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

Susanne Frick[1], Anna Brown[2], and 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

Author Note

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.

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

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

Word count: 13903

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.

insert Figure 1 about here

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 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:

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

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. $\Phi(x)$ 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):

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

where $\phi(x)$ 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 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 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.

insert Table 1 about here

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 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.

insert Tables 2 and 3 about here

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 ($r(\theta, \hat{\theta})$), 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 data[3]. 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 x 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 × Trait Correlations

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

Items per Trait × 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 × trait correlation combination. Second, the same item parameters were used for one replication within one block size × 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 dependencies[2] 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–loading^2.

Errors were sampled for each person on each item from $N(0, \text{uniqueness})$. 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.

insert Table 4 about here

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 $\hat{\eta}_d$, respectively, MAB and MSE were defined as follows:

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

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

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:

$$r_{empirical} = \frac{Var(\hat{\eta}) - Mean(SE^2)}{Var(\hat{\eta})}. \tag{5}$$

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.

insert Tables 5, 6 and 7 about here

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 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.

insert Tables 8 and 9 and Figure 2 about here

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.

insert Tables 10 and 11 and Figure 3 about here

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 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.

insert Table 12 about here

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: 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).

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 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).

insert Table 13 about here

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/≤ once a month/2-4 times a month/2-3 times a week/≥ 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.

insert Table 14 about here

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 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 (r_{MFC} = –.08, r_{TF} = –.41, difference in Fisher Z = 0.33) and agreeableness with quality of life (r_{MFC} = .11, r_{TF} = .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 (r_{MFC}= .14; r_{TF} = .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 (Rel_{MFC} = .83, Rel_{TF} = .89, difference in Fisher Z = .23).

insert Table 15 about here

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.

Footnotes

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.

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).

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.

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.

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.

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

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.

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.

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*(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

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.

Table 2

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

			Equal					Unequal 1(2)				
Block-size	Mixed com-parisons	Item keying	Trait 1	2	3	4	5	Trait 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

			Equal	Unequal 1(2)														
Block-size	Mixed com-parisons	Item keying	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.

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

Table 5

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

Hyp.	Factor / Contrast	$r(\theta, \hat{\theta})$	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	$r(\theta, \hat{\theta})$		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.

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.

Table 10

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

Hyp.	Factor	$r(\theta, \hat{\theta})$	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	$r(\theta, \hat{\theta})$		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.

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.

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.

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	≤ .001
	A	-0.08	-0.41	0.01	0.17	-0.08	-0.44	-0.33	medium	7.36	≤ .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	≤ .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	≤ .001
	C	0.21	0.35	0.04	0.12	0.21	0.36	-0.14	small	-3.03	≤ .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	≤ .001
	A	0.11	0.42	0.01	0.17	0.11	0.44	-0.30	medium	-6.78	≤ .001
	C	0.35	0.40	0.12	0.16	0.36	0.42	-0.05	negligible	-1.16	0.20
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	N	0.08	0.05	0.01	0.00	0.08	0.05	0.03	negligible	0.53	0.35

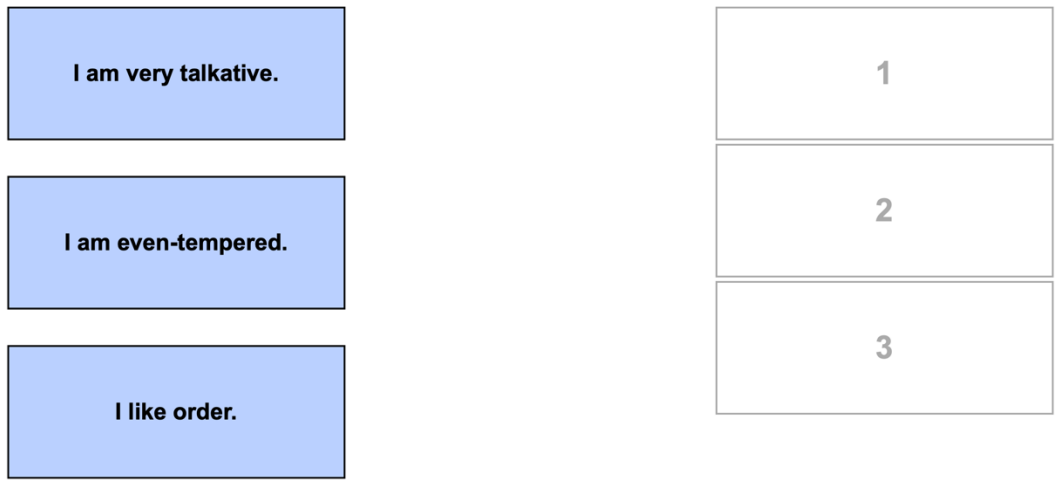
years											
	C	-0.08	-0.07	0.01	0.01	-0.08	-0.07	0.01	negligible	-0.26	0.39
Divorced	N	0.02	0.03	0.00	0.00	0.02	0.03	0.00	negligible	-0.08	0.40
	C	-0.04	-0.06	0.00	0.00	-0.04	-0.06	-0.02	negligible	0.43	0.36
Exercise regularly	E	0.07	0.12	0.00	0.01	0.07	0.12	-0.05	negligible	-0.99	0.25
	C	0.05	0.13	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
	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
	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
	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

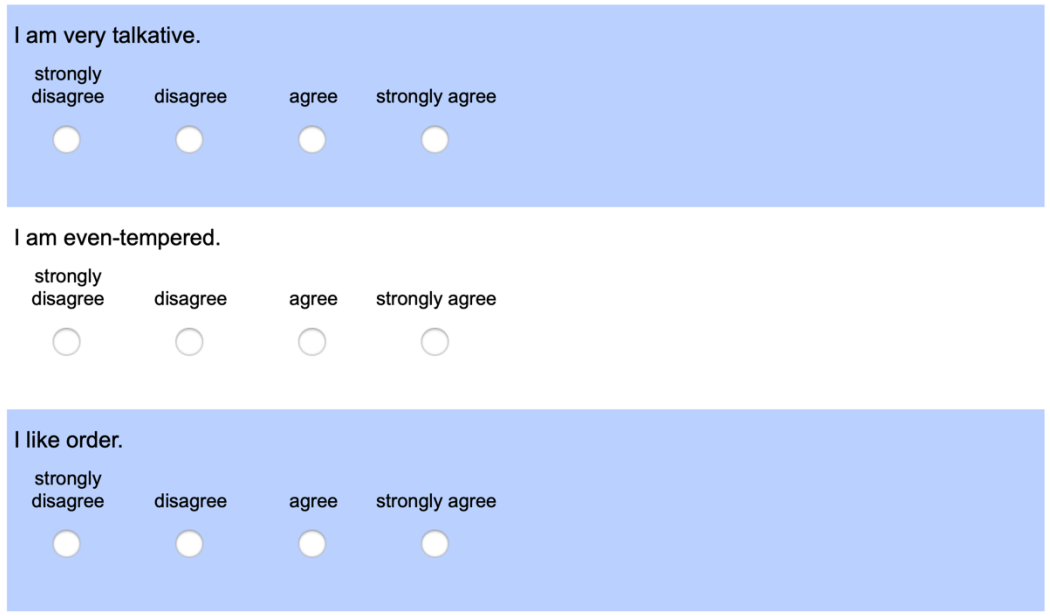
C	0.04	0.01	0.00	0.00	0.04	0.01	0.02	negligible	0.22	0.39

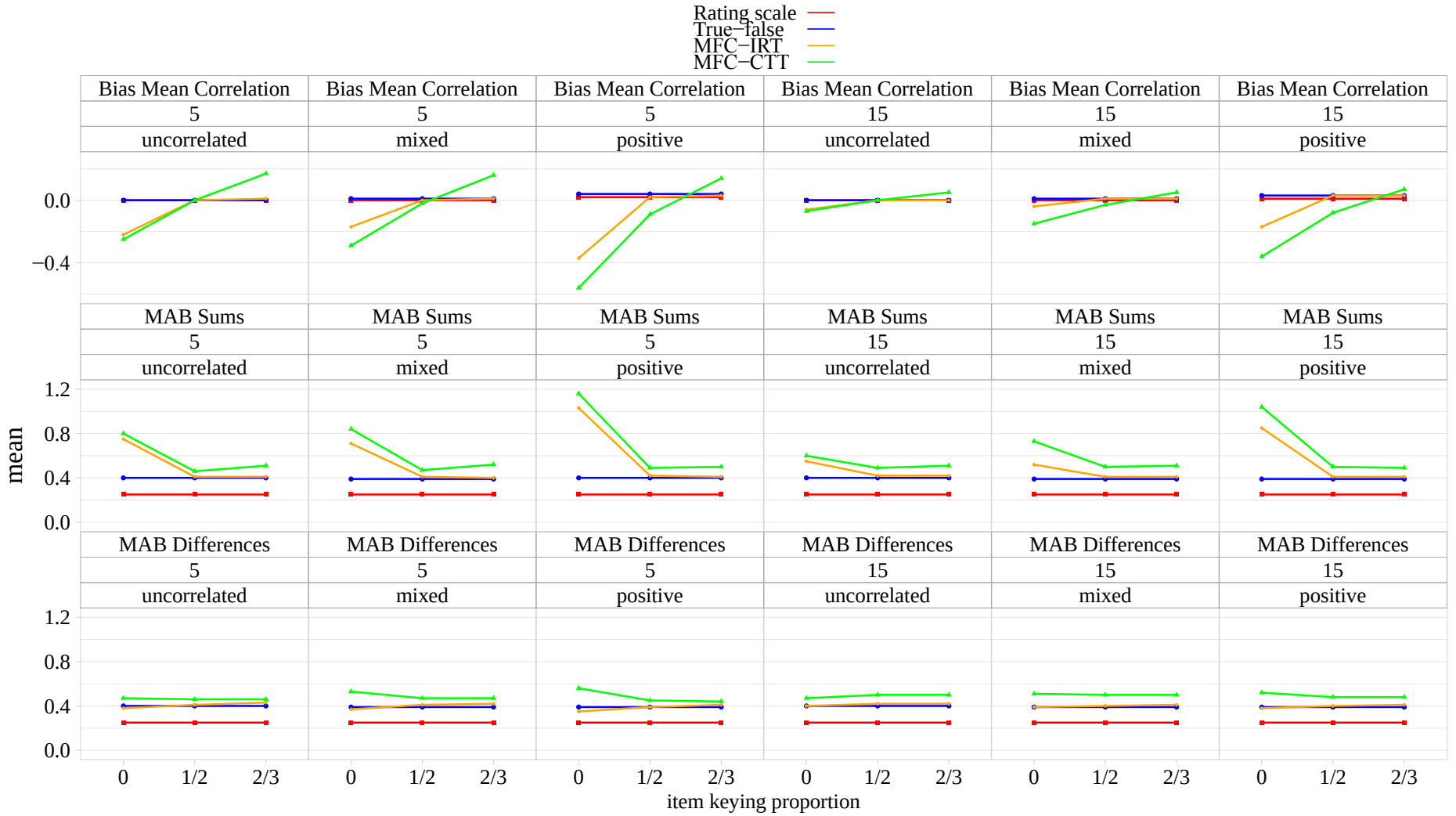
Note. MFC = multidimensional forced-choice format. TF = true-false format, N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness, CES-D short form = Center for Epidemiologic Studies–Depression Scale, SWLS = Satisfaction with Life Scale, WHO-QoL BREF = World Health Organization Quality of Life BREF. Only correlations that went in the predicted direction for both MFC and TF are shown.

A


Please rank the statements according to how well they describe you from *most like you* (1) to *least like you* (3).
I am very talkative.
I am even-tempered.
I like order.
1
2
3

B


Please select the answer that best corresponds to your agreement or disagreement to the following statements.
I am very talkative.
strongly disagree
disagree
agree
strongly agree
I am even-tempered.
strongly disagree
disagree
agree
strongly agree
I like order.
strongly disagree
disagree
agree
strongly agree


Rating scale
True−false
MFC−IRT
MFC−CTT
Bias Mean Correlation
MAB Sums
MAB Differences
5
15
uncorrelated
mixed
positive
mean
item keying proportion
0
1/2
2/3
0.0
−0.4
0.4
0.8
1.2

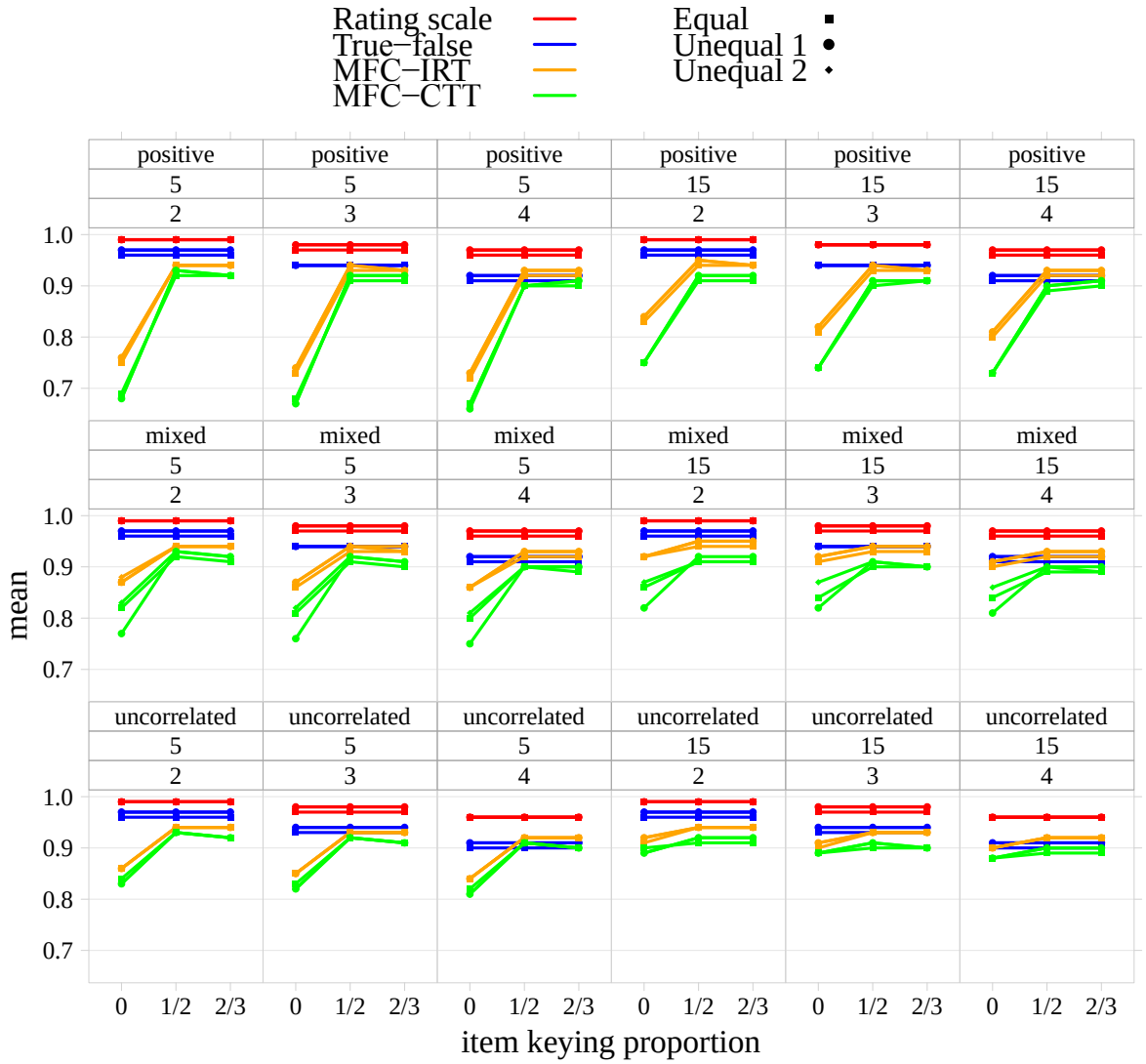

Rating scale
True−false
MFC−IRT
MFC−CTT
Equal
Unequal 1
Unequal 2
positive
mixed
uncorrelated
5
15
2
3
4
mean
1.0
0.9
0.8
0.7
0
1/2
2/3
item keying proportion

Figure captions

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.

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.

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.