender) and $N = 652$ participants in the group with the partly mixed version of the questionnaire, MFC-mixed (M(age) = 23.39, SD(age) = 4.34, 65% female, 0.2% transgender).

The procedure was identical in the two groups. Participants filled out an MFC personality questionnaire first under instructions to be honest. Then, they filled out other personality questionnaires and questions about external criteria. Afterwards, they received instructions to fake good and were asked to fill out the following questionnaire in accordance with the instructions. Then, they filled out the MFC personality questionnaire again. The fake good instructions asked them to imagine they were interested in a place in the Master's program of psychology at a German university and that the following personality questionnaire was part of the application procedure. The instructions detailed which attributes the university was looking for in their students, which amounted to low levels of neuroticism and high levels on the other Big Five Traits (i.e., extraversion, openness, agreeableness and conscientiousness). For a detailed description of the faking instructions, see Wetzel et al. (2021); for more information about the sample and other measures, see Wetzel and Frick (2020).

Measures

The Big Five Triplets (BFT; Wetzel & Frick, 2020), an MFC questionnaire measuring the Big Five traits, were used. The BFT are available in German and English from https://osf.io/ft9ud/. The BFT consist of 20 blocks of three items each. During test construction, the social desirability of the individual items in a larger item pool was rated by 33 psychology students on a 5-point rating scale. Social desirability was defined as fulfilling societal norms and expectations, accompanied by three examples. The blocks of the BFT were matched by these desirability ratings. For example, the first triplet contains items

that were all rated as socially undesirable. The BFT contain three socially undesirable, four neutral, and 13 socially desirable blocks. For the MFC-matched version, the original version of the BFT was used. Hence, the term *matched* in MFC-matched means that the items were matched by their desirability ratings. Whether those blocks are matched such that their true desirabilities are equal is an empirical question examined in the current study. To obtain the questionnaire version used for MFC-mixed, for each of the seven socially undesirable and neutral triplets, one item was replaced by a socially desirable item from the larger item pool that was not part of the original MFC-matched version. Thus, the two versions differed only in these seven items.

Data Analysis

First, the Thurstonian IRT model was fit to the data under the instructions to respond honestly, separately for MFC-mixed and MFC-matched, using Mplus. In contrast to Brown and Maydeu-Olivares (2011), the models were fit with a restriction on the intercepts to make the response probabilities on the block level add up to one. Second, the response probabilities were calculated with Equation (6). Third, the Faking Mixture model was fit to the faking data from MFC-mixed and MFC-matched. I will use the Faking Mixture model that was fit to the data from MFC-matched to illustrate how the model parameters can be interpreted.

To test whether fakability was higher for blocks with mixed desirability, I fit the Faking Mixture model to data from both MFC-matched and MFC-mixed. In this model, the rank-order parameters β_{kr} were set equal for blocks that contained the same items in the two versions (Blocks 8-20), and they were estimated separately for blocks that differed between the two versions (Blocks 1-7). The rank-order parameters for blocks containing different items should not be set equal, but they can still lead to approximately equal block fakability parameters α_k. To test for differences in the block fakability parameters α_k (for Blocks 1-7), these differences were included in the model so that 95%-posterior intervals could be obtained for them. This is preferable to testing for differences between parameter estimates outside the model. Next, I examined whether rank orders favoring the desirable item in MFC-mixed had higher rank-order probabilities than in MFC-matched.

Results

Convergence

The Thurstonian IRT models for MFC-matched and MFC-mixed showed excellent fits according to the RMSEA, with RMSEA values of .036 and .038, respectively, and acceptable fits according to the SRMR, with SRMR values of .089 and .093. Therefore, response probabilities when responding honestly could be calculated with these models. Due to esti-

mation errors in the Mplus parameters, about half of all response probabilities did not add up to one in the seventh decimal place. Therefore, all response probabilities were rescaled by dividing them by the sum of the probabilities across the rank orders.

For the Faking Mixture models, to achieve convergence in both MFC-matched and MFC-mixed, I fixed the faking trait variance $Var(\boldsymbol{\theta})$ to 0.25, which is a typical variance for a trait capturing response biases (Billiet & McClendon, 2000; Plieninger & Heck, 2018), and the variance of the rank-order parameters $Var(\boldsymbol{\beta})$ to 4, which allows both low and high fakability parameters for individual blocks. Thus, I used a hyperprior only for the faking trait mean $M(\boldsymbol{\theta})$. Six chains with 5,000 iterations were run, out of which the first 2,500 were discarded. I checked convergence via convergence criteria obtained from rstan, namely, $\hat{R} < 1.01$ and the effective sample size divided by the true sample size larger than .001. I also visually inspected plots of posterior densities, of the autocorrelation and of the running mean across iterations. If not otherwise stated, I report posterior medians and 95% posterior intervals.

Descriptive Results

The faking trait $\boldsymbol{\theta}$ was distributed with a mean of 2.97 $[2.82; 3.14]$ and a variance of 0.29 $[0.26; 0.33]$. To better grasp the extent of faking, one can calculate the percentage of faking probabilities $\Phi(\theta_j + \alpha_k)$ that were $> .5$ for each block. This shows how many participants are predicted to fake. The high faking trait mean along with the percentages of participants predicted to fake (Table 4) show that the faking instructions were effective in motivating participants to fake. Due to the low variance in the faking trait, estimates of θ_j are not very reliable (see also the simulation study). Therefore, they should not be used to investigate the validity of the faking trait via correlations with other traits or criteria.

The block fakabilities α_k show how well matching was achieved (i.e., whether it was clear which rank order to prefer when faking). As Table 4 shows, fakability was generally high, with 99 to 100% of participants predicted to fake for each block, but fakability still differed between the blocks. The rank-order probabilities provide additional information. Some exemplary rank-order probabilities are shown in Figure 2. The other rank-order probabilities are shown in Figures S17-S19. For all blocks, some rank orders were more desirable than others. Most blocks with intermediate fakability showed a pattern such as Block 3. Here, ranking the item "I am often sad" first was undesirable, whereas the remaining rank orders in which this item was ranked second or lowest were still desirable (Figure 2). Thus, for such blocks, the number of possible rank orders when faking was limited to four instead of six. Among the highly fakable blocks, there were a few for which one rank order was more desirable than the other five, such as Block 7. Here, the order "I tend to be very particular about things," followed by "I have a vivid imagination," and finally "I stay in the background" was most desirable. For most of the highly fakable

Table 4: Block fakabilities α_k and percentage of participants predicted to fake in MFC-matched

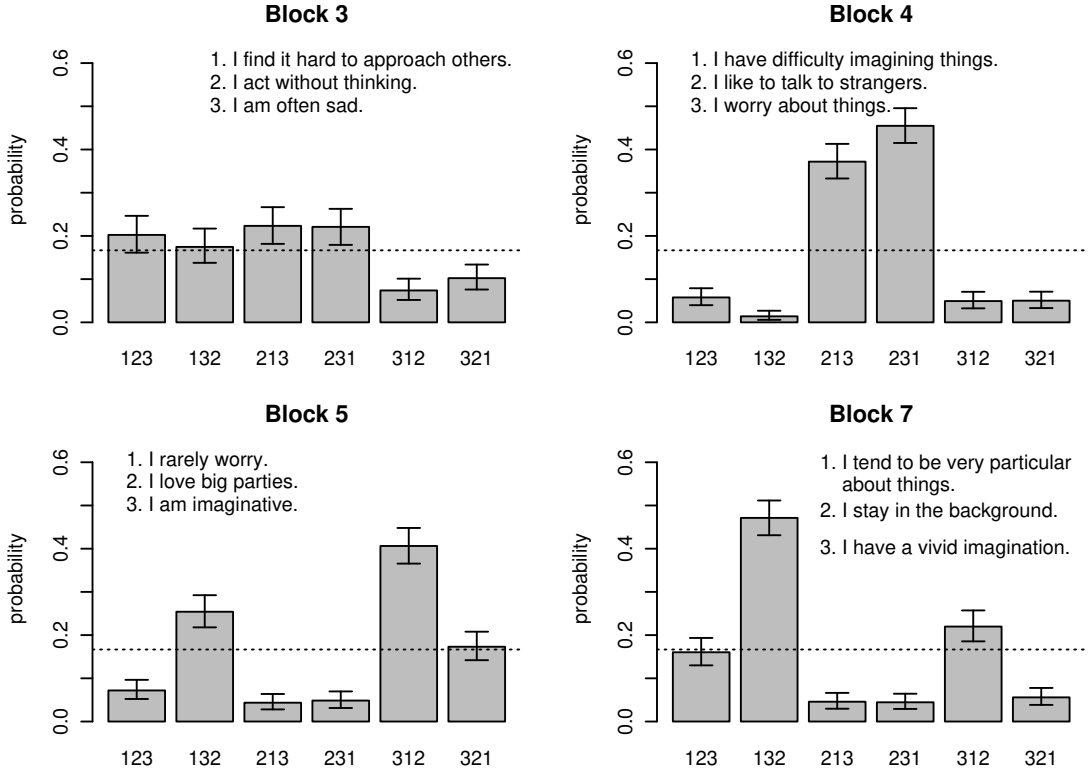
Block	Median	2.5%	97.5%	Predicted
3	-2.01	-2.19	-1.86	99
12	-1.95	-2.11	-1.79	99
1	-1.87	-2.02	-1.74	99
17	-1.86	-2.01	-1.72	99
10	-1.80	-1.95	-1.66	99
16	-1.77	-1.95	-1.62	100
9	-1.70	-1.83	-1.58	100
8	-1.68	-1.81	-1.56	100
11	-1.67	-1.80	-1.54	100
20	-1.61	-1.74	-1.50	100
13	-1.50	-1.65	-1.37	100
15	-1.41	-1.55	-1.29	100
5	-1.26	-1.38	-1.14	100
2	-1.25	-1.38	-1.14	100
14	-1.25	-1.37	-1.13	100
6	-1.17	-1.28	-1.06	100
18	-1.11	-1.23	-1.00	100
7	-1.09	-1.21	-0.97	100
19	-0.98	-1.08	-0.88	100
4	-0.88	-0.97	-0.80	100

blocks, one item (out of three) was clearly preferred or not preferred, limiting the number of rank orders when faking to two. For example, in Block 4, it was desirable to rank the item "I like to talk to strangers" first, but it was not clear which of the other two items "I have difficulty imagining things" and "I worry about things" should be preferred. The opposite tendency appeared, for example, in Block 5. Here, it was desirable to rank the item "I love big parties" last. Thus, for some blocks, participants agreed on which item should be ranked first, whereas for other blocks, they agreed on which item should be ranked last.

Fakability of Mixed versus Matched Blocks

The block fakability parameters α_k were higher in MFC-mixed than in MFC-matched for all seven mixed blocks (Figure 3). For Block 4, the block fakability parameters α_k were very high in both versions. The difference in block fakability parameters was descriptively higher for two out of the three socially undesirable Blocks 1 to 3, than for the neutral Blocks 4 to 7. Again, the rank-order probabilities derived from the rank-order parameters β_{kr} provide additional information. For all blocks, including a highly socially desirable item resulted in some rank-order probabilities being close to zero. Figure 4 shows three

Figure 2: Probabilities for rank orders when faking in MFC-matched



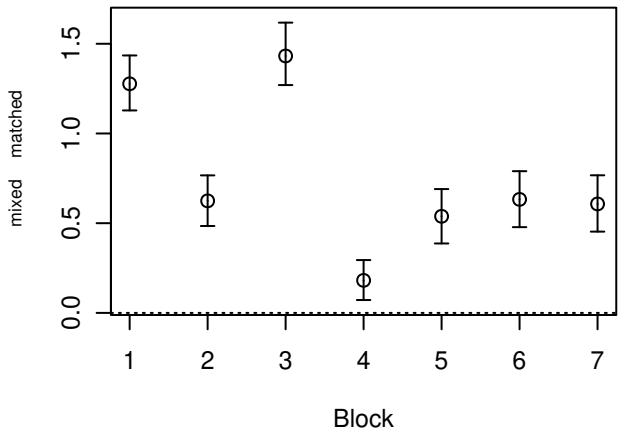
Note. The dotted line indicates where all rank orders are equally probable.

exemplary patterns (for all blocks, see Figure S20). For some blocks, both rank orders favoring the highly desirable item were more likely in MFC-mixed than in MFC-matched. However, the pattern differed only slightly, when the highly desirable item in MFC-mixed replaced an item that was already desirable in MFC-matched, such as for Block 4 (Figure 4). The pattern differed more markedly, when the highly desirable item in MFC-mixed replaced an undesirable item in MFC-matched, such as for Block 1. Here, in MFC-matched, the second item was most desirable and including a highly desirable third item in MFC-mixed led to overall high probabilities for the rank orders 3-1-2 and 3-2-1 (Figure 4). For other blocks, only one of the rank orders favoring the highly desirable item was more likely in MFC-mixed. For example, for Block 6, the rank order 1-3-2 was more likely, but not the rank order 1-2-3.

Discussion of Empirical Results

The sample had a strong tendency to fake as evidenced by a high faking trait mean. Such a sample is useful for evaluating maximum fakability during test construction. In real applicants or in clinical samples and in real instead of instructed faking situations, the

Figure 3: Differences in block fakability parameters α_k between MFC-mixed and MFC-matched



Note. The dotted line indicates no difference in block fakability parameters.

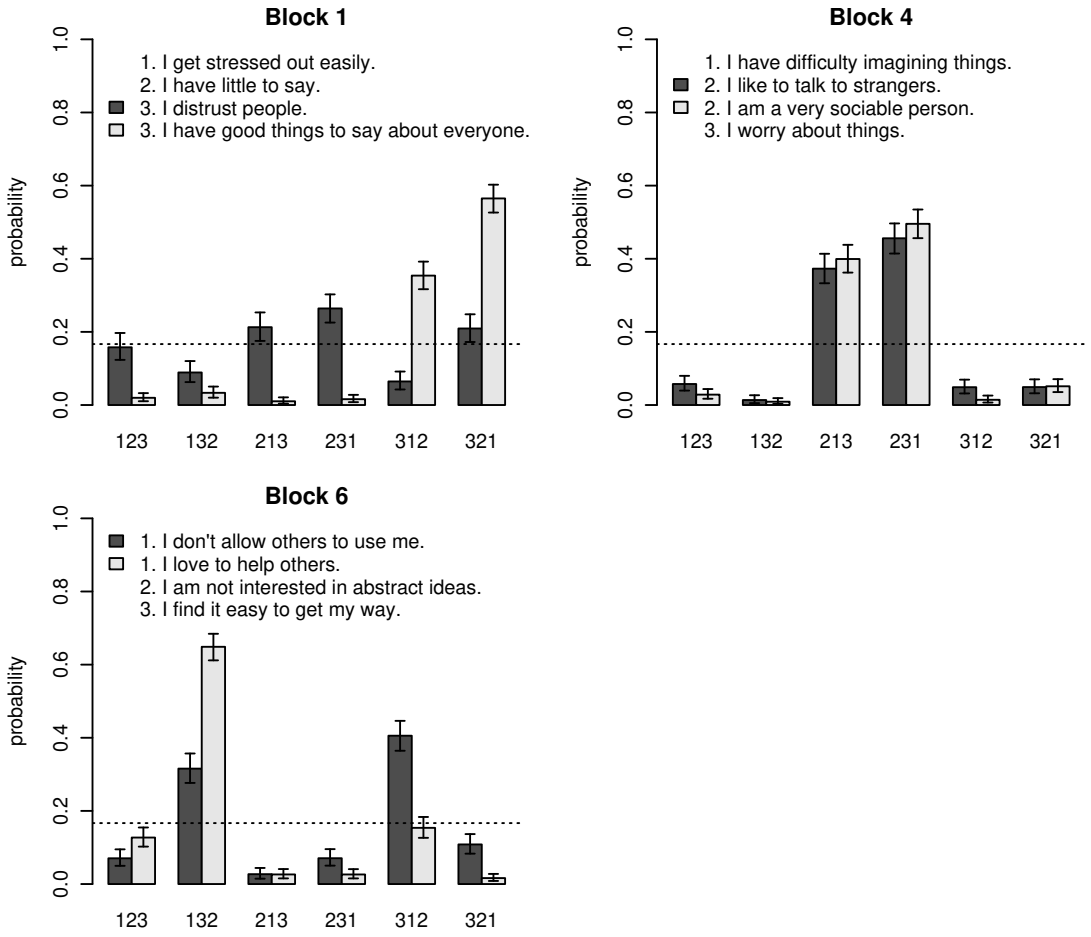
tendency to fake will most likely be smaller (MacCann et al., 2011), resulting in fewer faked responses than in the current study.

The descriptive analysis of the block fakability parameters showed that matching did not work out for all blocks. Though items within blocks were carefully matched for desirability, some blocks were highly fakable. This might indicate that participants evaluate item desirability in a more fine-grained manner when items are combined into blocks rather than presented individually (see also Feldman & Corah, 1960). More generally, this is additional evidence of item interactions in MFC questionnaires (Lin & Brown, 2017), which make modeling on the block-level necessary. A second reason for high fakability might be that the faking scenario (applying for a place in a psychology master program) differed from the one used to assess item desirability (general social desirability).

Some blocks had intermediate fakability where two rank orders were clearly undesirable. According to the Faking Mixture model, which of the remaining rank orders to select when faking is random. However, several empirical phenomena could underlie the randomness captured by the model: Participants might have ranked the items randomly, according to differences in perceived desirability that were unsystematic across individuals, or according to their levels of the content traits. In the last case, blocks with intermediate fakability are informative about some of the measured traits. This could be investigated via the construct and criterion validity of traits estimated from blocks with intermediate fakability in a high-stakes situation. In an exploratory analysis, Wetzel et al. (2021) found decreases in criterion validity in a simulated high-stakes situation.

The current analysis showed that beyond the overall block fakability α_k, the rank-order

Figure 4: Probabilities for rank orders when faking in MFC-matched versus MFC-mixed



Note. The dotted line indicates where all rank orders are equally probable. Results for MFC-matched are depicted in dark-grey, for MFC-mixed in light-grey.

probabilities provide additional information. They show which items are more desirable and which comparisons might still remain informative about the content trait. This information could be used during questionnaire development to modify blocks by removing or replacing items or to discard whole blocks. For example, when the rank-order probabilities show that one item is clearly preferred over the others, this item could be removed from the block.

The comparison of MFC-mixed and MFC-matched showed two things: First, matching was worth the effort, because matched blocks were less fakable than mixed blocks in all seven cases. Second, matching based on the desirability of the individual items was not sufficient, because item interactions were observed on the block-level. It is therefore recommended to first match items for desirability and then to examine the fakability of the resulting MFC blocks.

In line with the trait shift approach to estimating the fakability of MFC questionnaires (Wetzel et al., 2021), the analysis with the Faking Mixture model showed that the MFC-matched version of the questionnaire was fakable to some extent and that the MFC-mixed version was more fakable. Applying the Faking Mixture model additionally showed which particular blocks were fakable and that it was truly the unmatched items that increased fakability in MFC-mixed.

General Discussion

In this paper, I developed the Faking Mixture model, which is, to my knowledge, the first approach to modeling faking in the MFC format that allows the fakability of individual blocks to vary. A simulation study showed that the parameters of the Faking Mixture model could be recovered well under relevant conditions. Applying the Faking Mixture model to empirical data showed that matching based on the desirability ratings of the individual items was necessary but not sufficient to create an MFC questionnaire that can optimally reduce faking. Modeling fakability on the block-level allowed item context effects to be discovered within blocks. The Faking Mixture model can be used to reduce fakability during MFC test construction.

Faking on the Block-Level

For practical applications, the trait shift approach and the Faking Mixture model complement each other, because the former focuses on the person level and the latter on the item level: Whereas the trait shift approach can show effects of faking on trait estimates, the Faking Mixture model can show which blocks are more or less fakable. Moreover, the Faking Mixture model complements methods of matching and assessing item desirability because it allows fakability to be estimated on the basis of responses to MFC blocks. The

empirical validation showed that this approach is necessary because there were desirability differences that would not have been expected on the basis of the desirability ratings of the individual items. Using the results of the Faking Mixture model, test constructors can, for example, decide to keep only blocks with low fakability, or they can modify blocks by removing or replacing items that differ largely in their desirability from the others. Whereas the same can be achieved by tabulating frequencies of rank orders for honest versus faking conditions, the Faking Mixture model accounts for individual differences in the tendency to fake, which characterize most real assessment situations (Kuncel et al., 2012).

There are at least two differences between the response process for the rating scale versus the MFC format that make a mixture model on the block level parsimonious and necessary: First, because of the complexity of trait estimation, a shift model on the item level would be difficult if not impossible to identify. Second, a desirable rank-order cannot reasonably be determined without empirical data because item properties can change when items are combined into blocks (Lin & Brown, 2017).

Limitations

One limitation of the Faking Mixture model is that the tendency to fake does not differ across content domains, as there is only one faking trait that is uncorrelated with all other traits. However, this limitation is probably not very critical because several authors have argued that when socially desirable responding or faking is present, scales that are otherwise multidimensional end up showing a one-factor structure (e.g., Guenole et al., 2018; MacCann et al., 2011; van der Linden et al., 2010). Moreover, differential desirability of the content traits can still be captured by the rank-order probabilities.

In the Faking Mixture model, as in the current models for the faking of rating scales, perceived item desirability is fixed across individuals. A model that allows perceived item desirability to vary across individuals would be difficult to identify, even if the model is theoretically plausible.

Usually the fakability of a questionnaire is investigated either by using instructed (induced) faking or by comparing two samples (e.g., applicants and incumbents). Instructed faking allows researchers to estimate maximum fakability, but it is not fully representative of real applicant situations, for example, because real applicants might have goals other than faking (Kuncel et al., 2012). Comparing applicants and incumbents allows researchers to estimate how much faking occurs in real settings, but there is no guarantee that the differences are due only to faking. To apply the Faking Mixture model, given the current hardware and software resources, response probabilities under a non-faking situation are needed. These are most easily obtained with instructed faking. However, there are a few studies that have used within-subjects data from naturally occurring contexts. Gordon

and Stapleton (1956) compared the responses of high-school students who first filled out a questionnaire for guidance on a job search and some months later when they were actually seeking employment. Trent et al. (2020) compared the responses of army applicants with their later responses when they had been accepted into the army. Such a design would be optimal for the Faking Mixture model. Alternatively, a two-step procedure would be possible: First, instructed faking could be used to obtain estimates of item parameters and block fakabilities. Second, the faking trait and the content traits of another sample could be estimated with block and item parameters fixed to the estimates obtained in the first step. This is possible because, for the second step, only person parameters have to be estimated, which can be implemented in R. In this way, the Faking Mixture model could be used to obtain both the faking and the content trait estimates of real applicants.

The faking trait is incorporated in the Faking Mixture model to capture variance between individuals, but this variance is probably small in most cases, similar to response style traits (Böckenholt & Meiser, 2017). To assess the validity of the faking trait, a context in which it shows higher variance and therefore higher reliability is needed.

Future Research Directions

There are several ways to assess item desirability in the literature: Some researchers use item intercepts, estimated under linear factor models or ideal-point models, or raw item means, from a previous administration of the item set as indicators of item desirability (e.g., Guenole et al., 2018). Alternatively, they compute the difference between item intercepts obtained under instructions to respond honestly versus to fake (e.g., Lee et al., 2018; Ng et al., 2020). Others take a more explicit approach and have an external group rate the desirability of each item, either in general (e.g., Wetzel & Frick, 2020) or for a specific scenario, or they combine the two approaches (e.g., Heggestad et al., 2006; Jackson et al., 2000). The Faking Mixture model could be used to investigate which type of desirability estimate and matching provide the smallest fakability. Further, it could be used to investigate the effect of item keying on fakability because this issue has been raised by several authors (Bürkner et al., 2019; Morillo et al., 2016).

Due to its implementation in a Bayesian framework, the Faking Mixture model is flexible for incorporating and testing additional assumptions. For example, in the empirical validation, differences in the fakability of matched versus unmatched blocks were quantified and tested. Future research could continue this avenue, for example, by testing whether fakability and rank-order probabilities are similar across different faking contexts.

The Faking Mixture model incorporates the assumption that respondents are more likely to respond honestly when the block is less fakable. However, it is an empirical question whether they actually respond honestly, or whether they retrieve further information about desirability.

In an exploratory analysis, Wetzel et al. (2021) compared the criterion validity of responses on rating scales and MFC blocks of matched and mixed desirability under instructed faking. To further validate the Faking Mixture model, future research could investigate whether less fakable blocks show higher criterion or construct validities.

To my knowledge, this is the first mixture model for MFC data with a mixture on the block level. For rating scales, mixture models were proposed not only for faking and social desirability (Böckenholt, 2014; Leng et al., 2019) but also for other response biases, such as acquiescence (Plieninger & Heck, 2018). Models with a discrete mixture (constant across items) have been used for response styles in general (for an overview, see Henninger & Meiser, 2020). Future research could develop models for other response biases, such as careless responding in MFC data with a mixture on the block-level.

Further, the Faking Mixture model could be populated with other IRT models such as the generalized graded unfolding model for rank data (Hontangas et al., 2015; Lee et al., 2019), or models for rating scale data. For rating scale data, because the response probabilities can be estimated in most common software programs, the model could even be applied to data from a high-stakes context only. However, in this design, the model might capture not only faking but also other preferences for categories that are also constant across the sample.

In this paper, I proposed the Faking Mixture model, an IRT model for faking on MFC questionnaires. The empirical model validation showed that to construct an MFC questionnaire that can optimally reduce faking, we need to both match items on desirability and examine the fakability of the resulting MFC blocks. The Faking Mixture model can be used for the latter and may thus become a valuable tool for MFC test construction. I hope that the Faking Mixture model can provide avenues for future discussion and research on item desirability, faking, and the MFC format.

References

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

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

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

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

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., & 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. (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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). *Stan* : A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance*, *18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

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

Fox, J.-P. (2010). *Bayesian Item Response Modeling*. Springer New York. https://doi.org/10.1007/978-1-4419-0742-4

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., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities* (Vol. 195). Springer Science & Business Media.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). *Mvtnorm: Multivariate Normal and t Distributions (1.1-1)* (Computer Software; Version 1.1-1). http://CRAN.R-project.org/package=mvtnorm

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

Gordon, L. V., & Stapleton, E. S. (1956). Fakability of a forced-choice personality test under realistic high school employment conditions. *Journal of Applied Psychology*, *40*(4), 258–262. https://doi.org/10.1037/h0043595

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment*, *25*(4), 513–526. https://doi.org/10.1177/1073191116641181

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment

and faking resistance. *Journal of Applied Psychology*, *91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, *25*(5), 560–576. https://doi.org/10.1037/met0000249

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

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

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. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model: GGUM-RANK item and test information functions. *Journal of Educational Measurement*, *55*(3), 357–372. https://doi.org/10.1111/jedm.12183

Kuncel, N. R., & Borneman, M. J. (2007). Toward a New Method of Detecting Deliberately Faked Personality Tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*(2), 220–231. https://doi.org/10.1111/j.1468-2389.2007.00383.x

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2012). A Plea for Process in Personality Prevarication. *Human Performance*, *24*(4), 373–378. https://doi.org/10.1080/08959285.2011.597476

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

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., & 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

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

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

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

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

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–68). Erlbaum.

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., & 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

Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf

R Core Team. (2020). *R: A language and environment for statistical computing (3.6.3)* (Computer Program and Language; Version 3.6.3). Vienna, Austria. https://www.R-project.org/

Revelle, W. (2019). *Psych: Procedures for Personality and Psychological Research (1.8.12)* (Computer Software; Version 1.8.12). Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

Stan Development Team. (2020a). *RStan: The R interface to Stan (2.19.3)* (Computer Software; Version 2.19.3). http://mc-stan.org

Stan Development Team. (2020b). *Stan Reference Manual (2.22)* (Computer Program and Manual; Version 2.22). http://mc-stan.org

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

Trent, J. D., Barron, L. G., Rose, M. R., & Carretta, T. R. (2020). Tailored Adaptive Personality Assessment System (TAPAS) as an indicator for counterproductive work behavior: Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology*, *32*(1), 51–59. https://doi.org/10.1080/08995605.2019.1652481

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S (4)* (Computer Software; Version 4). New York.

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

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

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

Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial-Organizational Psychologist*, *49*(1), 29–36.

Block Information in the Thurstonian Item Response Model

Susanne Frick[1] *

[1]Department of Psychology, School of Social Sciences, University of Mannheim

The manuscript was submitted to Psychometrika.

*This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group „Statistical Modeling in Psychology“ (SMiP). The author acknowledges support by the state of Baden-Württemberg through bwHPC. The author would like to thank Eunike Wetzel and Thorsten Meiser for their helpful comments and discussions on previous versions of this manuscript, Safir Yousfi for sharing his R-code and Christoph Klauer for helpful discussions and the idea of using an R^2 metric.

Abstract

Multidimensional forced-choice (MFC) tests become increasingly popular but their construction is complex. The Thurstonian IRT model is most often used to score MFC tests with dominance items. Currently, information in the Thurstonian IRT model is computed for binary outcomes of pairwise comparisons, neglecting stochastic dependencies and item interactions. In this paper, it is shown how Fisher information on the block-level can be estimated. Several indices are proposed to summarize block information, which is a non-invertible matrix without a closed-form expression. A simulation study showed that standard errors based on block information are unbiased. Two other simulation studies found comparable performance of several block information summaries in automated test assembly. Thus, block information can aid the construction of reliable MFC tests.

Keywords: multidimensional forced-choice, Thurstonian item response model, information, standard errors, automated test assembly

Block Information in the Thurstonian Item Response Model

Test constructors aim to measure constructs reliably and valid. Most personality tests employ rating scales (e.g., strongly disagree, disagree...) for this purpose, but responses to rating scales are potentially biased, for example, by response styles (Henninger & Meiser, 2020; Krosnick, 1999; Wetzel et al., 2016). As an alternative, the multidimensional forced-choice (MFC) format is becoming increasingly popular. In the MFC format, several items measuring different attributes are presented simultaneously in blocks. The respondent's task is then to rank the items (for an example see Figure 1) or select those that he or she prefers most and/or least. This research is concerned with the former, termed full ranking.

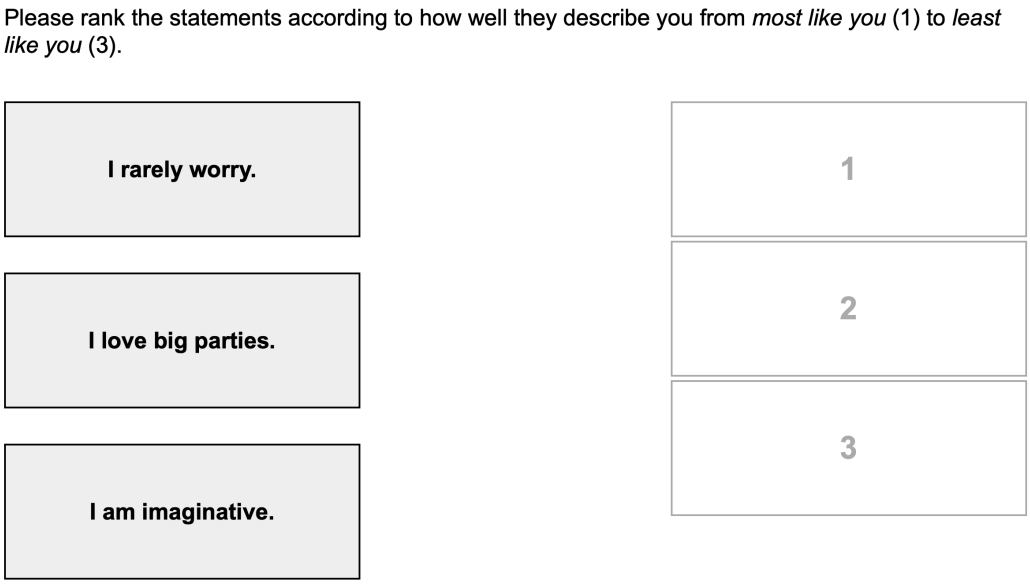


Figure 1: Example for the multidimensional forced choice format from the Big Five Triplet (Wetzel & Frick, 2020). The first item assesses neuroticism (reverse-coded), the second extraversion and the third openness.

In comparison to rating scales, the MFC format has the advantage to avoid or reduce several response biases. For example, overall faking is reduced (Cao & Drasgow, 2019; Wetzel et al., 2021) and uniform response biases such as halo effects are avoided (Brown

et al., 2017), for an overview see Brown and Maydeu-Olivares (2018a).

As interest in the MFC format increases, it becomes important how to construct such tests. Besides combinatorics, item interactions make the construction of MFC tests more complex than that of tests with a single-stimulus format such as the rating scale format. Item interactions occur because the items in a block are not evaluated independently, but are weighed against each other in giving a response. Research has found that the measured constructs slightly change when the same items are presented in an MFC versus a rating scale format (Guenole et al., 2018; Wetzel & Frick, 2020). Further, item desirability is evaluated differently in the context of MFC blocks than for single-stimulus items (Feldman & Corah, 1960; Hofstee, 1970). Most importantly, item parameters from item response theory (IRT) models differed depending on which items were combined to blocks (Lin & Brown, 2017).

To appropriately account for item interactions, item information must be calculated on the block level, henceforth termed *block information*. For MFC tests with ideal-point items, that is, where the preference for an item is highest at a certain trait level and decreases with increasing distance from it, block information can be calculated based on the generalized graded unfolding model for rank responses (Joo et al., 2018). It has been shown that this can be used to construct computerized adaptive tests (Joo et al., 2020). However, most tests employ dominance items, where the preference for an item increases or decreases monotonically with increasing trait levels. For MFC tests with dominance items, the Thurstonian item response model (Brown & Maydeu-Olivares, 2011) has become the most popular and widely applicable IRT model. The Thurstonian IRT model can incorporate different block sizes and different response instructions, such as ranking all items in a block or picking one of them.

However, in the Thurstonian IRT model, information is currently calculated for binary outcomes of pairwise comparisons of items within blocks (Brown, 2016; Brown & Maydeu-Olivares, 2011). For example, in a block of three items, there are three pairwise comparisons, namely between items 1 and 2, items 1 and 3, and items 2 and 3. Information for binary outcomes of pairwise comparisons has several disadvantages: First, from an empirical perspective, it does not fully reflect all item interactions within a block. Second, from a test constructor's perspective, it is unclear how to use the information for binary outcomes of pairwise comparisons to select items or blocks. Third, from a statistical perspective, information for binary outcomes of pairwise comparisons is problematic because it is not independent for block sizes larger than two (Brown & Maydeu-Olivares, 2011, 2018b). When test information is calculated based on binary outcomes of pairwise comparisons, it is slightly overestimated and accordingly standard errors for person parameter estimates are slightly underestimated. This effect increases with block size (Yousfi, 2018).

There is a formulation of the Thurstonian IRT model by Yousfi (2018), which allows

to calculate block information, but so far, it was not used for this purpose. Instead, the formulation was only used for trait estimation, that is, to calculate the likelihood of responses across blocks and to compute standard errors for trait estimates (Yousfi, 2018, 2020). Therefore, the aim of this paper is to show how to calculate block information for ranking responses in the Thurstonian IRT model and how it can be used for test construction. To accomplish this, in the first part, the formulation of the Thurstonian IRT model based on Yousfi (2018) is presented and formulas for block information are given. As MFC questionnaires measure several traits per block, block information is a matrix, which is difficult to interpret and to use for test construction. Therefore, in the second part, several information summaries are proposed, drawing from multidimensional computerized adaptive testing (CAT), that integrate block information across traits and it is shown how to plot the information summaries. There is no analytical formula for block information, instead, two steps of numerical approximation are involved. Therefore, it is essential to examine the accuracy of the estimation procedure. This is accomplished most easily by comparing standard errors based on expected and observed information to empirical standard errors. Therefore, the third part is a simulation study on standard error accuracy. However, the accuracy of standard errors on the questionnaire level is only partially informative about whether block information can be used for test construction. To approximate the test construction process more closely, automated test assembly (ATA) is simulated. In ATA, blocks can be selected so that test information is maximized. In addition, this allows to illustrate which ATA algorithms can be used to assemble MFC test based on block information and to compare the performance of algorithms and block information summaries. Hence, the fourth part consists of two simulation studies on ATA with MFC block information for two different scenarios.

Thurstonian IRT model

In the Thurstonian IRT model, for each item, there is a latent response tendency called utility. The utility t for person j on item i is a linear function of a latent trait θ_j:

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

where μ_i denotes the item intercept, λ_i the item loading and ε_{ij} the error term. The latent traits are assumed to be multivariate normally distributed: $\boldsymbol{\Theta} \sim N(\mathbf{M}_\theta, \boldsymbol{\Sigma})$ and the errors are independently normally distributed: $\boldsymbol{\varepsilon}_i \sim N(0, \psi_i)$.

Following Thurstone's Law of Comparative Judgment (Thurstone, 1927, 1931), participants order the items within each block according to the magnitude of their utilities. To express this mathematically, first, within each block k vectors of utilities $\boldsymbol{t}_k$ and of error variances $\boldsymbol{\psi}_k^2$ are sorted in descending order, according to the selected rank order r. Sec-

ond, differences between consecutive utilities are obtained by a comparison matrix $\mathbf{A}$. For example, if block size $B = 3$:

$$\boldsymbol{A}_{B=3} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{2}$$

With utilities sorted in descending order, each difference between two consecutive utilities is positive. Therefore, the probability to select rank order r is the area under the multivariate normal density of utilities where this applies (Yousfi, 2020):

$$P(X_{jk} = r) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty N\left(\boldsymbol{A}\boldsymbol{t}_{jk}(r), \boldsymbol{A}\boldsymbol{\psi}_k^2(r)\right) dt_1 dt_2 \dots dt_{B-1} \tag{3}$$

This multiple integral can be numerically approximated (Genz, 2004; Genz & Bretz, 2002). Note that, with estimated item parameters, Equation 3 is only true when the item intercepts μ_i were estimated instead of unrestricted thresholds for binary outcomes of pairwise comparisons which is the commonly used procedure (Brown & Maydeu-Olivares, 2011). As Equation 3 shows, the probability to select a certain rank order is dependent on all latent traits assessed in the block.

The Fisher information for a block and a single rank order r is the negative of the Hessian of the response probability for the latent traits, where $H(f)$ denotes the Hessian of function f.

$$\mathbf{I_{kr}} = -\mathbf{H}\left(P(X_{jk} = r)\right) \tag{4}$$

Obtaining the Hessian for a multidimensional response probability involves differentiating twice for each pair of traits in both orders. Hence, for F latent traits, $\mathbf{I}_{kr}$ is an $F \times F$ matrix. For example, to obtain the entry in the second row and first column, $\mathbf{I}_{kr}(1, 1)$, the response probability $P(X_{jk} = r)$ is first differentiated for Trait 1 and then for Trait 2. As for the response probability, there is no analytical solution for the Hessian, but numerical approximation is feasible. Expected block information $\mathbf{I}_k$ is calculated across all $R = B!$ possible rank orders:

$$\mathbf{I}_k = \sum_{r=1}^{R} \mathbf{I}_{kr} P(X_{jk} = r) \tag{5}$$

In the following, I refer to expected block information as block information if not explicitly indicated otherwise. When each item is only presented in one block, as is typically done in MFC tests, block information is additive across the test. Expected test information $\mathbf{I}_T^{exp}$ is obtained as the sum of expected block information across all $k = 1 \dots K$ blocks:

$$\mathbf{I}_T^{exp} = \sum_{k=1}^{K} \mathbf{I}_k \tag{6}$$

Observed test information $\mathbf{I}_T^{obs}$ is obtained as the sum of block information $\mathbf{I}_{kr}$ for the rank orders observed:

$$\mathbf{I}_T^{obs} = \sum_{k=1}^{K} \mathbf{I}_{kr} \tag{7}$$

Then, standard errors for a trait vector can be calculated by inverting expected or observed test information and taking the square root of its diagonal:

$$\sigma_T = \sqrt{\text{diag}\left(\mathbf{I}_T^{-1}\right)} \tag{8}$$

Standard errors based on the formulation in Equation 3 are unbiased (Yousfi, 2020) in contrast to the ones based on the typically used formulation with binary outcomes of pairwise comparisons.

Information Summaries

Dealing with MFC block information is challenging due to several properties of MFC tests: First, as outlined above, for F latent traits, the block information is an $F \times F$ matrix. Second, the information matrix for a single block is not invertible, because the latent trait space is only identified when there are several blocks and no linear dependencies between factor loadings λ, that is, when the matrix of factor loadings $\mathbf{\Lambda}$ has full rank (for details, see Brown & Maydeu-Olivares, 2018b). Third, the block information cannot be linearly approximated, because there is no closed-form expression for it. To cope with this, I propose several indices to summarize the information matrix for test construction. An R package implementing the estimation procedure, the information summaries and the ATA algorithms is available at GitHub: https://github.com/susanne-frick/MFCblockInfo. The R code for running and analyzing the simulations as well as the simulation results are available from the same GitHub repository.

Optimality Criteria

Optimality criteria originate from the optimal design literature and have been used in multidimensional CAT and sometimes in multidimensional ATA (Debeer et al., 2020). In MFC tests, usually all traits are of interest to the investigator. Therefore, I focus on those criteria that weigh all traits equally. Out of those, A-optimality and D-optimality performed best in an MFC CAT simulation (Lin, 2020). A-optimality is the sum of the sampling variances (Equation 8). D-optimality is the determinant of the information matrix (Equation 6). Hence, both A- and D-optimality depend on the information matrix being positive-definite. Therefore, with MFC tests, they can only be computed for several blocks at once (i.e., for test information). In contrast, T-optimality does not depend on

a positive-definite matrix. T-optimality is the sum of the diagonal entries of the information matrix (Equation 5). In such, it ignores the impact of trait correlations (Lin, 2020). T-optimality performed worst in an MFC CAT simulation (Lin, 2020). However, it has the advantages that it can be calculated for a single block and that it is additive across blocks.

Block R^2

To quantify how much a block contributes to the measurement of each trait, I propose a new information summary, which I will call block R^2. Block R^2 quantifies the proportional reduction in the sampling variances of the traits achieved by including this block. To compute block R^2, first, test information $\mathbf{I}_T$ is calculated for two sets of blocks: for a set T including the respective block and for a set $T \setminus k$ excluding it. Second, sampling variances are calculated for both sets following Equation 8. Third, block R^2 is obtained such that higher values indicate higher information:

$$\mathbf{R}_k^2 = 1 - \frac{\boldsymbol{\sigma}_T^2}{\boldsymbol{\sigma}_{T\setminus k}^2} \tag{9}$$

It follows from this procedure that block R^2 is relative to the set T of reference blocks. In practical applications, the set of reference blocks can be all blocks assessed or a subset of blocks forming a test that should be extended. As can be seen from Equation 9, block R^2 is related to A-optimality, but provides a trait-level index in a familiar metric.

Block Information Plots

To visualize block information, I propose 3D plots for block R^2 and expected *SE*s, and 2D plots for expected *SE*s, similar to Joo et al. (2018). For illustration, a test measuring five traits with 20 blocks of three items was simulated. The test design and distributions of the item parameters were identical to the simulation studies described later.

Figure 2 shows a 3D plot of block R^2, termed *block information plot*. The example shows block R^2 for Trait 1, for a block measuring Traits 1, 3, and 4. In each subplot, two traits measured by this block are varied continuously while the others take on fixed levels. For example, in the upper row of Figure 2, Traits 1 and 3 are varied on the horizontal axes. From left to right, Trait 4 is fixed at -1, 0 and 1 reflecting the mean $\pm 1SD$. The traits not measured by this block are fixed (at 0 in the example). In preliminary tests, this later level influenced the absolute size of block R^2 but not the differences between trait levels that are of interest. The block information plot can show how the size of block information depends on the traits in interaction. For example, in Figure 2, for low levels of Trait 4, information is highest for medium levels of Trait 1. However, for medium and high levels

of Trait 4, information is lowest for medium levels of Trait 1. Thus, this example illustrates that it can be illuminating to treat information in the MFC format as multidimensional on the block level instead of considering items in isolation.

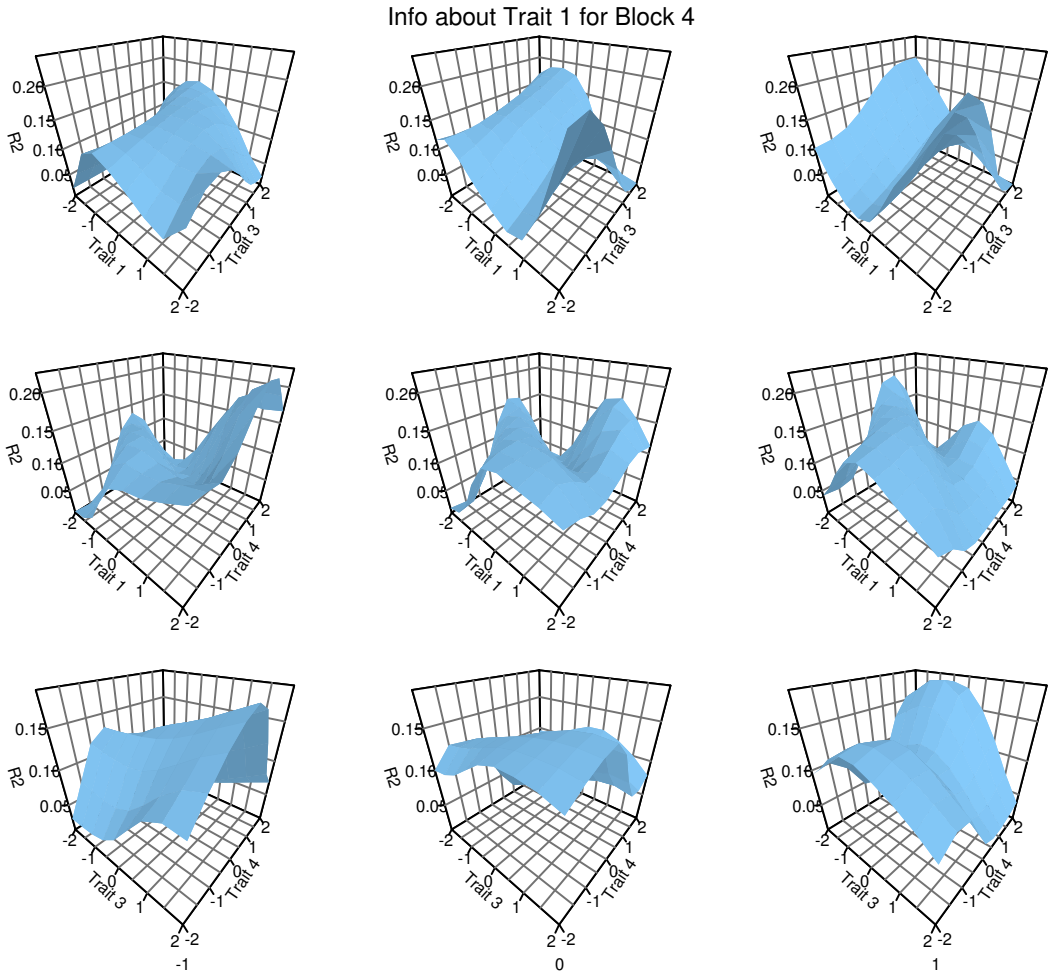


Figure 2: Block R^2 for Trait 1 from a simulated test block. Items 1-3 measured traits 1, 3, and 4, respectively. The simulated item parameters were: $\mu_1 \approx -0.22$, $\mu_2 \approx -0.96$, $\mu_3 \approx 0.82$, $\lambda_1 \approx -0.92$, $\lambda_2 \approx 0.77$, and $\lambda_3 \approx 0.89$. $\psi_i^2 = 1 - \lambda_i^2$.

Figure 3 shows a 3D plot of expected *SE*s, termed *test information plot*. Traits are varied pairwise while all other traits are fixed (at 0 in the example). The example in Figure 3 shows that *SE*s are lowest for medium trait levels, as to be expected. Dependencies on the other traits are small and mainly occur for extreme trait levels. For example, *SE*s for Trait 1 range from .34 to .41 for a Trait 1 level of 0, depending on the other traits, whereas they range from .35 to .55 for varying Trait 1 levels.

Figure 4 shows a 2D plot of expected *SE*s, termed *2D test information plot*. Here, only one trait varies systematically. Levels for the other traits are randomly drawn from a

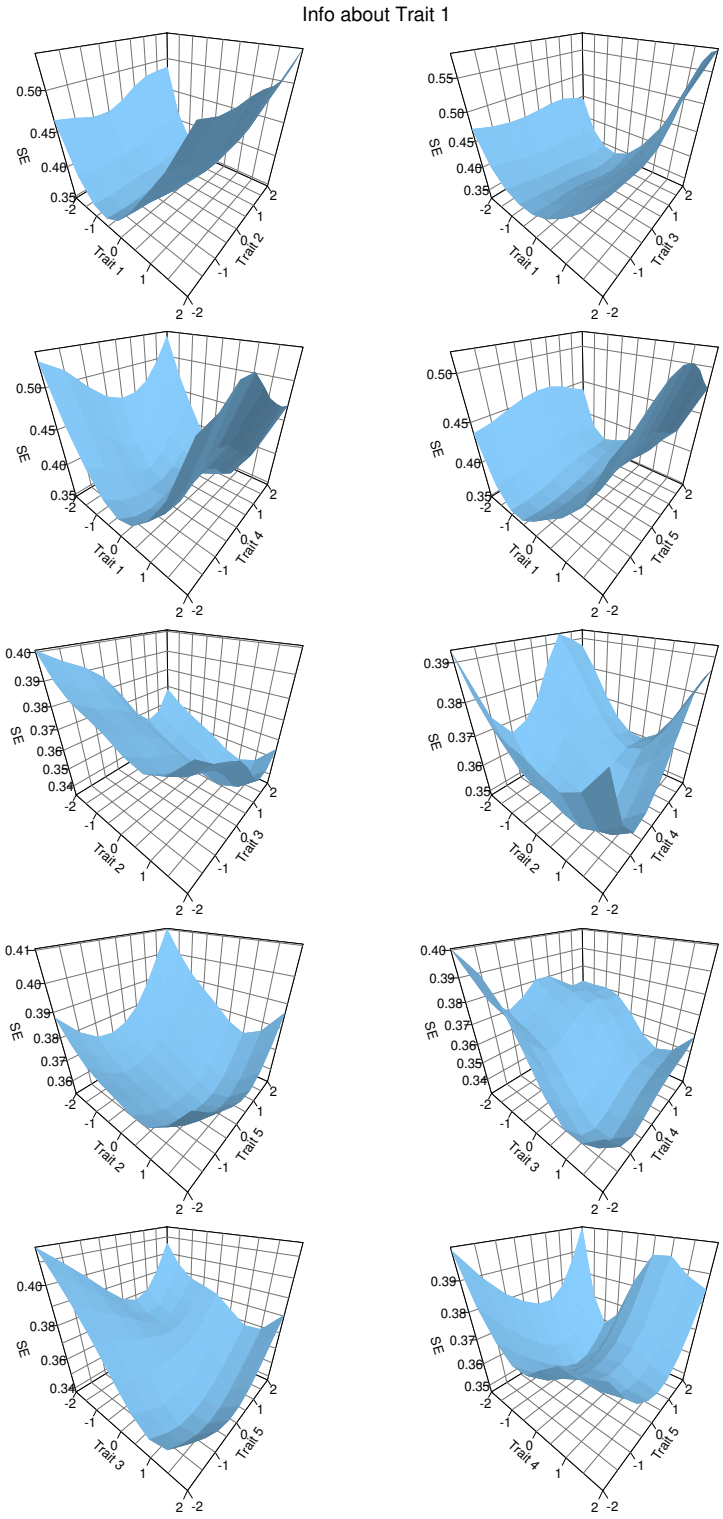


Figure 3: Expected *SE*s for Trait 1 from a simulated test measuring five traits with 20 blocks of block size three.

multivariate normal distribution for a specified number of respondents (100 in the example) so that the true trait correlations remain intact. The example in Figure 4 shows that *SE*s for Trait 1 are lowest for trait levels around the mean and 0.5 *SD* below it.

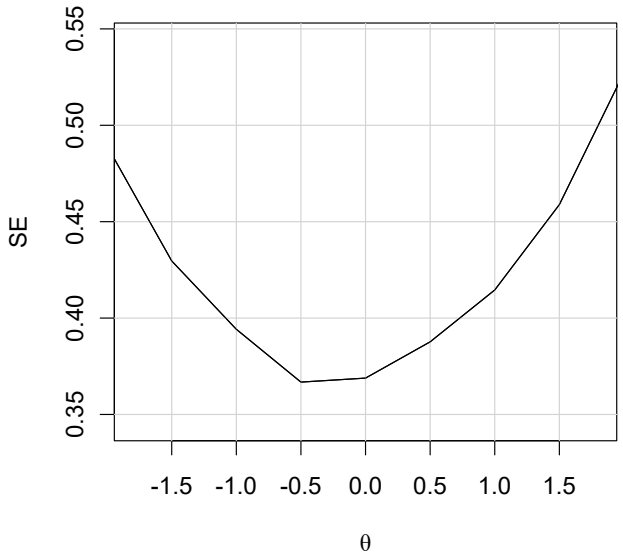


Figure 4: Expected *SE*s for Trait 1 from a simulated test measuring five traits with 20 blocks of block size three. For each level of Trait 1, *SE*s were averaged across 100 respondents with trait levels drawn from a multivariate normal distribution with a mean vector of zero and covariances given in Table 1.

Simulation Study 1: Simulation on Standard Error Accuracy

Both obtaining the response probability (Equation 3) and its Hessian (Equation 4) involve numerical approximation. Thus, block information is essentially an estimate. Therefore, in order to evaluate whether block information can be used for test construction, it is crucial to examine the accuracy of its estimation. The aim of this simulation study was to evaluate the accuracy of estimated block information. On the block level, there is no clear reference point for what constitutes accurate information. Instead, the accuracy of test information as the sum of block information (Equations 6 and 7), inverted to standard errors (Equation 8) was evaluated. Specifically, it was examined in how far empirical *SE*s and those based on observed and expected Fisher information correspond. Good correspondence provides evidence for the accuracy of estimated block information. The accuracy of *SE*s was examined under various conditions of test design influencing the

amount of information and for two types of estimators – maximum likelihood (ML) and maximum a posteriori (MAP). The MAP estimator is most often used for Thurstonian IRT models (e.g., Brown & Maydeu-Olivares, 2011; Wetzel & Frick, 2020).

Empirical *SE*s were defined as the standard deviation of trait estimates across R responses of the same person j (i.e., trait levels) to test q (cf., Ippel & Magis, 2020; Paek & Cai, 2014):

$$\mathrm{SE}_q\left(\theta_j\right) = \frac{\sum_{r=1}^{R}\left(\hat{\theta_{jr}} - \bar{\hat{\theta}}_j\right)^2}{R-1} \tag{10}$$

Both observed and expected *SE*s are based on test information at the trait estimate. For expected information, each possible rank order is weighted by its probability (Equation 5) and *SE*s obtained by setting $\mathbf{I}_T = \mathbf{I}_T^{exp}$ in Equation 8. Observed information is calculated only for the rank orders observed (Equation 7). Therefore, observed *SE*s are the diagonal of the inverse of the Hessian at the likelihood of the trait estimate. Equivalently, they can be calculated by setting $\mathbf{I}_T = \mathbf{I}_T^{obs}$ in Equation 8.

Methods

MFC responses were simulated for 5 traits, a test with block size three and 1/2 of pairwise item comparisons across the test involving items keyed in different directions (i.e., one positive, one negative factor loading). Item keying was chosen so that *SE* accuracy would not be confounded with ipsativity. Ipsativity with all positively keyed items was observed in simulations (e.g., Bürkner et al., 2019; Frick et al., 2021). Item intercepts μ_i were drawn from $U(-1, 1)$. Item uniquenesses ψ_i^2 were calculated as $1 - \lambda_i^2$. Errors were drawn from $N(0, \psi_i)$. For the second trait, trait levels varied from -2 to 2 in steps of .5. The other traits were fixed at 0. This yields 13 trait levels. Traits were estimated with box constraints to be within the range of $[-3, 3]$.

Three factors were varied and completely crossed: First, *size of factor loadings*: High factor loadings were drawn from $U(.65, .95)$, and low factor loadings were drawn from $U(.45, .75)$. Second, the *type of estimator* was either ML or MAP. For the MAP estimator, a multivariate normal prior with a mean vector of zero and correlations based on meta-analytic correlations between the Big Five (D. van der Linden et al., 2010) were used. The correlations are shown in Table 1. Third, *test length* was either short (20 blocks) or long (40 blocks).

Simulation Procedure

All data generation and analysis was carried out in R (R Core Team, 2020), involving the R packages doMPI (Weston, 2017), mvtnorm (Genz et al., 2020), numDeriv (Gilbert & Varadhan, 2019), psych (Revelle, 2019), gridExtra (Auguie, 2017), and ggplot2 (Wickham, 2016). For each combination of test design, estimator and trait level, 500 tests were simu-

Table 1: Correlations used in the simulation study

Trait	E	O	A	C
N	−.36	−.17	−.36	−.43
E		.43	.26	.29
O			.21	.20
A				.43

Note. N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness. These are meta-analytic correlations between the Big Five as reported by D. van der Linden et al., 2010.

lated, yielding a total of $2 \times 2 \times 13 \times 500 = 26000$ tests. Within each test, item parameters were drawn according to the test design and $R = 500$ response vectors were simulated. Traits were estimated for each response vector. Then, the three types of *SE*s (empirical, expected and observed) were computed at the trait estimate. Trait recovery and accuracy of *SE*s for the second trait were assessed by mean bias (MB) and root mean square error (RMSE) according to the following formulas, where ξ denotes the true parameter and $\hat{\xi}$ its estimate:

$$MB(\xi) = \frac{\sum_{r=1}^{R} \hat{\xi} - \xi}{R} \tag{11}$$

$$RMSE(\xi) = \sqrt{\frac{\sum_{r=1}^{R} \left(\hat{\xi} - \xi\right)^2}{R}} \tag{12}$$

MB and RMSE were computed for the latent traits θ and their expected and observed *SE*s. For the *SE*s, the empirical *SE* according to Equation 10 served as true parameter.

Results

Overall, bias was lower and *SE*s were smaller for medium trait levels and for the long test (Figure 5). For example, for the short test, the RMSE ranged from .24 to .99, for the long test it ranged from .18 to .65. The ML estimator showed a slight outward bias (e.g., mean MB($\theta = 2$) = .13), the MAP estimator showed a more pronounced inward bias (e.g., mean MB($\theta = 2$) = -.43, Figure 5). Both MB and RMSE were smaller for high loadings and for the MAP estimator. The RMSE was especially high for low loadings combined with the ML estimator.

For the ML estimator, empirical *SE*s were smaller for medium trait levels (Figure 5). For the MAP estimator, empirical *SE*s were approximately similar across trait levels. This might be attributed to the inward bias of the estimates. Empirical *SE*s were highest for the ML estimator combined with low loadings, similar to RMSE. For both estimators, empirical *SE*s were smaller for trait levels of ±2 than for ±1.5. This might be due to the

box constraints: For true trait levels of ±2, many estimates were ±3, resulting in smaller standard deviations of the estimates, that is, empirical *SE*s.

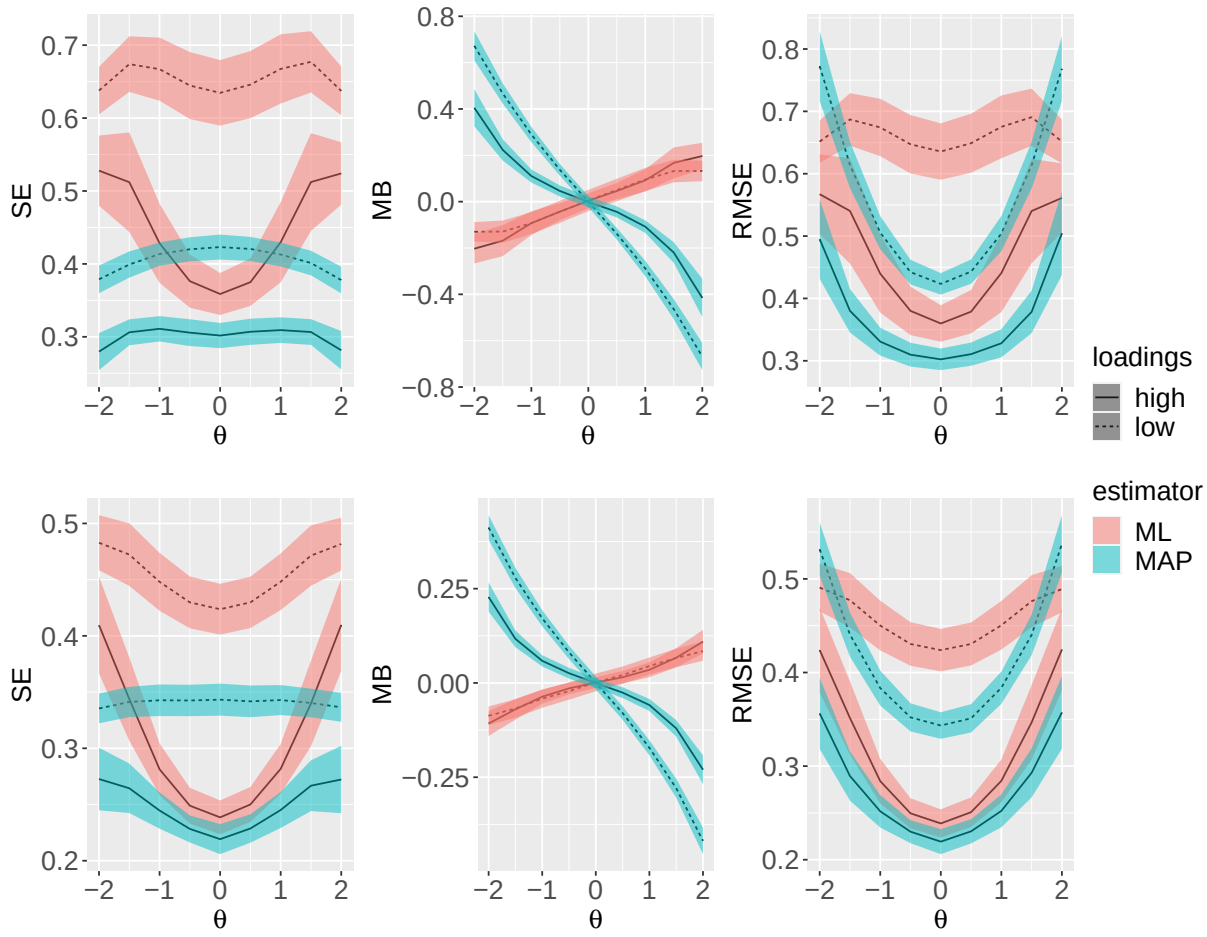


Figure 5: Trait recovery and empirical *SE*s in simulation study 1 on standard error accuracy. The top row shows results for the short test (20 blocks) and the bottom row shows results for the long test (40 blocks). Shaded areas show ±1SD around the mean (line). SE = empirical Standard Error, MB = Mean Bias, RMSE = Root Mean Square Error, ML = Maximum Likelihood, MAP = Maximum a Posteriori.

In 8% of the cases, the information-based *SE*s could not be estimated for up to 1114 (out of 2500) cases per condition and test (mean = 23, SD = 113). The MB of the information-based *SE*s was generally low (mean = .02, SD=.06, Table 2). The difference in the MB of observed and expected *SE*s was negligible, explaining 0% of variance across trait levels (Table 3, Figures 6 and 7). For the ML estimator, information-based *SE*s had a small negative (mean = -.01), for the MAP estimator a small positive MB (mean = .06, Table 2). The MB of information-based *SE*s was especially high for the MAP estimator combined with low loadings. The RMSE of information-based *SE*s was smaller for longer

tests, explaining 22% of variance across trait levels (Table 3). Differences related to the type of estimator and loadings were less pronounced than for MB.

Table 2: Means of bias for information-based standard errors by condition in simulation study 1 on standard error accuracy

Method	Length	Estimator	Loadings	MB		RMSE	
expected	short	ML	high	-0.01	(0.05)	0.09	(0.07)
			low	-0.01	(0.06)	0.07	(0.03)
		MAP	high	0.05	(0.05)	0.06	(0.05)
			low	0.12	(0.03)	0.12	(0.03)
	long	ML	high	-0.02	(0.02)	0.04	(0.03)
			low	-0.01	(0.02)	0.03	(0.01)
		MAP	high	0.01	(0.03)	0.03	(0.02)
			low	0.05	(0.02)	0.06	(0.02)
observed	short	ML	high	0.00	(0.06)	0.10	(0.10)
			low	0.00	(0.06)	0.08	(0.04)
		MAP	high	0.06	(0.05)	0.06	(0.05)
			low	0.10	(0.04)	0.10	(0.04)
	long	ML	high	-0.02	(0.02)	0.04	(0.04)
			low	-0.01	(0.02)	0.03	(0.01)
		MAP	high	0.01	(0.03)	0.03	(0.02)
			low	0.05	(0.02)	0.05	(0.02)

Note. MB = Mean Bias, RMSE = Root Mean Squared Error, ML = Maximum Likelihood, MAP = Maximum a posteriori. Standard deviations are given in parentheses.

Table 3: Variance in bias for information-based standard errors explained in % by the manipulated factors in simulation study 1 on standard error accuracy

Factor	MB	RMSE
loadings	5	1
estimator	34	0
length	6	22
loadings×estimator	3	7
estimator×length	3	0
loadings×estimator×length	0	1
Residuals	48	69

Note. MB = Mean Bias, RMSE = Root Mean Squared Error, ML = Maximum Likelihood, MAP = Maximum a posteriori. Expected vs. observed explained less than 1% of variance.

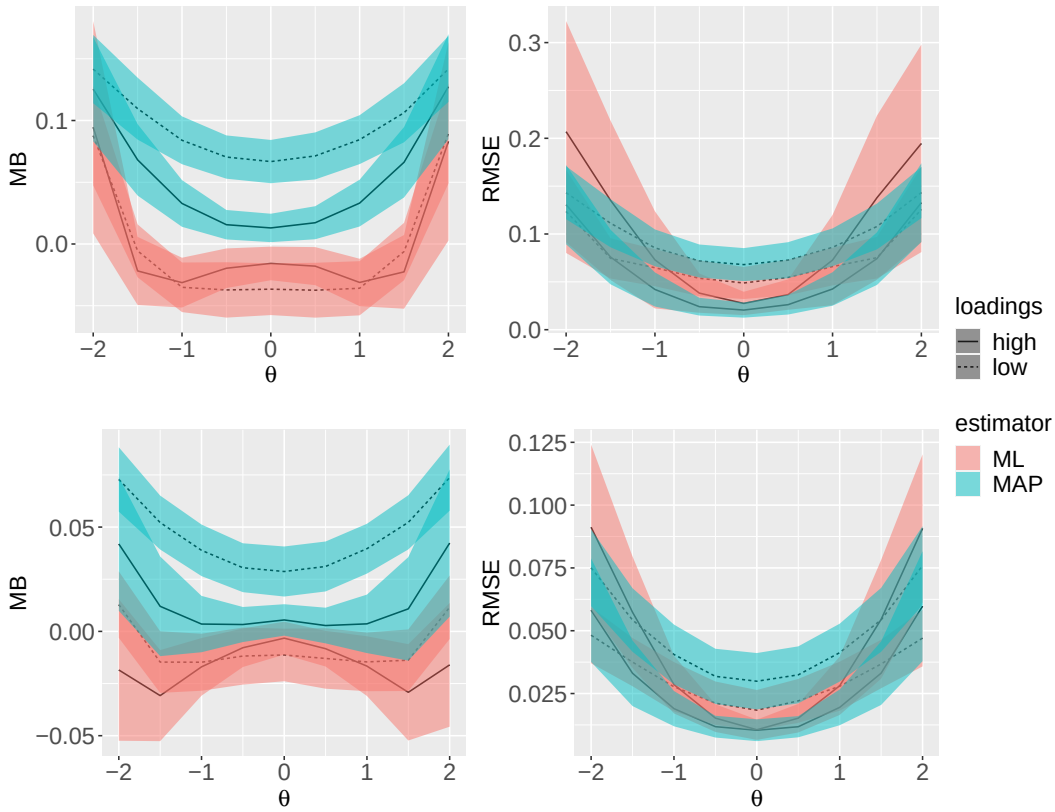


Figure 6: Bias of observed standard errors in simulation study 1 on standard error accuracy. The top row shows results for the short test (20 blocks) and the bottom row shows results for the long test (40 blocks). Shaded areas show $\pm 1SD$ around the mean (line). MB = Mean Bias, RMSE = Root Mean Square Error, ML = Maximum Likelihood, MAP = Maximum a Posteriori.

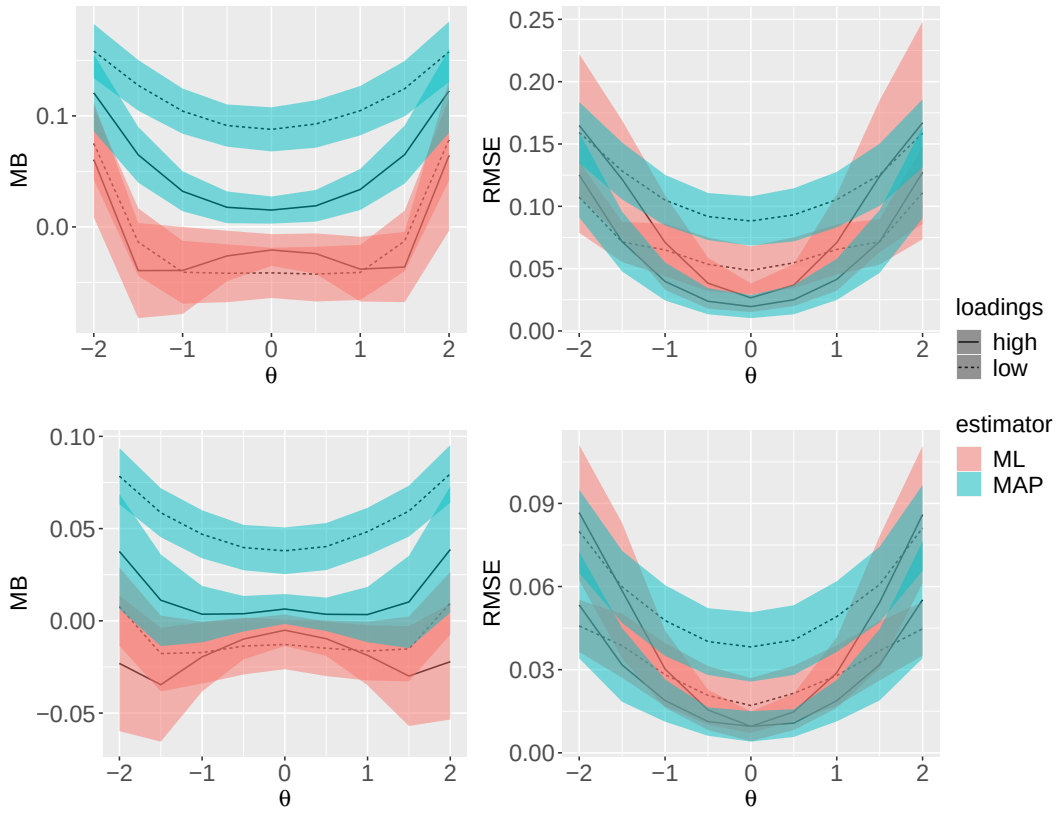


Figure 7: Bias of expected standard errors in simulation study 1 on standard error accuracy. The top row shows results for the short test (20 blocks) and the bottom row shows results for the long test (40 blocks). Shaded areas show $\pm 1SD$ around the mean (line). MB = Mean Bias, RMSE = Root Mean Square Error, ML = Maximum Likelihood, MAP = Maximum a Posteriori.

Discussion

As to be expected, the results of this simulation showed that bias was lower and *SE*s smaller for medium trait levels, longer tests and higher loadings. Thus, higher loadings and longer tests are recommended, because both trait estimates and their *SE*s are more accurate.

Regarding the comparison of estimators, especially with small loadings, *SE*s were more accurate for the MAP estimator than for the ML estimator. Similarly, Lin (2020) recommends using the MAP estimator based on a comparison of several trait estimators under various test designs with a large multivariate normal sample. In the current simulation, the accuracy of *SE*s for the MAP estimator with suboptimal test designs was underestimated by the information methods, that is, estimated *SE*s were larger than empirical *SE*s. Thus, the advantage of the MAP estimator in terms of precision might not be detectable in empirical applications.

Further, the results showed that observed test information was as accurate as expected test information. Thus, when only test level information is of interest, researchers can rely on the observed information at the trait estimate which saves computational time and resources. However, when block level information is of interest, expected information might still be preferred.

In this simulation, I focused on test design factors relevant for the level of information, keeping other design factors such as number of traits, trait correlations, and number of comparisons between mixed keyed items constant. Future studies varying these test design factors might yield more pronounced differences between information estimation methods.

Information Summaries for the Automated Assembly of MFC tests

On the one hand, standard errors are only partially informative about the accuracy of block information, because their computation involves summing across blocks. Moreover, the performance of the block information summaries was not evaluated so far. On the other hand, constructing MFC tests can be a combinatorial challenge, because it may not only involve maximizing information, but also balancing item keying and the numbers of items per trait and social desirability matching (e.g., Brown & Maydeu-Olivares, 2011; Wetzel et al., 2020). Therefore, automated test assembly (ATA) is particularly promising.

In automated test assembly (ATA), items are selected from a pool so that a criterion is maximized (or minimized) and certain restrictions are fulfilled (W. J. van der Linden, 2005). For example, information is maximized while keeping the number of items per trait equal. Practical applications of ATA include constructing parallel test forms with similar information curves or a test with peaked information at a certain trait level for selection

purposes. For example, employers might be interested in selecting all applicants who score two standard deviations above the mean. In contrast, in CAT, the focus is on maximizing information at the trait level of the individual respondent. For an introduction to ATA, see W. J. van der Linden (2005).

In the following part, I outline which ATA algorithms and block information summaries can be combined. Two further simulation studies were conducted to compare the performance of the block information summaries in conjunction with ATA algorithms. The simulation results can provide first insights as to which block information summaries (and potentially ATA algorithms) should be preferred for test construction.

Information Summaries for Mixed Integer Programming

Mixed integer programming (MIP) algorithms are the first choice for ATA, because they can find the optimal solution if it exists. Moreover, they can incorporate a *maximin* criterion, which has good properties and is particularly suited to IRT (W. J. van der Linden, 2005). For MIP, a test assembly problem has to be framed as a (constrained) linear optimization problem. Next, I describe how assembling an MFC test from a block pool can be framed for MIP with a block information summary as a relative maximin criterion.

First, $g = 1, \ldots, G$ trait levels are defined at which information is to be computed. In the multidimensional case, typically, a grid of trait levels is selected, for example, all combinations of -1, 0, and 1 across five traits (e.g., Debeer et al., 2020; Veldkamp, 2002). To obtain a relative criterion, for each grid point, the information summary s is summed across all K blocks and weighted by this sum for an arbitrary reference grid point, say θ_1 to obtain a weight w_g:

$$w_g = \frac{\sum_{k=1}^{K} s_k(\boldsymbol{\theta}_g)}{\sum_{k=1}^{K} s_k(\boldsymbol{\theta}_1)} \tag{13}$$

Whether block k is included in the test is encoded in a decision vector $\mathbf{x} = (x_1, \ldots, x_K)'$, taking on a value of 1 if the block is included and 0 otherwise. Then, the task is to find the values of $\mathbf{x}$ for which the summary y at reference point θ_1 is maximized while constraints ensure that the summary at the other points is close to proportional to their value in the block pool:

$$\text{maximize } y \tag{14}$$

subject to

$$\sum_{k=1}^{K} (s_k(\theta_g) x_k) - w_g y \geq 0 \quad \text{for all} \quad g \tag{15}$$

Additional constraints can be added to the ATA problem. The blocks' values on the constrained attributes are encoded in a $K \times N$ matrix $\mathbf{C}$ and the minimum values for the

constraints are encoded in a vector $\mathbf{d} = (d_1, \dots d_N)'$.

$$\sum_{k=1}^{K} c_{kn} x_k \geq d_n \quad \text{for all} \quad n \tag{16}$$

For example, if the first constraint is that the test should include at least 5 blocks measuring trait 1, the first column of $\mathbf{C}$ would take on a value of 1 for all blocks measuring trait 1 and 0 otherwise and $d_1 = 5$.

However, MIP methods are only applicable to optimization criteria that are linear across items (or blocks). In the multidimensional case, linear approximations to item information can be used (e.g., Debeer et al., 2020; Veldkamp, 2002), but linear approximation is not possible with MFC block information because there is no closed-form expression for it. Of the optimality criteria described above, only T-optimality can be used to construct MFC tests with MIP because it is additive (and accordingly linear) across blocks.

Information Summaries for Heuristics

Because T-optimality performed worst in MFC CAT simulations (Lin, 2020), I also investigate algorithms for ATA that can be used with A- and D-optimality, criteria that performed well in previous simulations (Brown, 2012; Lin, 2020; Mulder & van der Linden, 2009). These algorithms are heuristics which can be combined with all optimality criteria described above. In contrast to MIP methods, heuristics are guaranteed to find a solution, but the solution is not guaranteed to be optimal (W. J. van der Linden, 2005).

The simplest heuristics are constructive heuristics, which sequentially select a locally optimal item (or block). For example, Veldkamp (2002) compared the performance of a greedy heuristic for ATA with multidimensional items to that of MIP (with a linear approximation of item information). More sophisticated heuristics are local search heuristics which introduce randomness into the selection process to prevent the search from being trapped in a suboptimal space, often inspired by natural processes. For example, Olaru et al. (2015) compared amongst others a genetic algorithm and ant colony optimization for the assembly of a short scale. However, local search heuristics are more specifically tailored to a certain problem than MIP.

As outlined in the introduction, A- and D-optimality cannot be calculated for a single block. In ATA, they can only be used to extend an existing test. However, this "existing test" may be as small as three blocks (see Simulation 3). Hence, similar to some CAT scenarios, a small number of blocks selected by another method can be used as a starting point. Then, heuristics for A- and D-optimality can still be used to build the main part of the test.

Simulation Studies on Automated Test Assembly

Overview

Having examined theoretically how ATA algorithms and block information summaries can be combined for different purposes, their performance is compared in two simulation studies. As heuristics for A- and D-optimality can only be used for test extension, I conducted two separate simulation studies: Simulation Study 2 on Automated Test Construction and Simulation Study 3 on Automated Test Extension. Admittedly, test extension is not a typical ATA scenario so far, however the properties of MFC tests make it quite likely to occur in practice. In both simulations, the composition of the block pool was ideal with respect to the balancing of traits and item keying. Thus, the aim of the simulation studies is to gain a first impression of the performance of the criteria and algorithms in a simple setting.

Since local search heuristics are specifically tailored to certain problems, in this simulation, I use a simple greedy heuristic instead, to gain a broadly applicable estimate. This greedy heuristic sequentially selects the block with highest A- or D-optimality, weighted across trait levels. The results can serve as a benchmark of what might be achieved with these criteria and a more elaborate local search heuristic.

In both simulations, the performance of the proposed optimality criteria in conjunction with ATA algorithms is compared to that of mean block R^2, mean absolute loadings within blocks, and random block selection. The mean of absolute loadings within blocks serves as an approximation to the practice of selecting items (primarily) based on the size of their loadings. Block R^2 is calculated with the whole block pool as references set T, which makes it independent on the previously selected items. In this setting, the optimal solution for mean block R^2 and mean loadings is the one with the highest values on the respective criterion. Random block selection serves as a benchmark. Any algorithm should perform better than random block selection in order to be worth using.

Simulation Study 2: Simulation on Automated Test Construction

The first ATA simulation focuses on the assembly of a new test comparing the performances of MIP with T-optimality as a criterion, block R^2, mean loadings and random block selection.

Methods

In this simulation, an initial pool of 80 blocks is reduced to 1/4, that is, 20 blocks. The tests each measured 5 traits, with block size three. Across the block pool, 1/2 of pairwise item comparisons involved items keyed in different directions (i.e., one positive, one negative factor loading). Item intercepts μ_i were drawn from $U(-2, 2)$. Item loadings λ_i were drawn

from $U(.45, .95)$. Item uniquenesses ψ_i^2 were calculated as $1 - \lambda_i^2$. Errors were drawn from $N(0, \psi_i)$. The ranges of the item parameter distributions were larger than in simulation study 1 on *SE* accuracy so that the algorithms could improve trait recovery as compared to random block selection. Information was calculated over a grid of points. Trait levels were set at -1, 0, and 1 and fully crossed for the five traits. This yielded $3^5 = 243$ grid points.

Two factors of the ATA problem were varied: First, the *target* information curve was either proportional to that of the block pool or flat. For the flat target, all trait levels were weighted equally. Second, the problem was either unconstrained or constrained. In the unconstrained problem, the only constraint was test length. In the constrained problem, in addition, the blocks were selected such that the numbers of items per trait were equal, there were at least 2/3 of blocks including a negatively keyed item and at least one negatively keyed item per trait.

Algorithms

MIP with T-optimality For MIP with T-optimality (MIP T), a maximin criterion was chosen to select the combination of blocks so that T-optimality is maximal, while, across grid points, it is close to proportional to the target T-optimality. The MIP solver used was lpSolve with the R package lpSolveAPI (lp_solve et al., 2020) as an interface (see Diao & van der Linden, 2011, for an illustration of how to use lpSolveAPI for MIP with single-stimulus items).

Block R^2 To obtain one value per block, block R^2 was averaged across traits. For the weighted target, instead of using a maximin criterion, block R^2 was weighted across grid points by the sum of sampling variances (i.e., A-optimality) in the block pool. The 20 blocks with the highest weighted mean block R^2 were selected.

Mean Loadings For mean loadings, the blocks with the highest mean absolute loadings were selected.

Random For random block selection, the blocks were selected randomly.

For all algorithms, the constrained problem was implemented as an MIP problem. This was possible because selecting the blocks with the highest values on a criterion (for block R^2 and mean loadings) is equivalent to selecting the blocks that maximize the sum of this criterion across blocks. For random, in the constrained problem, a random value drawn from $U(0, 1)$ was used as a criterion.

Procedure 500 replications were conducted. All data simulation and analysis was carried out in R, using the same R packages as in simulation study 1 on *SE* accuracy, in addition to lpSolveAPI. First, item parameters were drawn. Second, information was estimated for the grid points. Third, a test was assembled by each of the four algorithms. Fourth, for the weighted target, true trait levels were drawn from a multivariate normal distribution with a mean vector of 0 and covariances based on meta-analytic correlations between the Big Five (Table 1) for 500 respondents. For the equal target, the grid points served as trait levels. To achieve a comparable sample size to the weighted target, each grid point was duplicated, yielding 486 respondents. Responses for these respondents on the block pool were simulated. Fifth, trait levels were estimated as MAP estimates for each of the four assembled tests based on true item parameters and the Big Five correlations. To assess trait recovery, three summary measures were calculated: the correlation between true and estimated traits $r(\theta, \hat{\theta})$, RMSE (Eq. 12, with $\xi = \theta$), and mean absolute bias (MAB):

$$MAB(\theta) = \frac{\sum_{r=1}^{R} |\hat{\theta} - \theta|}{R} \tag{17}$$

Trait recovery was summarized via means and *SD*s by condition and variance explanation for the contrasts between conditions was calculated within an ANOVA framework. For the ANOVA, $r(\theta, \hat{\theta})$ was Fisher Z transformed.

Results

All MIP models converged. Trait recovery was worse for random block selection (e.g., mean MAB = 0.37) than for MIP T, mean loadings and block R^2 together (mean MAB = 0.30, Table 4, Figure 8), explaining 32% to 46% of total variance (Table 5). Descriptively, recovery was slightly worse for mean loadings (e.g., mean MAB = 0.31) than for MIP T and block R^2 (mean MAB = 0.30). However, this difference only explained 1% of variance and was negligible in absolute size. Trait recovery was worse for the equal (mean MAB = .32) than for the weighted target (mean MAB = .31), explaining 30% of variance in $r(\theta, \hat{\theta})$ and 2% to 3% in MAB and RMSE. The difference between free and constrained problems and all interactions were negligible.

Table 4: Mean trait recovery by condition in simulation study 2 on test construction

Algorithm	Constraints	Target	$r(\theta, \hat{\theta})$		$r(\theta, \hat{\theta})^2$		MAB		RMSE	
Equal	Constrained	MIP T	0.89	(0.02)	0.80	(0.03)	0.30	(0.02)	0.14	(0.02)
		Block R^2	0.89	(0.02)	0.80	(0.03)	0.29	(0.02)	0.13	(0.02)
		Mean Loadings	0.88	(0.02)	0.78	(0.03)	0.31	(0.02)	0.15	(0.02)
		Random	0.84	(0.03)	0.70	(0.05)	0.36	(0.03)	0.20	(0.03)
	Free	MIP T	0.89	(0.02)	0.79	(0.04)	0.30	(0.03)	0.14	(0.03)
		Block R^2	0.90	(0.02)	0.80	(0.04)	0.29	(0.03)	0.13	(0.02)
		Mean Loadings	0.88	(0.03)	0.78	(0.05)	0.31	(0.03)	0.15	(0.03)
		Random	0.83	(0.04)	0.69	(0.07)	0.36	(0.04)	0.21	(0.05)
Weighted	Constrained	MIP T	0.92	(0.01)	0.85	(0.02)	0.31	(0.02)	0.15	(0.02)
		Block R^2	0.92	(0.01)	0.85	(0.02)	0.31	(0.02)	0.15	(0.02)
		Mean Loadings	0.92	(0.01)	0.84	(0.02)	0.32	(0.02)	0.16	(0.02)
		Random	0.89	(0.02)	0.78	(0.04)	0.37	(0.03)	0.22	(0.03)
	Free	MIP T	0.92	(0.02)	0.85	(0.03)	0.31	(0.03)	0.15	(0.03)
		Block R^2	0.92	(0.01)	0.85	(0.03)	0.30	(0.03)	0.15	(0.03)
		Mean Loadings	0.92	(0.02)	0.84	(0.03)	0.32	(0.03)	0.16	(0.03)
		Random	0.88	(0.03)	0.78	(0.05)	0.37	(0.04)	0.22	(0.05)

Note. MAB = Mean Absolute Bias, RMSE = Root Mean Squared Error, MIP T = Mixed Integer Programming with T-optimality. Standard deviations are given in parentheses.

Table 5: Variance in trait recovery explained in % by algorithm, target and constraints in simulation study 2 on test construction

Factor	$r(\theta, \hat{\theta})$	MAB	RMSE
Algorithm vs. Random	32	46	46
Info vs. Mean Loadings	1	1	1
Weighted vs. Equal	30	2	3
Residuals	37	51	50

Note. MAB = Mean Absolute Bias, RMSE = Root Mean Squared Error, MIP T = Mixed Integer Programming with T-optimality. $r(\theta, \hat{\theta})$ was Fisher Z transformed.

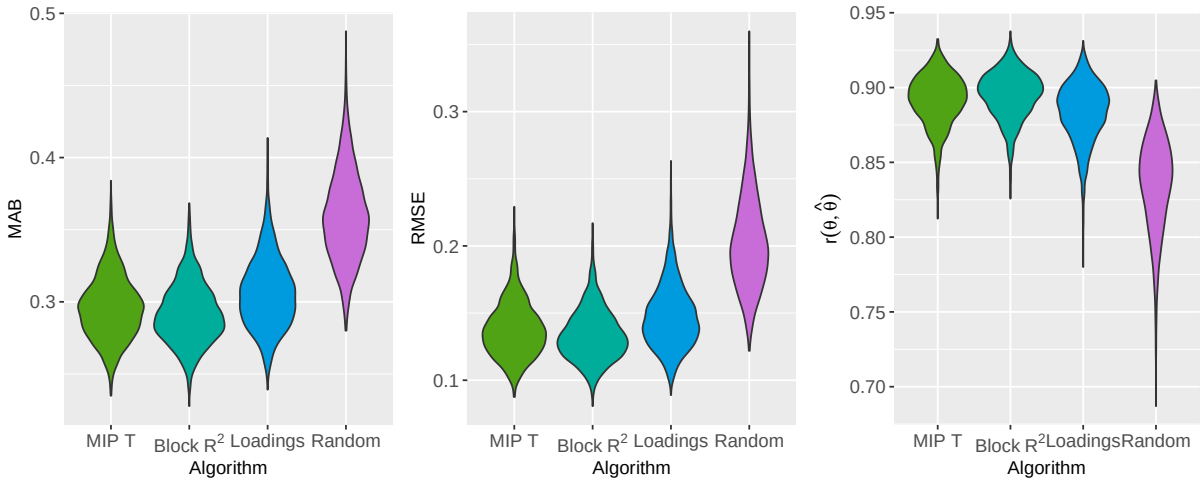


Figure 8: Trait recovery by algorithm, for trait levels weighted equally and the constrained problem, in simulation study 2 on test construction. MIP T = Mixed Integer Programming with T-optimality, MAB = Mean Absolute Bias, RMSE = Root Mean Square Error.

Discussion

The results of this simulation showed that MIP T, block R^2, and mean loadings performed better than random block selection, thus they are worth using. However, they performed on par with each other. Surprisingly, at least in the limited conditions examined, the mean loadings proved as a good alternative to the information summaries to be used in test construction. The mean loadings do not require any considerable computational effort (besides model fitting which is needed for any method). Block R^2 might be useful when information is of interest for each trait individually, although this was not examined in this simulation which focused on increasing precision for all traits simulateneously. T-optimality in conjuction with an MIP algorithm should be preferred if the other advantages of MIP are needed, such as matching a test information curve for selection purposes or parallel

test forms. In these settings, a minimax criterion has better properties than a weighted criterion (W. J. van der Linden, 2005).

The target information surface with trait levels weighted equally resulted in lower trait recovery than the target that was proportional to information in the block pool. This is not surprising because in the latter case there are both many items and many persons with the same trait level.

Further, the composition of the block pool was rather ideal with all combinations of three out of five traits occurring equally often and 1/2 of comparisons between mixed keyed items. Varying the block pool or constraining the ATA problem should have similar effects on the performance of the algorithms. In this simulation, both a constrained and an unconstrained ATA problem were simulated. This did not result in differences with respect to trait recovery or to the performance of the algorithms. Probably, the constraints were mild and well suited to the block pool.

This simulation only examined a limited set of conditions. Specifically, block size was fixed to three and five traits were simulated. Although these settings might be representative for some applied tests (e.g., Brown & Maydeu-Olivares, 2011; Wetzel & Frick, 2020), more research is needed on how well the proposed methods perform under different test designs and for more complex ATA problems.

Simulation Study 3: Simulation on Automated Test Extension

In the second ATA simulation the extension of a test was simulated in order to compare the performances of a greedy algorithm for A- and D-optimality, MIP with T-optimality, block R^2, mean loadings and random block selection. It extends the previous simulation by including A- and D-optimality, which performed better than T-optimality in previous ATA and CAT simulations. The conditions with a constrained ATA problem were dropped because developing a sophisticated greedy algorithm or local search heuristic is beyond the scope of this paper (for examples of such algorithms, see Kreitchmann et al., 2021; Luecht, 1998; Olaru et al., 2015). In addition, note, that the differences in recovery between the free and constrained problems were negligible in the previous Simulation 2 on test construction.

Methods

A basis test of three blocks was extended by 17 blocks out of a pool of 77 blocks, yielding a final test size of 20 blocks. The only constraint was test length. Besides that, the simulation design and procedure were identical to simulation study 2 on automated test construction. The three blocks that served as a basis were composed of items loading on Traits a) 1-2-3, b) 1-2(negatively keyed)-4, and c) 1-2(negatively keyed)-5. This was sufficient to obtain a positive-definite test information matrix, but it left enough room for performance differences between the algorithms.

Algorithms

Greedy algorithm based on A-optimality For the greedy algorithm based on A-optimality (Greedy A), for each block not in the current test, A-optimality achieved by adding this block to the current test was calculated for each grid point. For the weighted target, A-optimality was weighted by A-optimality in the block pool for this grid point. Weighted or unweighted A-optimality was then averaged across grid points, yielding mean A-optimality. The block with the lowest mean A-optimality was added to the current test. This procedure was repeated until the final test length of 20 blocks was reached.

Greedy algorithm based on D-optimality The greedy algorithm based on D-optimality (Greedy D) was identical to that for A-optimality, except that D-optimality was used instead.

MIP T The MIP algorithm with T-optimality was identical to simulation study 2 on test construction, except that T-optimality for the basis test of three blocks was added.

Mean loadings and Random The algorithms for mean loadings and random block selection were identical to simulation study 2 on test construction.

Results

All MIP models converged. The information-based algorithms, block R^2 and mean loadings together performed better (e.g., mean MAB = 0.30) than random block selection (mean MAB = 0.35), explaining 22 to 32 % of total variance (Table 7). Mean loadings and MIP T performed slighlty worse (e.g., mean MAB = .30) than Greedy A, Greedy D and block R^2 (mean MAB = .29, Table 6, Figure 9). However the absolute size of this difference was negligible. Moreover, the difference between mean loadings and the information-based algorithms explained only 1% of variance (Table 7). Recovery for A- and D-optimality showed slightly smaller variance (*SD* MAB = 0.02) than for MIP T, mean loadings and block R^2 (*SD* MAB = 0.03, Table 7, Figure 9). Trait recovery was worse for the equal (mean MAB = .31) than for the weighted target (mean MAB = .30), explaining 31% of variance in $r(\theta, \hat{\theta})$ and 4% to 6% in MAB and RMSE. All interactions were negligible.

Table 6: Mean trait recovery by algorithm in simulation study 3 on test extension

Target	Algorithm	$r(\theta, \hat{\theta})$		$r(\theta, \hat{\theta})^2$		MAB		RMSE	
Equal	Greedy A	0.90	(0.01)	0.81	(0.03)	0.28	(0.02)	0.13	(0.02)
	Greedy D	0.90	(0.02)	0.81	(0.03)	0.28	(0.02)	0.13	(0.02)
	MIP T	0.90	(0.02)	0.80	(0.04)	0.29	(0.03)	0.13	(0.03)
	Block R^2	0.90	(0.02)	0.81	(0.04)	0.29	(0.03)	0.13	(0.02)
	Mean Loadings	0.89	(0.03)	0.79	(0.04)	0.30	(0.03)	0.14	(0.03)
	Random	0.85	(0.04)	0.72	(0.06)	0.34	(0.04)	0.19	(0.04)
Weighted	Greedy A	0.93	(0.01)	0.86	(0.02)	0.30	(0.02)	0.14	(0.02)
	Greedy D	0.93	(0.01)	0.86	(0.02)	0.30	(0.02)	0.15	(0.02)
	MIP T	0.92	(0.02)	0.85	(0.03)	0.31	(0.03)	0.15	(0.03)
	Block R^2	0.93	(0.01)	0.86	(0.03)	0.30	(0.03)	0.15	(0.03)
	Mean Loadings	0.92	(0.02)	0.85	(0.03)	0.31	(0.03)	0.16	(0.03)
	Random	0.89	(0.02)	0.80	(0.04)	0.36	(0.04)	0.20	(0.04)

Note. MAB = Mean Absolute Bias, RMSE = Root Mean Squared Error, A = A-optimality, D = D-optimality, MIP T = Mixed Integer Programming with T-optimality. Standard deviations are given in parentheses.

Table 7: Variance in trait recovery explained in % by algorithm and target in simulation study 3 on test extension

Factor	$r(\theta, \hat{\theta})$	MAB	RMSE
Algorithm vs. Random	22	31	32
Info vs. Mean Loadings	1	1	1
Weighted vs. Equal	31	4	6
Residuals	46	63	61

Note. MAB = Mean Absolute Bias, RMSE = Root Mean Squared Error. $r(\theta, \hat{\theta})$ was Fisher Z transformed.

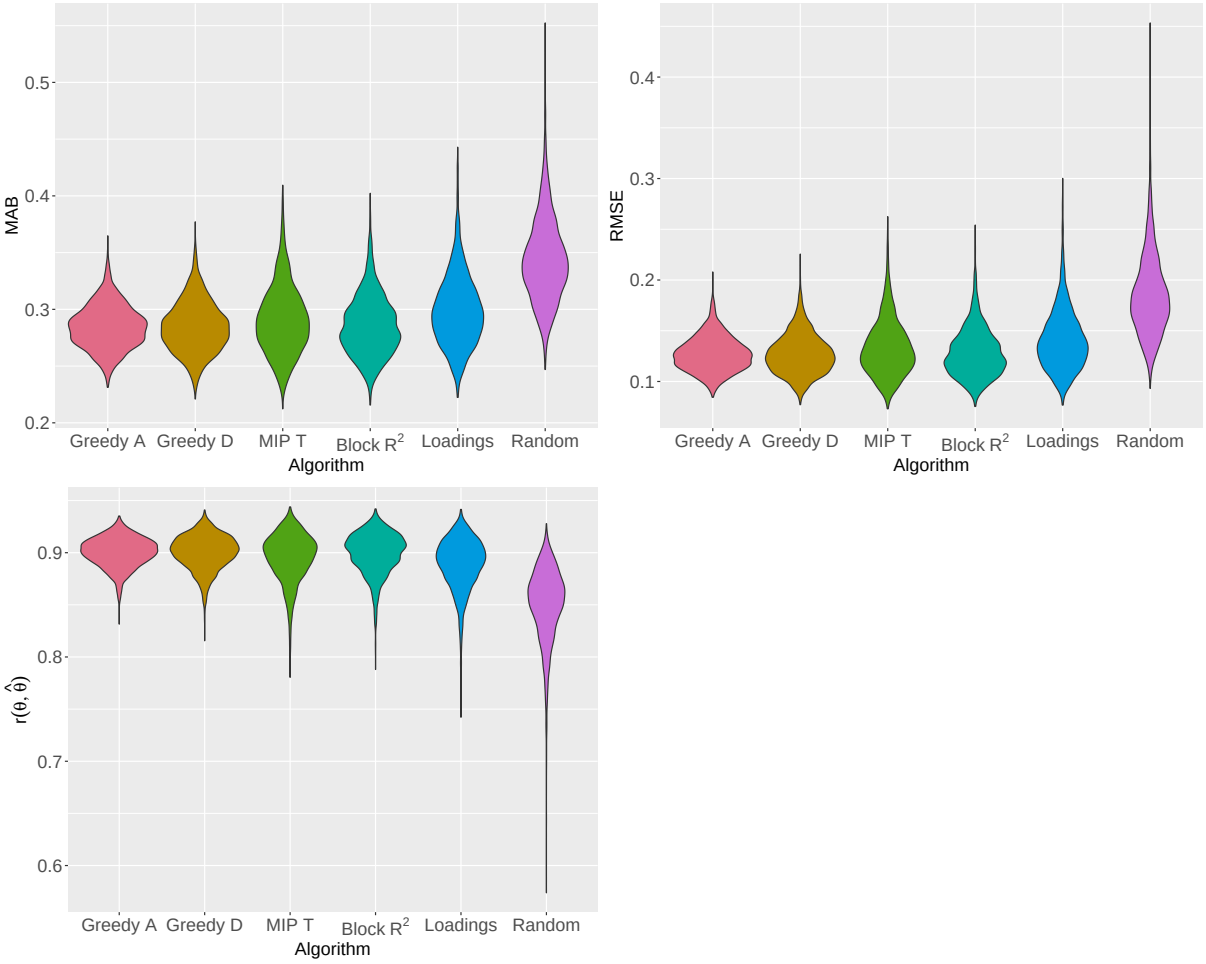


Figure 9: Trait recovery by algorithm, for trait levels weighted equally in simulation study 3 on test extension. MIP = Mixed Integer Programming, A = A-optimality, D = D-optimality, T = T-optimality, MAB = Mean Absolute Bias, RMSE = Root Mean Square Error.

Discussion

The results of simulation study 3 on test extension confirmed the results of simulation study 2 on test construction. The information-based algorithms and mean loadings performed all on par but better than random block selection. Thus, also A- and D-optimality proved useful for test construction. Further, the greedy algorithms showed slightly smaller variance. Like the other algorithms, heuristics can be adapted to include more constraints in addition to test length. The performance of a greedy algorithm gives a lower-bound estimate to that of a more elaborate heuristic. Thus, A- and D-optimality are promising information summaries for the development of a local search heuristic or a more elaborate constructive heuristic. D-optimality is computationally less intensive than A-optimality.

General Discussion

In this paper, I have shown how Fisher information in Thurstonian IRT models can be calculated on the block level. Because Fisher information for a block is a non-invertible matrix, I have proposed several indices to summarize block information: block R^2, A-, D-, and T-optimality. A simulation study showed that observed and expected standard errors based on block information were accurate. Two other simulation studies showed that the proposed information summaries in conjunction with different test assembly algorithms can be used to create tests that are more reliable than those assembled by chance. In addition, the mean of absolute item loadings within a block proved to be a good alternative, albeit it does not allow to weigh precision by trait level. In the following, I will outline possible applications of block information in research and practice.

Statistical Improvements

With Fisher information on the block level, unbiased expected and observed *SE*s can be obtained for block sizes > 2 (Yousfi, 2020). Although the overestimation of reliability based on information for binary outcomes of pairwise outcomes is small (Brown & Maydeu-Olivares, 2011; Frick et al., 2021), it increases with increasing block size. Block information allows to calculate unbiased optimality criteria that can be used in test construction. Future research could compare the proposed method to results obtained with the simplifying assumption of local independence.

Focus on the Block Level

The proposed method for estimating Fisher information for Thurstonian IRT blocks extends previous methods by a focus on the block level. First, a focus on the block level as compared to the item level better reflects the response options available to participants and thus captures the relative nature of MFC responses.

Second, relatedly, MFC tests have an inseparable design. Thus, all traits measured in a block mutually interact at influencing ranking preferences and accordingly Fisher information. As illustrated in the section on block information plots, the information summaries and plots proposed in this manuscript can account for and visualize those mutual influences.

Third, focusing on the block level allows capturing item interactions. The estimation of Thurstonian IRT models became possible when rank orders were recoded as binary outcomes whose dependencies could be modeled in a structural equation framework (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares, 1999; Maydeu-Olivares & Brown, 2010). However, items in MFC blocks were sometimes observed to function differently between different block compositions (Lin & Brown, 2017) or response contexts, for example, simulated

low- and high-stakes contexts (Lee & Joo, 2021). At least as long as the extent of item interactions and item parameter invariance between different block compositions is unclear, a focus on the block level seems a useful supplement.

Investigating the MFC format

Block-level Fisher information can yield further insights into how item content and statistical pecularities of the MFC format influence the precision of trait estimates. For example, in simulations with MFC tests comprised of all positively keyed items, trait recovery was decreased (Bürkner et al., 2019; Schulte et al., 2020) and evidence of ipsativity was observed (Frick et al., 2021). Comparing block information between mixed and equally keyed blocks might yield further insights into how item keying contributes to the recovery of normative trait levels.

Moreover, differences in item social desirability might lead to certain rank orders being more frequent. For example, it has been reported that agreement as to which rank order should be preferred increased the more the items within blocks differed in their social desirability (Hughes et al., 2021). Certain rank orders being more frequent due to socially desirable responding might make the whole block less informative about the content traits. Further empirical studies could investigate the effect of item matching on the size of block information.

Benefits for MFC Test Assembly

The newly proposed information summaries can be used to assemble MFC tests that maximize trait estimation precision. For manual test assembly, block information is easier to interpret and incorporate than standardized item loadings which may differ by binary outcome (e.g., Wetzel & Frick, 2020). Moreover, in the current simulations on test construction and test extension, the mean loadings within blocks performed almost as good as block information. If trait-level information is of interest, for example because the measurement precision for a single trait should be increased, block R^2 is the preferred information summary in an interpretable metric.

Further, the current simulations showed that and illustrated how block information can be used for the automated assembly of fixed tests. Considering the complexity of assembling MFC tests, including balancing of traits, item keying and item desirability, automated test assembly might prove particularly valuable. Examining minimal restrictions for test composition, the current simulations serve as a proof of concept showing that the proposed block information summaries can be used for ATA. The full advantages might be observed with more complex restrictions, more specific test information goals, and more sophisticated heuristics.

Lastly, the proposed summaries can be used in computerized adaptive testing, where

tests are assembled for each participant, based on their answers. In later stages of computerized adaptive testing, A- and D-optimality can be used and might be preferable. These criteria performed best in a simulation on computerized adaptive testing where MFC blocks were assembled from separate items (Lin, 2020).

A drawback of using information for blocks instead of items is that whole blocks have to be removed from the item pool. The selection of whole blocks costs more items and therewith more time of participants and more research funds than newly assembling blocks from separate items. Future research and applications will show how practicable and necessary this procedure is.

In this paper, I have proposed a method to estimate Fisher information for multidimensional forced-choice blocks assessed with the Thurstonian IRT model. Several ways to summarize the information matrix for test construction were presented and evaluated. I hope this paper will improve the construction of MFC tests and encourage further investigation of their properties.

References

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics* (Computer Software; Version 2.3). https://CRAN.R-project.org/package=gridExtra

Brown, A. (2012). *Multidimensional CAT in non-cognitive assessments*. Conference of the International Test Comission, Amsterdam.

Brown, A. (2016). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research*, *51*(2-3), 345–356. https://doi.org/10.1080/00273171.2016.1150152

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-Degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148. https://doi.org/10.1177/1094428116668036

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

Brown, A., & Maydeu-Olivares, A. (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

Debeer, D., van Rijn, P. W., & Ali, U. S. (2020). Multidimensional Test Assembly Using Mixed-Integer Linear Programming: An Application of Kullback–Leibler Information. *Applied Psychological Measurement*, *44*(1), 17–32. https://doi.org/10.1177/0146621619827586

Diao, Q., & van der Linden, W. J. (2011). Automated Test Assembly Using lp_Solve Version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398–409. https://doi.org/10.1177/0146621610392211

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., 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

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). *Mvtnorm: Multivariate Normal and t Distributions (1.1-0)* (Computer Software; Version 1.1-0). http://CRAN.R-project.org/package=mvtnorm

Gilbert, P., & Varadhan, R. (2019). *numDeriv: Accurate Numerical Derivatives* (Computer Software; Version 2016.8-1.1). https://CRAN.R-project.org/package=numDeriv

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

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

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

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

Ippel, L., & Magis, D. (2020). Efficient Standard Errors in Item Response Theory Models for Short Tests. *Educational and Psychological Measurement*, *80*(3), 461–475. https://doi.org/10.1177/0013164419882072

Joo, S.-H., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model: GGUM-RANK item and test information functions. *Journal of Educational Measurement*, *55*(3), 357–372. https://doi.org/10.1111/jedm.12183

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

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

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Lee, P., & Joo, S.-H. (2021). A New Investigation of Fake Resistance of a Multidimensional Forced-Choice Measure: An Application of Differential Item/Test Functioning. *Personnel Assessment and Decisions*, *7*(1). https://doi.org/10.25035/pad.2021.01.004

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

lp_solve, Konis, K., & Schwendiger, F. (2020). *lpSolveAPI: R Interface to 'lp_solve' Version 5.5.2.0* (Computer Software; Version 5.5.2.0-17.7). https://CRAN.R-project.org/package=lpSolveAPI

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

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

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

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

Paek, I., & Cai, L. (2014). A Comparison of Item Parameter Standard Error Estimation Procedures for Unidimensional and Multidimensional Item Response Theory Modeling. *Educational and Psychological Measurement*, *74*(1), 58–76. https://doi.org/10.1177/0013164413500277

R Core Team. (2020). *R: A language and environment for statistical computing (3.6.3)* (Computer Program and Language; Version 3.6.3). Vienna, Austria. https://www.R-project.org/

Revelle, W. (2019). *Psych: Procedures for Personality and Psychological Research (1.8.12)* (Computer Software; Version 1.8.12). Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

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

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

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

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

Veldkamp, B. P. (2002). Multidimensional Constrained Test Assembly. *Applied Psychological Measurement*, *26*(2), 133–146. https://doi.org/10.1177/01421602026002002

Weston, S. (2017). *doMPI: Foreach Parallel Adaptor for the Rmpi Package* (Computer Software; Version 0.2.2). https://CRAN.R-project.org/package=doMPI

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

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer. https://ggplot2.tidyverse.org

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

This dissertation was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group “Statistical Modeling in Psychology" (SMiP).