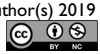
Article

On the Statistical and Practical Limitations of Thurstonian IRT Models

Educational and Psychological Measurement
2019, Vol. 79(5) 827–854



DOI: 10.1177/0013164419832063
journals.sagepub.com/home/epm


Paul-Christian Bürkner[1], Niklas Schulte[1] and Heinz Holling[1]

Abstract

Forced-choice questionnaires have been proposed to avoid common response biases typically associated with rating scale questionnaires. To overcome ipsativity issues of trait scores obtained from classical scoring approaches of forced-choice items, advanced methods from item response theory (IRT) such as the Thurstonian IRT model have been proposed. For convenient model specification, we introduce the thurstonianIRT R package, which uses Mplus, lavaan, and Stan for model estimation. Based on practical considerations, we establish that items within one block need to be equally keyed to achieve similar social desirability, which is essential for creating forced-choice questionnaires that have the potential to resist faking intentions. According to extensive simulations, measuring up to five traits using blocks of only equally keyed items does not yield sufficiently accurate trait scores and inter-trait correlation estimates, neither for frequentist nor for Bayesian estimation methods. As a result, persons' trait scores remain partially ipsative and, thus, do not allow for valid comparisons between persons. However, we demonstrate that trait scores based on only equally keyed blocks can be improved substantially by measuring a sizable number of traits. More specifically, in our simulations of 30 traits, scores based on only equally keyed blocks were non-ipsative and highly accurate. We conclude that in high-stakes situations where persons are motivated to give fake answers, Thurstonian IRT models should only be applied to tests measuring a sizable number of traits.

Keywords

forced-choice format, Thurstonian IRT model, ipsative data, multidimensional IRT, Stan, lavaan, Mplus, R

[1]University of Münster, Münster, Germany

Corresponding Author:
Paul-Christian Bürkner, Institute of Psychology, Westfälische Wilhelms-Universität Münster, Fliednerstr. 21, 48149 Münster, Germany.
Email: paul.buerkner@gmail.com

Psychometric questionnaires have long used rating scales as a response format from which to glean information from, but rating items are prone to several response biases such as social desirability, acquiescence, and extremity bias (Wetzel, Böhnke, & Brown, 2016). A promising alternative to rating items with Likert scales has been to use questionnaires with forced-choice (FC) items, as FC scales may solve these problems (e.g., Brown & Bartram, 2013; Hontangas et al., 2015; Paulhus, 1991; Saville & Willson, 1991). Application contexts best suited to FC items are high-stakes situations in which respondents are motivated to give fake answers, such as in personnel selection (Christiansen, Burns, & Montgomery, 2005; Stewart, Darnold, Zimmerman, Parks, & Dustin, 2010) and in studies where respondents have varying levels of the aforementioned response styles—for example, in cross-cultural research (He, Bartram, Inceoglu, & Vijver, 2014; Johnson, Kulesa, Cho, & Shavitt, 2005; Ross & Mirowsky, 1984; Rudmin & Ahmadzadeh, 2001). FC scales may also be applicable in a number of other fields, such as value measurement (Döring, Blauensteiner, Aryus, Drögekamp, & Bilsky, 2010; J. A. Lee, Soutar, & Louviere, 2008; Schwarz & Cieciuch, 2016) and market research (Parvin & Wang, 2014).

Yet traditional scoring methods for FC items yield ipsative person parameters; that is, the score of an individual depends on the individual's score on other variables. Derived scores are therefore inappropriate for interindividual comparisons (Cattell, 1944; Hicks, 1970). A discussion of related psychometric problems can be found in Baron (1996), while issues in employee selection are detailed in Johnson, Wood, and Blinkhorn (1988) and Meade (2004).

To overcome these problems, a new scoring method for FC items has recently been introduced: Thurstonian item response theory (T-IRT) models (Brown & Maydeu-Olivares, 2011). In contrast to traditional scoring methods for FC items, estimates obtained from T-IRT models are designed to be non-ipsative and therefore to allow for comparisons between individuals. Although other non-ipsative scoring methods based on IRT techniques have been proposed (for an overview, see Brown, 2016), T-IRT models are the most comprehensive and best-established approach for dominance items. They allow test constructors to model choices between two or more items measuring multiple dimensions.

However, as already mentioned by Brown and Maydeu-Olivares (2011) in their initial publication, the goodness of T-IRT person parameter estimates depends on several test characteristics—for instance, whether items are all keyed in the same direction. It is still unknown if the model actually produces more valid parameter estimation in situations it is designed to be used for—for instance, in personnel selection settings or other contexts where individuals are motivated to endorse every item. We will argue that this may be rooted in the fact that for applied contexts, tests can hardly be constructed meeting the tests characteristics recommended by Brown and Maydeu-Olivares (2011). Therefore, it has to be established how severe the estimation problems are when taking into consideration practical limitations in test construction, such as equally keyed items, a limited number of traits, and a limited number of items. Our article seeks to address this issue and contribute to the

literature in four different ways. After a brief summary of the model, we (a) review the literature on the model's properties and validity, (b) discuss minimal requirements to construct faking resistant FC tests, (c) introduce a software package implementing the method in R, and then (d) use simulation studies to investigate potential estimation problems as well as biases in depth. Last, we discuss implications for the perspectives of the model.

The Thurstonian IRT Model

As the T-IRT model has been thoroughly described (Brown & Maydeu-Olivares, 2011, 2012), we do not attempt to reiterate every detail here; instead, we focus on the most important points to understand the model. In FC tests, items are presented in blocks of two or more items, and test takers can receive different instructions for how to respond to the items. For example, participants can be asked to rank order all items, choose the most and least preferred item, or select only one item. In the following, we will focus on rankings, as all other formats represent partial rankings and follow the same rationale (for a detailed introduction, see Brown & Maydeu-Olivares, 2011). Ranking tasks can be thought of as if participants are comparing each stimulus (i.e., each item) with every other item in the block in separate judgments (Maydeu-Olivares & Böckenholt, 2005; Thurstone, 1931). This procedure yields to $\tilde{n} = n(n-1)/2$ comparisons of two items in a block of n items. For example, for a block of the three items $\{A, B, C\}$, the three comparisons $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$ can be derived. Responses to these paired comparisons (and therefore also to rankings) can be coded in a binary pattern. Let y_{ik} be the observed response of a certain respondent for whether he or she prefers item i over item k in a given block. If item i is preferred, we set $y_{ik} = 1$. Else, if item k is preferred, we set $y_{ik} = 0$.

To model these responses, T-IRT draws on the *Law of Comparative Judgment* (Thurstone, 1927). Under this law, every item i elicits a utility, which we will call t_i. The utility is a latent variable that can be thought of as the psychological value or desiredness of an item (Maydeu-Olivares & Böckenholt, 2005). Item i is preferred over item k if and only if the utility of item i is greater than or equal to the utility of item k, that is if $t_i \geq t_k$. Each latent utility t_i is assumed to be a linear function of the latent traits η. Throughout, we assume that each item loads only on one trait, which is the usual assumption when fitting the model in practice. Thus, t_i will only load on one latent trait η_a. That is, we model the utilities as

$$t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i \tag{1}$$

with μ_i being the mean of the latent utility t_i, λ_i being the factor loading of item i on the latent trait η_a, and ε_i being an independently normally distributed error with variance ψ_i^2. Furthermore, we assume the latent traits to be multivariate normally distributed with mean zero and correlation matrix Φ. For identifiability, the variances of the latent traits are fixed to one. These assumptions result in a structural equation model exemplified in figure 1 of Brown and Maydeu-Olivares (2011).

Since the latent utilities t_i are not of interest themselves in an IRT context, we can marginalize over them and directly model the difference of the utilities $y^*_{ik} = t_i - t_k$ with $y_{ik} = 1$ if and only if $y^*_{ik} \geq 0$. Following Equation (1), we can write

$$y^*_{ik} = \mu_i + \lambda_i \eta_a + \varepsilon_i - \mu_k - \lambda_k \eta_b - \varepsilon_k. \tag{2}$$

The resulting structural equation model is exemplified in figure 2 of Brown and Maydeu-Olivares (2011). The intercept of the above model can naturally be defined as $\gamma_{ik} = \mu_i - \mu_k$. However, following Brown, we do not impose this restriction and just let γ_{ik} be free to vary for all item comparisons ik. Also, for consistency with Brown, we use $-\gamma_{ik}$ as the intercept instead of γ_{ik}. Since both η and ε are assumed to be normally distributed, the item-characteristic function can be written as

$$P(y_{ik} = 1 | \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_{ik} + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \tag{3}$$

with Φ being the standard normal distribution function in the above equation. When using Equation (3) as the pointwise likelihood, the data are not conditionally independent given the parameters on the right-hand side. More precisely, if y_{ik} and y_{ij} are from the same block and contain the same item i, then the residual correlation of y_{ik} and y_{ij} given the above model is $\mathrm{cor}(y_{ik}, y_{ij}) = \pm\psi_i^2$. We may extend the pointwise likelihood to directly include this residual correlation, so that the data are conditionally independent given the parameters:

$$P(y_{ik} = 1 | \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_{ik} + \lambda_i \eta_a - \lambda_k \eta_b + \nu_i - \nu_k}{\sqrt{\psi_i^2 + \psi_k^2}}\right). \tag{4}$$

We can understand the ν parameters as random effects with $\nu_i \sim N(0, \psi_i)$ capturing the residual correlation of data related to the same item i. The latter expression of the T-IRT model is helpful when implementing it outside a structural equation modeling framework.

Previous Validation Studies

T-IRT models were developed to estimate non-ipsative person parameters based on FC questionnaires. Thus, FC questionnaires analyzed by means of T-IRT are assumed to restrict the influence of the aforementioned response biases while meeting all common psychometric requirements. The promise behind the model, therefore, is to finally provide unbiased and consequently more valid person parameter estimates in contexts where these are hard to obtain, such as in situations where respondents are motivated to give fake answers. Therefore, many test developers have adopted this technique and constructed tests that rely on T-IRT models (e.g., Anguiano-Carrasco,

MacCann, Geiger, Seybert, & Roberts, 2015; Brown & Bartram, 2013; Guenole, Brown, & Cooper, 2018; Lewis, 2015; Merk, 2016).

At the same time, several studies have investigated different aspects of the model's validity. In a think-aloud task when completing FC items, respondents indeed reported making pairwise comparisons between all items of a triplet—that is, a block of three items (Sass, Frick, Reips, & Wetzel, 2018). Thus, one of the key components of T-IRT models, the transformation of rankings into pairwise comparisons, seems to describe respondents' underlying decision-making processes reasonably well. However, this study also draws attention to a limitation of this modeling technique: Only 76% of the participants reported having no difficulty in keeping the information for all statements in mind to appraise the utility of an item relative to the utilities of all other items in the block. Furthermore, as the Sass et al. (2018) study only used triplets, this limitation likely leads to severe violations of the model assumptions when blocks with more items are employed.

A second line of studies has compared T-IRT models with results of rating scales and other FC scoring procedures. FC-based T-IRT scores correlated with results from rating scales substantially (on average .76 in Guenole et al., 2018; and .79 in P. Lee, Lee, & Stark, 2018). However, in combination with differing loading patterns between both conditions, these results suggest that the two response formats (FC vs. rating) produce different trait meanings (Guenole et al., 2018).

Compared with other scoring procedures for FC items, T-IRT scores yield equal or better convergent and discriminant validity results (Anguiano-Carrasco et al., 2015; P. Lee et al., 2018). In comparison to rating items, three studies (Anguiano-Carrasco et al., 2015; Brown & Maydeu-Olivares, 2013; P. Lee et al., 2018) reported that T-IRT scores had slightly lower validity estimates. However, the validation scales used in two of these studies (Anguiano-Carrasco et al., 2015; P. Lee et al., 2018) had a rating format and thus might have led to a common method effect with the rating version of the validated questionnaires. Moreover, none of these studies was conducted under high-stakes conditions and therefore participants had limited incentive to distort their answers; under these conditions, FC items cannot derive a benefit from their potential to combat insincere answers. In a situation with a higher motivation to distort (i.e., in a 360-degree feedback) and when the criterion itself was an FC test, T-IRT yielded better validity estimates than conventionally scored rating items (Brown, Inceoglu, & Lin, 2017). However, for two reasons these results need to be interpreted with caution. First, in an aforementioned study by Brown and Maydeu-Olivares (2013), the median correlation between IRT scored rating items and T-IRT scored FC items was .70 for the same questionnaire. In contrast, the median correlation for ipsative FC and T-IRT FC scores was .88. This suggests a considerable influence of the response format. For this reason, it is problematic to compare validity estimates of FC versus rating scores when the criterion itself is an FC score. Second, Brown et al. (2017) modeled responses to rating items not only conventionally but in a second condition with an additional bias factor, too. Rating items analyzed with this method yielded equal convergent validities compared with FC-based T-IRT scores.

In sum, there is still a lack of convincing validation studies in which FC-based T-IRT scores achieve higher validity estimates than their rating scale counterparts.

Equally Versus Unequally Keyed Items

As mentioned earlier, FC items are supposed to reduce a variety of response biases associated with Likert scales. In many high-stakes contexts such as personnel selection, however, preventing socially desirable responding is of primary interest. The rationale behind using FC items for this purpose is to force respondents to choose from several equally attractive options (McCloy, Heggestad, & Reeve, 2005). In this way, it is—in contrast to rating items—impossible for participants to fully endorse every attractive item. Thus, the mere fit between the person and the item is assumed to elicit the choice of a specific item.

Evidently, for FC techniques to work as they are supposed to, all items within a block need to be equally desirable. For this reason, Brown and colleagues (2017, p. 142) clearly state, ''Careful matching on desirability of behaviors within each block is strongly recommended to minimize nonuniform response distortions.'' Thus, FC tests will only resist distortion attempts if they are constructed reasonably, that is, if respondents cannot detect a ''best'' answer.

Typically, attractiveness ratings from pretests are used to obtain blocks of equally desirable items (Christiansen et al., 2005). Many studies on FC tests do not report whether test items are matched or obviously do not match them. Where matching is taken into account, the matching process is done only once, and from that point on, blocks remain unchanged. However, the desirability of items depends on the context (Waters, 1965). In application processes, the statement ''I am very talkative'' might be more desirable for sales positions than for auditors. Similarly, students might perceive certain attributes as attractive, whereas older applicants with work experience may perceive other items as more favorable. Therefore, the combination of items in FC tests should be based on desirability ratings (a) for the situation in which they are to be administered and (b) by raters from the same populations as the test takers. Note that a single FC aptitude test for all occupations will hardly be fake resistant in all contexts of application. Moreover, other items within the same block can shift the desirability of a specific item (Feldman & Corah, 1960) and influence the item parameters in the T-IRT model (Lin & Brown, 2017). According to Lin and Brown (2017), only a small proportion of items is affected by other items of the block (in the sense that they change the item parameters) and the authors suggest strategies to reduce the occurrence of these changes. However, to clearly assure equal desirability of all items within a block, one would have to measure the desirability of an item in the context of a given block and then modify the combination of items accordingly. As this is an iterative process, it might require several waves of data collection with the specific population for which the test is constructed.

Besides the large amount of effort needed to conduct a pretest, there is another reason why test developers refrain from desirability matching. Namely, first

simulation studies indicate that latent trait recovery is far more accurate when a questionnaire consists of equally and unequally keyed blocks (Brown & Maydeu-Olivares, 2011). For tests with only equally keyed blocks (i.e., blocks that contain only positively keyed or only negatively keyed items), simulation results prompt the question of whether this condition can even allow for the construction of sufficiently reliable questionnaires. However, the use of unequally keyed blocks (where positively and negatively keyed items are present within the same block) in FC questionnaires is problematic for at least four reasons:

First, the use of negatively keyed items can increase the cognitive demand on test takers. One widespread technique to create negative factor loadings is to negatively phrase items—that is, to use negations. While they are known to be problematic in single-stimulus items already, the issue becomes more severe in FC questionnaires due to the increased cognitive load they provoke: Assuming that the Law of Comparative Judgment describes the response processes in FC questionnaires reasonably well, a block of n items leads to the respondent having to make $\tilde{n} = n(n-1)/2$ comparisons when answering a single block of items. Recall that only 76% of respondents reported having no difficulties in processing the three comparisons necessary for blocks with three items (Sass et al., 2018); but blocks with four items require participants to process six comparisons. The use of negatively keyed items can further increase this already high cognitive demand. When the block size and therefore the cognitive load is high, and when negatively keyed items further increase the cognitive load, it is especially questionable whether the Law of Comparative Judgement describes the response processes appropriately. Even if T-IRT models correctly describe response processes under such conditions, test results might depend on the working memory capacity of respondents. This, in turn, limits the potential construct validity such a test can reach.

A second problem of including negatively keyed items in a block is their potential to result in substantial methodological variance (Dueber, Love, Toland, & Turner, 2018). This can affect the covariance structure of the items because the negatively keyed items may effectively form a separate method factor (Lucía et al., 2014).

While the above issues only suggest that including negatively keyed items is challenging, a third and more severe issue concerns how including such items affects FC tests' ability to resist faking and therefore raises doubts about the reasonableness of the approach in general. Again, the rationale behind FC as a method to prevent fake answers is to force respondents to choose between several equally attractive items. Suppose a given test measures three traits for which high values are desirable; therefore, to form a block (e.g., a triplet) with both positively and negatively keyed items, one could include two positively and one negatively keyed item. Now, the positively keyed items both contain statements that correspond to the desirable end of the trait continuum, while the negatively keyed item, in contrast, describes the undesired end of the trait. By looking at the items, participants will be able to clearly identify that the negatively keyed item is less socially desirable than the other two. As such, this violates the rationale for how FC tests are resistant to fake answers. If the quality of

T-IRT person parameter estimates critically depends on responses to these unequally keyed blocks (see Brown & Maydeu-Olivares, 2011), it is implausible to believe that such tests can prevent socially desirable responding. In sum, we argue that items within a block cannot be matched for desirability (as correctly requested by Brown et al., 2017) and at the same time unequally keyed (as suggested based on the simulations by Brown & Maydeu-Olivares, 2011).

This being said, the use of negatively keyed items leads to a closely related fourth issue: If participants can identify one item as optimal, almost all will choose this item in a high-stakes situation. Consequently, the respective block will contribute little information to the parameter estimation (Wang, Qiu, Chen, Ro, & Jin, 2017). Thus, the precision that has been suggested by previous simulations (Brown & Maydeu-Olivares, 2011) will not be obtained in many practical implementations due to the restricted validity of the simulations' generative model. In case of unequally keyed items within one block, the social desirability of the items will influence test takers' choices much more strongly than is accounted for in the simulations. Therefore, questionnaires including unequally keyed items will not only be fakable but likely yield considerably lower measurement precision than simulations for this condition suggest.

From a practical perspective, the model's applicability therefore largely depends on its ability to estimate persons' trait scores correctly based on equally keyed items only. This property has only been investigated by a single simulation study so far (see Brown & Maydeu-Olivares, 2011). Understandably, this initial investigation was not able to cover all conditions relevant for test construction.

One aspect we want to deepen our understanding of is the recoverability of the true correlations between traits. In empirical applications of the model, results showed that the covariance structures of the estimated traits are a critical issue sometime leading to unexpected results (Morillo et al., 2016). In other cases, the estimation needed to be facilitated by manual interventions (Anguiano-Carrasco et al., 2015; Brown et al., 2017; Guenole et al., 2018).

Another limitation of the initial investigation by Brown and Maydeu-Olivares (2011) was that their model evaluation only involved the comparison of reliability estimates. These estimates are based on the simplifying assumption of constant measurement error across individuals and linearity of the relationship between true and estimated trait scores. One may argue that lowered reliabilities are acceptable if trait scores are unbiased in exchange, but the unbiasedness still needs to be investigated. Moreover, existing implementations estimate T-IRT models solely by means of frequentist statistics even though Bayesian statistics offers some key advantages—for example, the ability to incorporate prior information and to estimate the full joint posterior distribution of all model parameters (Gelman et al., 2013). Among others, the latter allows to obtain the individual-specific distributions of trait scores without having to make any simplifying assumptions.

In this context, our subsequently reported simulations include more differentiated conditions than previous simulation studies. Also, they provide several different

criteria of model accuracy such as the root mean squared error (RMSE), estimated inter-trait correlations, and nonlinear relationships between true and estimated trait scores. Together, these procedures seek to gain further insights into the statistical properties of the estimates obtained from T-IRT models.

Implementation in R

For the purpose of fitting T-IRT models in R (R Core Team, 2018), we have developed a new package called *thurstonianIRT* (Bürkner, 2018), which is available on GitHub (https://github.com/paul-buerkner/thurstonianIRT). It also contains functions for convenient data preparation as well as functions to simulate data. Within the package, one can choose between lavaan (Rosseel, 2012), Mplus (Muthén & Muthén, 2015), and Stan (Carpenter et al., 2017) as underlying engines for model fitting. Both Mplus and lavaan implement (mostly frequentist) structural equation modeling. The thurstonianIRT package autogenerates Mplus and lavaan code based on information provided by the user. We extensively validated this code against the Mplus code provided by Brown and Maydeu-Olivares (2012).

The third engine, Stan, is fundamentally different from the other two. Stan is a probabilistic programming language that fits Bayesian models using Markov chain Monte Carlo sampling (for more details, see Carpenter et al., 2017; Hoffman & Gelman, 2014). As compared with frequentist methods, fitting models in a Bayesian framework provides several advantages. First, we can estimate the complete (joint) posterior distribution of the parameters. This is not only fully consistent with probability theory but also much more informative than a single point estimate and an approximate measure of uncertainty commonly known as ''standard error'' in frequentist statistics. Second, we usually see less convergence issues, which is in part because we can specify hard boundaries on naturally bounded parameters such as variances or correlations. Third, Bayesian models allow to explicitly incorporate prior information into the model by means of specifying prior distribution on the parameters (e.g., see Gelman et al., 2013). Our Stan implementation of T-IRT models uses weakly informative (i.e., rather wide) default priors that improve sampling efficiency and convergence without considerable influencing parameter estimates.

To obtain individual trait scores, the thurstonianIRT package uses maximum a posteriori (MAP) estimation in Mplus and lavaan and expected a posteriori (EAP) estimation in Stan. As the posterior distribution of the trait scores can be expected to be highly symmetric and unimodal, MAP and EAP estimators are likely to yield similar results.

Simulations

The goal of the present simulations is to extend the results of Brown and Maydeu-Olivares (2011) on the properties of T-IRT models. To us, it is of primary interest to evaluate the accuracy of subjects' estimated trait scores, because those are the

parameters that really matter in practice. In Brown and Maydeu-Olivares (2011), only the reliability of estimated trait scores was investigated, which is their squared product–moment correlation with the corresponding true scores:

$$\mathrm{Rel}\left(\hat{\theta}, \theta\right) = \mathrm{Cor}(\hat{\theta}, \theta)^2. \tag{5}$$

For reasons explained above, we believe that only investigating the reliability of the trait scores is unsatisfactory. Therefore, we will also investigate the RMSE, which is a common measure of absolute fit defined as

$$\mathrm{RMSE}\left(\hat{\theta}, \theta\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)^2}. \tag{6}$$

In the above equation, n is the number of persons over which to compute the RMSE. Because the scale of traits is not essential, we will be computing the RSME for standardized trait scores (i.e., z scores). In addition to the trait scores themselves, their intercorrelations and individual precision estimates will be investigated as well.

Below we describe the simulation design used in the simulations described in the following two subsections. For each simulation trial, data of $n = 2000$ subjects were simulated and only blocks of three items (i.e., triplets) were considered. The number of traits n_T varied between $n_T = 3$ and $n_T = 5$, and the number of blocks in which items of a given trait were administered (n_{BT}; read *blocks per trait*) varied between $n_{BT} = 9$ and $n_{BT} = 27$ in steps of 6. Within a triplet, each item measured a different trait. The true correlation matrix Φ between the traits took on five different values: (a) all correlations set to 0 (i.e., independent traits), (b) all correlations set to 0.3, (c) all correlations set to -0.3,[1] as well as two real-world correlation matrices (d) from the NEO-PIR (McCrae & Costa, 1992), and (e) from one of our own projects[2] (see Table 1). When simulating scores for fewer than five traits, the first three or four rows and columns of these matrices were selected.

The true values of the standardized factor loadings λ took on (a) random values sampled uniformly between 0.3 and 0.7, (b) random values sampled uniformly between 0.65 and 0.95 (both conditions are referred to as positive factor loadings or equally keyed triplets), and (c) same as (b) but with half of the values having the opposite sign (referred to as mixed factor loadings or unequally keyed triplets). The last two λ conditions correspond to the approach used by Brown and Maydeu-Olivares (2011), while the first condition aims to test the robustness of the model against smaller factor loadings. As we simulated standardized factor loadings, the error variances ψ were computed as $\psi = 1 - \lambda^2$. For all simulation conditions, γ values were sampled uniformly between -1 and 1. We used a fully crossed design, which results in 180 unique simulation conditions to be evaluated.

Our simulations consist of three parts. In the first part, only Mplus is used for the model fitting and $N = 100$ simulation trials per condition are evaluated. In the second part, Mplus, lavaan, and Stan are used for a single trial per simulation condition so that the whole simulation can finish in reasonable time (i.e., roughly half a year on

Table 1. Two Real-World Correlation Matrices of Big Five Personality Scores.

	NEO-PIR				Own Project			
	N	E	C	A	N	E	C	A
E	−0.21				−0.33			
C	−0.53	0.27			−0.43	0.30		
A	−0.25	0.00	0.24		−0.37	0.32	0.27	
O	0.00	0.40	0.00	0.00	−0.04	0.16	0.05	0.13

Note. N = neuroticism; E = extraversion; C = conscientiousness; A = agreeableness; O = openness to experiences.

our simulation machines). The third part does not follow the above described simulation conditions completely, as we simulate data of 30 traits (instead of 3 to 5 traits) under conditions similar to those found in the Occupational Personality Questionnaire (OPQ) test (Brown & Bartram, 2011). All simulations were done with the thurstonianIRT package described above. The complete R code as well as all simulation results can be found on OSF (Open Science Framework) (https://osf.io/df5yq/).

Results of Mplus Only Simulations

For reasons of readability, numeric results are reported only for selected simulation conditions that are representative of the complete simulations. As summarized in Table 2, and in line with the results of Brown and Maydeu-Olivares (2011), tests including unequally keyed triplets yield much higher reliability for predicted trait scores than tests with only equally keyed triplets. The RMSE points in the same direction. For other simulation conditions held constant, tests with equally keyed triplets had roughly twice the RMSE than tests with unequally keyed triplets. Moreover, in terms of absolute fit, we perceive the RMSE values found for equally keyed triplets to be worryingly high (RMSE > 0.5 standard deviation units in most cases). Increasing the number of traits from 3 to 5 as well as the number of triplets per trait from 9 to 27 leads to improvements in the univariate accuracy measures. However, the accuracy of trait scores based on only equally keyed triplets remains unsatisfactory throughout the simulation conditions.

With regard to variation of the inter-trait correlation matrix Φ, the accuracy of trait scores again depends on the type of factor loadings. For unequally keyed triplets, Φ seems to have little to no influence both on reliability and RMSE. However, for equally keyed triplets, the influence of Φ is not negligible. In particular, the accuracy is reduced considerably if the true inter-trait correlations are set to 0.3 (see Table 2).

Looking at the accuracy in the estimation of Φ, again the most important aspect is the type of factor loadings. While inter-trait correlations are unbiased in conditions

Table 2. Selected Simulation Results of T-IRT Models Fitted With Mplus.

Conditions			Traits = 3				Traits = 5			
BpT	λ	Φ	Conv	Rel	RMSE	Φ-Bias	Conv	Rel	RMSE	Φ-Bias
9	$+^*$	0	0.97	0.54	0.73	−0.44	0.94	0.60	0.67	−0.22
9	$+^*$	0.3	1.00	0.39	0.87	−0.66	1.00	0.43	0.83	−0.42
9	$+^*$	NEO	0.87	0.64	0.63	−0.27	0.90	0.64	0.63	−0.16
9	+	0	0.84	0.59	0.68	−0.44	0.93	0.73	0.54	−0.22
9	+	0.3	0.90	0.44	0.82	−0.70	0.97	0.56	0.71	−0.44
9	+	NEO	0.70	0.69	0.58	−0.26	0.83	0.77	0.49	−0.15
9	+/−	0	0.98	0.88	0.35	0.01	0.98	0.90	0.32	0.00
9	+/−	0.3	0.97	0.88	0.35	0.00	0.99	0.90	0.32	0.00
9	+/−	NEO	0.97	0.88	0.35	0.01	0.95	0.90	0.32	0.00
27	$+^*$	0	1.00	0.67	0.60	−0.40	1.00	0.75	0.51	−0.20
27	$+^*$	0.3	1.00	0.56	0.71	−0.52	1.00	0.64	0.64	−0.32
27	$+^*$	NEO	1.00	0.76	0.50	−0.25	1.00	0.79	0.47	−0.15
27	+	0	0.94	0.69	0.58	−0.38	0.94	0.82	0.44	−0.19
27	+	0.3	0.98	0.60	0.67	−0.50	0.97	0.73	0.54	−0.28
27	+	NEO	0.96	0.77	0.49	−0.25	0.93	0.85	0.40	−0.14
27	+/−	0	0.61	0.95	0.22	0.00	0.30	0.96	0.19	0.00
27	+/−	0.3	0.56	0.95	0.23	−0.02	0.28	0.96	0.20	0.00
27	+/−	NEO	0.69	0.95	0.23	0.01	0.38	0.96	0.20	0.00

Note. Results were computed based on 100 simulation trials per condition. BpT = blocks per trait; Conv = convergence rate; Rel = reliability; RMSE = root mean squared error of z scores; Φ-Bias = bias in the correlations of estimated trait scores; $\lambda(+^*)$ = positive factor loadings in [0.3, 0.7]; $\lambda(+)$ = positive factor loadings in [0.65, 0.95]; $\lambda(+/-)$ = mixed factor loadings; $\Phi(0)$ = independent traits; $\Phi(0.3)$ = traits correlated by 0.3; Φ(NEO) = correlations taken from the NEO-PIR.

with unequally keyed triplets, correlations show a strong negative bias in conditions with only equally keyed triplets. This bias persists even for larger tests with 5 traits and 27 triplets per trait, although the strength of bias varies across conditions of Φ (see Table 2).

Using positive factor loadings of reduced size (i.e., $\lambda \in [0.3, 0.7]$) yielded slightly worse reliability and RMSE than when using $\lambda \in [0.65, 0.95]$, but the overall pattern described above remains the same. Furthermore, using Φ from our own project yielded similar results than Φ from the NEO-PIR. For reasons of brevity, these results are not shown in detail but can be found on OSF (https://osf.io/df5yq/).

To evaluate the uncertainty in our results due to simulation error, we computed the standard deviations of the quantities of interest (i.e., reliability, RMSE, and inter-trait correlation bias) across simulation trials within each condition. As can be seen in Table 3, reliability and RMSE were highly consistent in all conditions. The inter-trait correlation bias showed substantial variation across trials for small number of blocks, but it also became more stable for larger number of blocks. Together, these results indicate that $N = 100$ simulation trials yielded sufficient precision in the

Table 3. Standard Deviations of Selected Simulation Results of T-IRT Models Fitted With Mplus.

Conditions			Traits = 3			Traits = 5		
BpT	λ	Φ	*SD* (Rel)	*SD* (RMSE)	*SD* (Φ-Bias)	*SD* (Rel)	*SD* (RMSE)	*SD* (Φ-Bias)
9	$+^*$	0	0.03	0.03	0.17	0.03	0.03	0.10
9	$+^*$	0.3	0.05	0.04	0.21	0.05	0.05	0.13
9	$+^*$	NEO	0.04	0.04	0.20	0.03	0.03	0.09
9	+	0	0.04	0.03	0.24	0.02	0.02	0.10
9	+	0.3	0.03	0.03	0.24	0.03	0.02	0.11
9	+	NEO	0.04	0.04	0.22	0.01	0.02	0.09
9	+/−	0	0.01	0.02	0.03	0.01	0.01	0.02
9	+/−	0.3	0.01	0.02	0.03	0.01	0.01	0.02
9	+/−	NEO	0.01	0.01	0.03	0.01	0.01	0.02
27	$+^*$	0	0.02	0.02	0.08	0.02	0.02	0.05
27	$+^*$	0.3	0.02	0.02	0.08	0.02	0.02	0.05
27	$+^*$	NEO	0.01	0.02	0.09	0.01	0.02	0.05
27	+	0	0.02	0.02	0.09	0.01	0.01	0.05
27	+	0.3	0.02	0.02	0.09	0.01	0.01	0.04
27	+	NEO	0.02	0.02	0.12	0.01	0.01	0.04
27	+/−	0	0.00	0.01	0.01	0.00	0.00	0.01
27	+/−	0.3	0.00	0.01	0.01	0.00	0.00	0.01
27	+/−	NEO	0.00	0.01	0.01	0.00	0.00	0.01

Note. Standard deviations were computed across on 100 simulation trials per condition. T-IRT = Thurstonian item response theory; BpT = blocks per trait; Conv = convergence rate; Rel = reliability; RMSE = root mean squared error of *z* scores; Φ-Bias = bias in the correlations of estimated trait scores; $\lambda(+^*)$ = positive factor loadings in [0.3, 0.7]; $\lambda(+)$ = positive factor loadings in [0.65, 0.95]; $\lambda(+/-)$ = mixed factor loadings; $\Phi(0)$ = independent traits; $\Phi(0.3)$ = traits correlated by 0.3; Φ(NEO) = correlations taken from the NEO-PIR.

average quantities shown in Table 2. That is, the variation across conditions was much larger than the uncertainty due to simulation error.

Comparison Between Model Implementations

In this section, we compare the results of T-IRT models fitted with Mplus, lavaan, and Stan based on a single simulation trial per condition. If no convergence was reached, the trial was repeated. Again, for reasons of readability, numeric results are reported for selected simulation conditions that are representative of the complete simulations (see Table 4). For unequally keyed triplets, trait estimates computed by Mplus and Stan show very similar reliability and RMSE, while the accuracy of estimates obtained by lavaan are slightly worse. For equally keyed triplets, none of the three implementations performs uniformly better than the others. In particular, the problems seen in the Mplus-only simulation above can also be found when fitting T-

Table 4. Selected Simulation Results of T-IRT Models Fitted With Mplus, lavaan, or Stan.

Conditions		Mplus			lavaan			Stan		
λ	Φ	Rel	RMSE	Φ-Bias	Rel	RMSE	Φ-Bias	Rel	RMSE	Φ-Bias
$+^*$	0	0.67	0.61	−0.39	0.64	0.64	−0.49	0.66	0.61	−0.29
$+^*$	0.3	0.55	0.72	−0.62	0.50	0.77	−0.74	0.58	0.69	−0.38
$+^*$	NEO	0.76	0.50	−0.26	0.75	0.52	−0.28	0.76	0.50	−0.20
+	0	0.70	0.58	−0.39	0.62	0.65	−0.48	0.67	0.61	−0.48
+	0.3	0.57	0.70	−0.49	0.47	0.79	−0.75	0.44	0.82	−0.40
+	NEO	0.78	0.48	−0.21	0.73	0.54	−0.24	0.76	0.50	−0.29
+/−	0	0.95	0.22	−0.01	0.93	0.26	−0.01	0.95	0.22	0.00
+/−	0.3	0.95	0.23	−0.02	0.92	0.28	0.12	0.95	0.22	−0.01
+/−	NEO	0.95	0.23	0.01	0.92	0.28	−0.02	0.95	0.22	0.01

Note. Results are averaged across 3 traits each measured in 27 triplets. T-IRT = Thurstonian item response theory; Rel = reliability; RMSE = root mean squared error of *z* scores; Φ-Bias = Bias in the correlations of estimated trait scores; $\lambda(+^*)$ = positive factor loadings in [0.3, 0.7]; $\lambda(+)$ = positive factor loadings in [0.65, 0.95]; $\lambda(+/-)$ = mixed factor loadings; $\Phi(0)$ = independent traits; $\Phi(0.3)$ = traits correlated by 0.3; Φ(NEO) = correlations taken from the NEO-PIR.

IRT models with lavaan or Stan. Apparently, even a fully Bayesian implementation—at least one with weakly informative priors—cannot considerably improve accuracy in tests with only equally keyed triplets.

As exemplified in Figure 1, all three implementations show a similar nonlinearity in the relationship between true and estimated trait value. More specifically small trait scores are overestimated and large trait scores are underestimated. These tendencies seem to be somewhat stronger for lavaan as compared with Mplus and Stan, which may explain the slightly inferior performance of lavaan with regard to reliability and RMSE. Comparing the left- and right-hand side of Figure 1, we see that the nonlinearity is even stronger when using equally keyed triplets as compared with using unequally keyed triplets. Moreover, the uncertainty in the estimated trait scores seems to be considerably larger in the case of equally keyed triplets as indicated by the wider 95% confidence intervals around the regression curve.

So far, comparisons have focused only on univariate measures of accuracy for each trait. However, as we have seen in the above section, more severe problems are uncovered if we investigate the traits simultaneously by looking at their correlations. In Tables 5 and 6, we see example simulation results for three independent traits estimated based on 27 triplets. If unequally keyed triplets are used, estimated trait correlations are very close to the true correlation, which is approximately zero for the shown condition. In contrast, if only equally keyed triplets are used, estimated trait correlations are way off in all three packages showing a strong negative bias (see Table 5). None of the software packages seems to perform notably better than the others with regard to trait correlations independent of the number of estimated traits, the number of triplets per trait, or the true trait correlations (see OSF for results of all

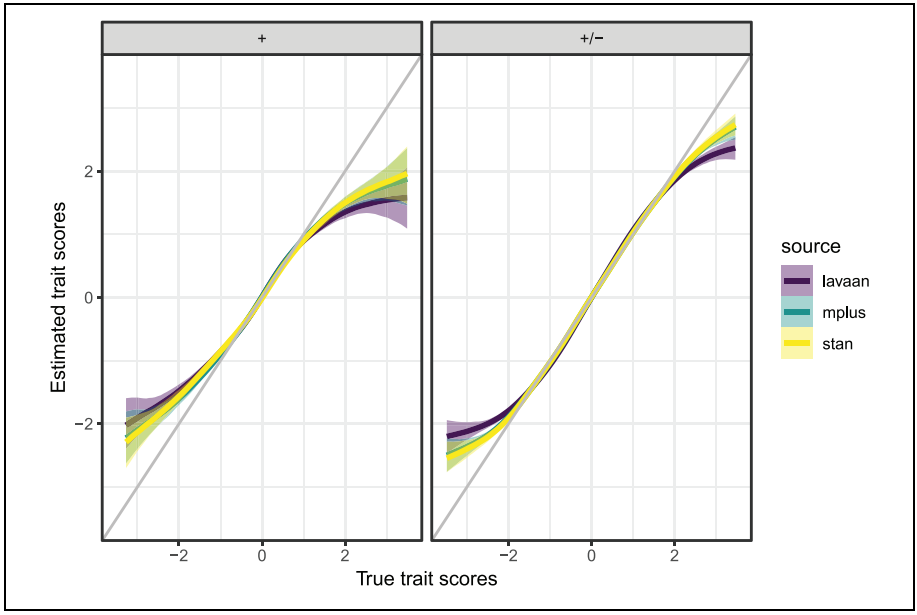


Figure 1. Relationship between true and estimated trait scores for one of three independent traits using 27 blocks per trait. Plots on the left- and right-hand sides show results for equally and unequally keyed triplets, respectively. Shaded areas are 95% confidence intervals. Regression curves were estimated by thin-plate splines.

conditions). Still, the packages closely agree with each other in the sense that trait estimates of the same traits obtained with different packages highly correlate under all conditions. Furthermore, the agreement of estimates from different packages is higher than the agreement of these estimates with the true values (see Table 6).

Last, we turn our attention toward the uncertainty with which persons' trait scores are estimated. Under the assumption of (approximate) normality, standard errors provide a sufficient statistic for the uncertainty of a quantity of interest, in this case the individual trait scores. For T-IRT models, such standard errors can be obtained in both frequentist and Bayesian frameworks regardless of the particular implementation (Dueber et al., 2018; Gelman et al., 2013; Maydeu-Olivares & Brown, 2010). As shown in Figure 2, standard errors of trait scores vary considerably. We see that standard errors of more extreme (very low or high) scores are higher on average than those of medium scores. The minimal standard errors are only about one fourth of the maximal ones. Moreover, even very similar trait scores may have considerably different standard errors. This demonstrates that summary statistics, such as the reliability or the RMSE, are not necessarily appropriate for each individual trait score, but rather describe an average accuracy.

Table 5. Correlations of Trait Scores Coming From the Same Source.

		Equally keyed triplets			Unequally keyed triplets		
Source	Trait	Trait1	Trait2	Trait3	Trait1	Trait2	Trait3
Truth	Trait1	1.00			1.00		
	Trait2	0.00	1.00		0.00	1.00	
	Trait3	0.00	0.00	1.00	0.00	0.00	1.00
Mplus	Trait1	1.00			1.00		
	Trait2	−0.26	1.00		0.01	1.00	
	Trait3	−0.47	−0.48	1.00	−0.05	−0.05	1.00
lavaan	Trait1	1.00			1.00		
	Trait2	−0.37	1.00		0.02	1.00	
	Trait3	−0.59	−0.52	1.00	−0.09	−0.03	1.00
Stan	Trait1	1.00			1.00		
	Trait2	−0.42	1.00		0.02	1.00	
	Trait3	−0.53	−0.51	1.00	−0.05	−0.05	1.00

Note. Estimated inter-trait correlations are based on three independent traits using 27 blocks per trait.

Table 6. Correlations of Trait Scores Coming From Different Sources.

		Equally keyed triplets				Unequally keyed triplets			
Trait	Source	Truth	Mplus	lavaan	Stan	Truth	Mplus	lavaan	Stan
Trait1	Truth	1.00				1.00			
	Mplus	0.83	1.00			0.98	1.00		
	lavaan	0.78	0.97	1.00		0.97	0.99	1.00	
	Stan	0.81	0.99	0.98	1.00	0.98	1.00	0.99	1.00
Trait2	Truth	1.00				1.00			
	Mplus	0.83	1.00			0.98	1.00		
	lavaan	0.78	0.96	1.00		0.97	0.99	1.00	
	Stan	0.81	0.99	0.98	1.00	0.98	1.00	0.99	1.00
Trait3	Truth	1.00				1.00			
	Mplus	0.85	1.00			0.98	1.00		
	lavaan	0.81	0.96	1.00		0.96	0.99	1.00	
	Stan	0.83	0.98	0.98	1.00	0.98	1.00	0.99	1.00

Note. Estimated inter-trait correlations are based on three independent traits using 27 blocks per trait.

Comparing equally keyed to unequally keyed triplets (left and right, respectively, in Figure 2), we see that, on average, unequally keyed triplets lead to more certain estimation of the factor scores, which is in line with the true accuracies being higher for unequally keyed triplets. Moreover, the distributions of factor scores for different traits seem to be slightly shifted when only equally keyed triplets are used. Presumably, this is because the model is not able to estimate the absolute position of factor scores with sufficient accuracy in that condition, as already discussed above.

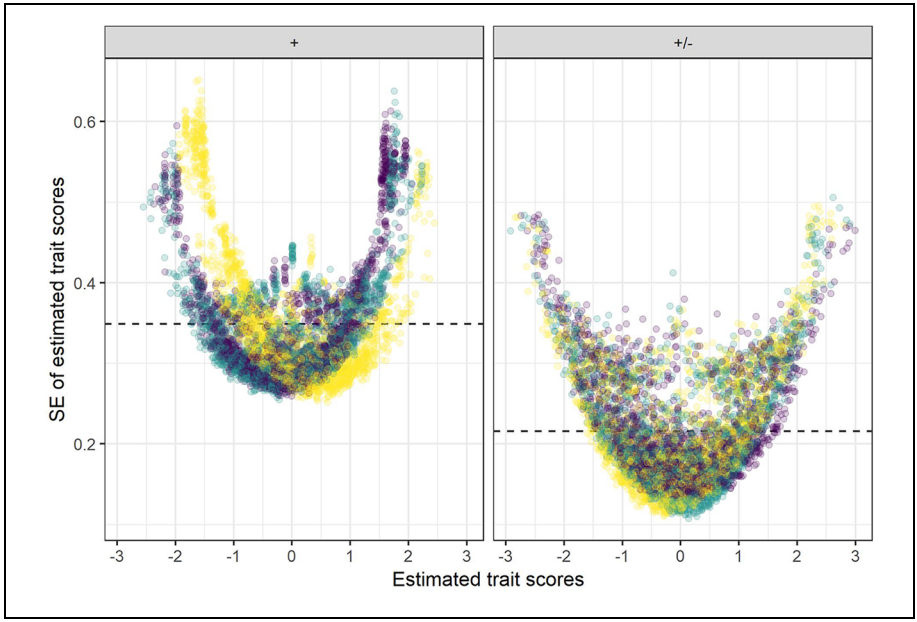


Figure 2. Relationship between estimated trait scores and their standard errors computed with Stan based on three independent traits and 27 triplets. Plots on the left- and right-hand sides show results for equally (+) and unequally (+/−) keyed triplets, respectively. Each point corresponds to a particular person's trait score and corresponding standard error. Horizontal dashed lines indicate mean standard errors. Colors indicate different traits.

Simulations of 30 Traits

Several reports in the literature claim that the above-described problems with tests consisting of only equally keyed items will vanish if a large number of traits are measured (e.g., Bartram, 1996; Saville & Willson, 1991). Usually, measuring 30 traits is said to be sufficient. Yet we have not found a simulation study where T-IRT models have been applied to that many traits. To fill this gap, we focus on six particularly relevant simulation conditions. With regard to the factor loadings λ, we sample uniformly between 0.65 and 0.95 for the equally keyed triplets and randomly change the sign of half of them for the unequally keyed condition. Off-diagonal entries of the inter-trait correlation matrix Φ are either (a) all set to 0, (b) all set to 0.3, or (c) taken from the OPQ CM 4.2 (Saville, Holdsworth, Nyfield, Cramp, & Maybey, 1992), which represents a real correlation matrix of 30 traits with correlations ranging from -0.41 to 0.50.[3] Each of the 30 traits was measured in 9 triplets and compared randomly with other traits, resulting in a total of 90 triplets. This structure is similar to the latest OPQ tests (Brown & Bartram, 2011; Joubert, Inceoglu,

Table 7. Simulation Results of T-IRT Models for 30 Traits Fitted With Stan.

Conditions		Factor scores		Bias of inter-trait correlations			
λ	Φ	Rel	RMSE	*M*	*SD*	Minimum	Maximum
+	0	0.87	0.36	−0.03	0.02	−0.09	0.02
+	0.3	0.70	0.57	−0.30	0.03	−0.39	−0.22
+	OPQ	0.91	0.30	0.00	0.02	−0.08	0.07
+/−	0	0.90	0.32	0.00	0.02	−0.05	0.05
+/−	0.3	0.91	0.31	0.03	0.02	−0.03	0.08
+/−	OPQ	0.92	0.29	0.00	0.02	−0.05	0.07

Note. Results are averaged across 30 traits each measured in 10 triplets. T-IRT = Thurstonian item response theory; Rel = reliability; RMSE = root mean squared error; *M* = mean; *SD* = standard deviation; λ(+) = positive factor loadings; λ(+/−) = mixed factor loadings. Φ(0) = independent traits; Φ(0.3) = traits correlated by 0.3; Φ(OPQ) = traits correlated according to the OPQ CM 4.2.

Bartram, Dowdeswell, & Lin, 2015; Lin & Brown, 2017). Responses of 1,000 participants were simulated.

T-IRT models were fit only in Stan with a single simulation trial per condition. When trying to fit those models with lavaan or Mplus, we ran into serious memory issues and thus were not able to obtain any estimates with those packages. Both programs required more than 32 GB of working memory, which none of our machines could offer when these simulations were run. On personal communication with developers of lavaan, we think that this is because the Hessian matrix of the model parameters—or matrices built on top of that during the optimization procedure—becomes too large. Apparently, to fit T-IRT models of that size with structural equation modeling software, a cluster solution with excess RAM is required.

The Stan results were as follows. As summarized in Table 7, equally keyed triplets seem to perform similarly to unequally keyed triplets with respect to both reliability and RMSE under the above-described conditions—provided that the traits are either uncorrelated or have a mixed correlation pattern as found in the OPQ. As indicated in the last four columns of Table 7, these correlations also seem to be estimated very precisely. However, if the true correlations are 0.3 and only equally keyed triplets are administered, the reliability and RMSE fall off notably. Arguably, though, the assumption of 30 pairwise positively correlated traits is quite unrealistic and thus hardly of practical relevance. Together, the results indicate that inference based on tests with only equally keyed items can be improved considerably by measuring a large number of traits. However, standard errors of estimated trait scores are still rather high in particular for persons with more extreme trait scores as displayed in Figure 3.

Discussion

In the present article, we investigated the T-IRT model from both a practical and a statistical perspective and found that these perspectives are difficult to bring into

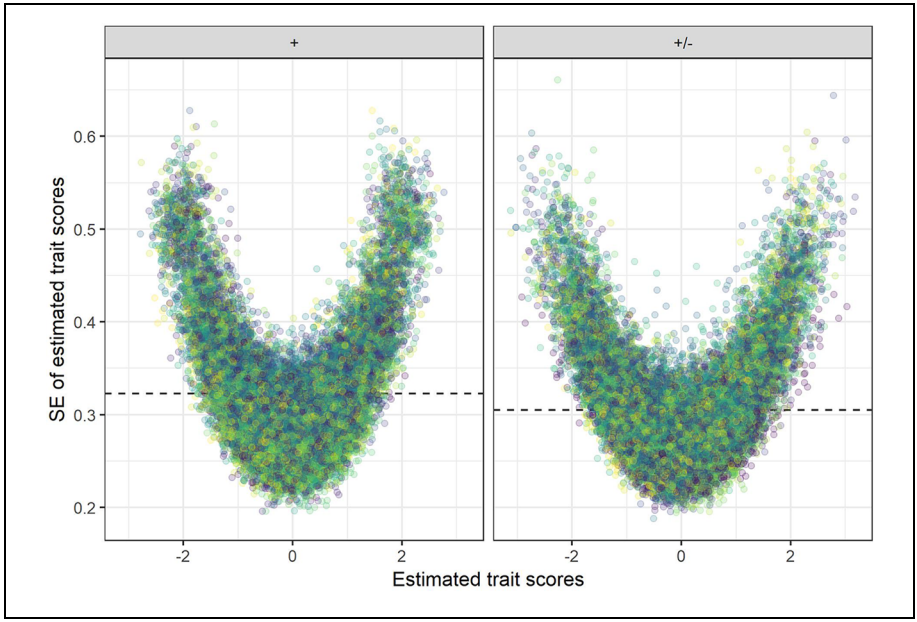


Figure 3. Relationship between estimated trait scores and their standard errors computed with Stan based on 30 traits correlated according to the OPQ CM 4.2 and each measured in 9 triplets. Plots on the left- and right-hand sides show results for equally (+) and unequally (+/−) keyed triplets, respectively. Each point corresponds to a particular person's trait score and corresponding standard error. Horizontal dashed lines indicate mean standard errors. Colors indicate different traits.

alignment. According to our simulations and practical considerations, severe problems arise with using the model in settings that are appropriate for implementing in practice. That is, when only using equally keyed triplets and not too many traits (5 or less), we get such inaccurate estimation of subjects' trait scores that we would not recommend the T-IRT model for practical applications. In contrast, when a large number of traits (30) are administered, subjects' trait scores are highly accurate for sets of traits with mixed correlation patterns even when using only equally keyed triplets. As such, if one can afford to measure that many traits, T-IRT models may indeed be practically applicable even under high-stakes situation where participants are motivated to give fake answers.

Simulation Results

Overall, our simulations can be seen as extending those provided in Brown and Maydeu-Olivares (2011). The main differences are summarized in the following.

First, we simulated from a larger range of conditions, in particular with regard to the true inter-trait correlations, factor loadings, number of traits as well as number of blocks per trait. In addition to the reliability, we also investigated the RMSE of the factor scores as an absolute measure of accuracy. Furthermore, the estimated inter-trait correlations were investigated in more detail than before. When measuring 5 or less traits, we found substantial bias particularly in the inter-trait correlations when only equally keyed triplets were administered. The problem of unrealistic inter-trait correlation estimates was also found in empirical studies of real data (Morillo et al., 2016); furthermore, other studies even had to fix some inter-trait correlations to achieve sensible results (Anguiano-Carrasco et al., 2015; Guenole et al., 2018). We did not find such problems when simulating from unequally keyed triplets. This result is in line with the simulations of Frick (2017) who varied the percentage of unequally keyed triplets from 33% to 50% to 66% and found no substantial differences between those conditions with regard to the accuracy of the T-IRT model's estimates.

Second, in addition to fitting models with Mplus, we also used lavaan and Stan, both of which are—in contrast to Mplus—open source and free to use. All three packages can conveniently fit T-IRT models when using the thurstonianIRT package as an interface (Bürkner, 2018). The overall accuracy of the obtained estimates was similar across packages, although lavaan had more problems in estimating extreme factor scores. Both Mplus and lavaan turned out to have considerable convergence problems under some simulation conditions, especially conditions of larger tests that we were not able to sort out despite getting support from the package maintainers. Obtaining negative variance estimates was a particularly common issue, which prevented Mplus from returning any factor scores at all. Fixing some inter-trait correlations (Anguiano-Carrasco et al., 2015; Guenole et al., 2018) or factor loadings (Brown et al., 2017) to facilitate convergence was not an option as it would have invalidated the simulation results. In addition to convergence issues, Mplus and lavaan require an enormous amount of working memory to fit T-IRT model for larger data sets, which is unlikely to be available to most researchers. It seems that T-IRT models cannot necessarily be fit reliably using frequentist structural equation modeling. We believe this to be highly problematic, as researchers invest considerable time, effort, and money into studies applying FC questionnaires and need to have confidence in the statistical analysis they are planning to use. We did not encounter similar convergence issues in Stan, presumably because it allows users to put hard boundaries on naturally bounded parameters such as variances and set weakly informative priors on model parameters to regularize inference. Also, no memory issues were observed. Moreover, Stan performs fully Bayesian inference rather than just point estimation, which is much slower but may considerably increase the obtained information in particular for nonnormally distributed parameters (Gelman et al., 2013).

Third, to our knowledge, we are the first to systematically investigate via simulations the accuracy of T-IRT models when applied to a very large number of traits

(i.e., 30 traits), a situation that can be found in practice primarily in various versions of the OPQ personality test (Brown & Bartram, 2011). Our results indicate that, when measuring 30 traits, subjects' trait scores are highly accurate under realistic inter-trait correlations even when using only equally keyed triplets. This provides evidence that indeed, as claimed in the literature (e.g., Brown & Bartram, 2013), FC questionnaires measuring a sizable number of traits, such as the OPQ, have the potential to provide valid inference even if no unequally keyed triplets are administered.

Limitations

Despite our simulations being quite extensive, some limitations need to be discussed. First, we only investigated blocks of three items (i.e. triplets), although more items per block may improve accuracy of factor scores mainly because the number of comparisons increases substantially with the number of items being compared (Brown & Maydeu-Olivares, 2011). It thus remains unclear the degree to which larger blocks can reduce the estimation problems of factor scores found in the present study. We decided to focus on triplets because we believe them to be the most widely applicable in practice; larger blocks may increase test takers' cognitive load too much. Second, we used simulated data from 1,000 to 2,000 subjects, which may be more subjects than can be included in typical studies applying T-IRT models. However, as the performance of T-IRT models is expected to decrease with smaller samples, the issues we raised may be even more apparent in smaller samples. Third, as all former studies, we did neither simulate nor model cross-factor loadings of items and instead assumed each item to load only on a single trait.

Another limitation of our simulations is that the number of trials per condition was rather small compared with what may be considered as common in simulation studies. We simulated 100 trials in the Mplus-only simulations, and we expected this to be sufficient given the size of each single simulated data set and the highly similar summary estimates obtained from each trial. Unfortunately, we were not able to run more than one trial per condition for simulations involving lavaan or Stan, as both (but Stan in particular) are much slower than Mplus in the fitting of T-IRT models, such that each trial required several days to complete. Still, due to the size of the simulated data sets and the consistency across similar conditions, we have no reason to believe that our simulation results are unstable or depend on simulation error in a way that distorts our conclusions.

Our simulation results indicate insufficient accuracy for tests with only equally keyed blocks measuring up to 5 traits but high accuracy in the same scenario when measuring 30 traits. This raises the obvious question of what happens between those two conditions. Asked differently, how many traits are required so that T-IRT scores become non-ipsative and sufficiently accurate when administrating only equally keyed blocks? Based on the observations we made in our 30 traits simulations, this will likely require Stan or similar software to not run into the working memory issues we found when trying to fit larger T-IRT models in Mplus or lavaan. Running the

necessary simulations will thus require a substantial amount of time, even when using cluster solutions, in particular as several replications per condition are highly desirable. As such, these simulations are out of the scope of the present article but will be an important topic to study in future research. We hope our theoretical and practical considerations as well as the tools we provide as part of this article—the thurstonianIRT package and all codes used in our simulations (https://osf.io/df5yq/)—will help researchers answer this and related research questions about T-IRT models.

Practical Implications for Test Construction

Concerning test characteristics, our results suggest that the most important feature is whether or not a test includes blocks of unequally keyed items. If unequally keyed items are included, both the reliability estimates as well as the RMSEs indicate sufficient to excellent measurement precision of T-IRT models. However, as pointed out earlier, this condition cannot be applied to tests that are meant to reduce social desirability bias because it does not force respondents to choose between equally attractive items.

For practitioners considering the development of a questionnaire including unequally keyed items regardless of this issue, we need to state that it is questionable whether the precision observed in our simulations can be achieved in the field. If a desirable answer can be easily identified, most test takers will inevitably choose this answer in a high-stakes situation. Thus, the information contained in such a block of items will be very small (or even zero) and, in any case, be much smaller than the simulations may suggest.

Less fundamental but still of practical relevance is the fact that we observed serious convergence issues in Mplus and lavaan. If a high number of blocks per trait, of which some contain unequally keyed items, are used (leading to excellent reliability values), models did converge only in approximately half of the cases. This is a serious problem given the time-consuming test construction procedure and the high-stakes situations in which these tests are used. Test developers constructing tests under these conditions risk ending up without estimated person parameters unless they are using full Bayesian inference.

In contrast, tests based on only equally keyed items converge well in most conditions. However, when using only few traits (i.e., 5 or less), they show reliabilities and RMSEs that are clearly beyond an acceptable range in most conditions. In our simulations, we only reached sufficient reliabilities when we measured 5 traits with 27 blocks per trait. Under these conditions, we obtained acceptable reliabilities for traits correlated as the NEO-PIR scales and for uncorrelated traits. Yet the RMSE was inappropriately high and the reliability dropped when inter-trait correlations were 0.3. Moreover, as in all other conditions with only equally keyed items and 5 or less traits, the model failed to retrieve the inter-trait correlations accurately. Unfortunately, this bias causes the same issue the model was intended to solve: Individuals' scores on one scale again depend on their scores on other scales even if the latent traits are

actually uncorrelated. The bias, of course, considerably reduces the construct validity of T-IRT scales. However, the validity increase thought to be achieved through response biases reduction has been the central motivation for constructing T-IRT scales in the first place.

Using only equally keyed triplets yields sufficient precision only if the number of traits is very large (i.e., 30 in our simulations). For research purposes, this condition will rarely be relevant. In inductive test development, exploratory factor analyses suggesting such a high number of dimensions are quite unrealistic. For deductively developed tests, researchers will have a hard time to argue that 30 constructs do not violate the principle of parsimony. There is, however, a demand for tests with a highly differentiated factor structure, as the economically successful example of the OPQ (Brown & Bartram, 2013) shows. Of course, tests with such a high number of traits have to meet the same requirements to be faking resistant—for instance, combine items that are equally desirable in the specific context of application.

Beyond the general problems described above, results on the precision of individual scores reveal an additional issue for applications where individual scores are of interest, such as personnel selection. In most selection procedures, employers are interested in excluding particularly unqualified applicants or selecting highly qualified applicants. In most cases, these aptitude levels are associated with very low and very high values of the measured trait. Yet the standard errors of more extreme trait scores tend to be much higher than those of average trait scores. This tendency may be even more pronounced when all items are equally keyed. Thus, with respect to the measurement error, T-IRT models perform worst under exactly those conditions that are most relevant in practice. In many applications, it will be impossible to separate the best applicants from those who perform slightly above average.

To summarize our statistical and practical considerations, we conclude that in high-stakes situations where test takers are motivated to give fake answers, T-IRT models—despite all the promises—have to be applied with great care. They require a sizable number of measured traits to provide non-ipsative and sufficiently accurate estimates of participants' characteristics. In contrast, when measuring only up to five traits, as is arguably the most common case, T-IRT models are unlikely to yield non-ipsative and sufficiently accurate estimates in practice. This study thus draws a more nuanced picture of T-IRT models by demonstrating that non-ipsativity is not simply a global property of this model.

Authors' Note

To conduct the presented analyses and create this article, we used the programming language R (R Core Team, 2018) through the interface RStudio (RStudio Team, 2018). Furthermore, the following R packages were crucial (in alphabetical order): brms (Bürkner, 2017), ggplot2 (Wickham, 2016), knitr (Xie, 2014), lavaan (Rosseel, 2012), mgcv (Wood, 2011), mplusAutomation (Hallquist & Wiley, 2018), papaja (Aust & Barth, 2018), rmarkdown (Allaire et al., 2018), rstan (Carpenter et al., 2017), thurstonianIRT (Bürkner, 2018), and tidyverse (Wickham, 2017).

Acknowledgment

We thank two anonymous reviewers for their helpful comments and suggestions on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Paul-Christian Bürkner https://orcid.org/0000-0001-5765-8995

Notes

1. While running the simulations, we realized that setting all correlations to −0.3 yields an invalid correlation matrix for $n_T = 5$. Thus, no results were obtained for this combination of conditions.
2. The data from which we computed correlations of Big Five scores were obtained from a collaboration with a public agency.
3. After finishing our analysis with the correlation matrix of the OPQ CM 4.2, we also got the correlation matrices of the latest OPQ versions, OPQ32n and OPQ32r, from SHL. These versions measure 32 instead of 30 traits but otherwise have a similar correlation pattern. Simulations available on OSF (https://osf.io/df5yq/) indicate that using the correlation from OPQ32n and OPQ32r yield highly similar results than those based on the OPQ CM 4.2.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Chang, W. (2018). *Rmarkdown: Dynamic documents for R*. Retrieved from https://CRAN.R-project.org/package=rmarkdown

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment*, *33*, 83-97. doi:10.1177/0734282914550387

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

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

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology*, *69*, 25-39. doi:10.1111/j.2044-8325.1996.tb00597.x

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*, 135-160. doi:10.1007/s11336-014-9434-9

Brown, A., & Bartram, D. (2011). *OPQ32r technical manual*. Thames Ditton, England: SHL.

Brown, A., & Bartram, D. (2013). *The occupational personality questionnaire revolution: Applying item response theory to questionnaire design and scoring*. Retrieved from http://www.humandevelopmentsolutions.com/views/archives/pdf/White-Paper-OPQ32r.pdf

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

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

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*, 1135-1147. doi: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*, 36-52.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*. Retrieved from https://www.jstatsoft.org/article/view/v080i01

Bürkner, P.-C. (2018). *thurstonianIRT: Thurstonian IRT models in R*. R package Version 0.5.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*. Retrieved from https://www.jstatsoft.org/article/view/v076i01

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review*, *51*, 292-303. doi:10.1037/h0057299

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*, 267-307. doi:10.1207/s15327043hup1803_4

Döring, A. K., Blauensteiner, A., Aryus, K., Drögekamp, L., & Bilsky, W. (2010). Assessing values at an early age: The picture-based value survey for children (PBVS-C). *Journal of Personality Assessment*, *92*, 439-448. doi:10.1080/00223891.2010.497423

Dueber, D. M., Love, A. M., Toland, M. D., & Turner, T. A. (2018). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement*, *79*, 108-128. doi:10.1177/0013164417752782

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

Frick, S. (2017). *Deriving normative trait estimates from multi-dimensional forced-choice data: A simulation study* (Unpublished bachelor's thesis). University of Konstanz, Konstanz, Germany.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Guenole, N., Brown, 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*, 513-526. doi:10.1177/1073191116641181

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*, 621-638. doi: 10.1080/10705511.2017.1402334

He, J., Bartram, D., Inceoglu, I., & Vijver, F. J. R. van de. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology*, *45*, 1028-1045. doi:10.1177/0022022114534773

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167-184. doi:10.1037/h0029780

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593-1623.

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

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

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264-277. doi:10.1177/0022022104272905

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and Likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, *23*, 92-97. doi: 10.1111/ijsa.12098

Lee, J. A., Soutar, G., & Louviere, J. (2008). The best-worst scaling approach: An alternative to Schwartz's values survey. *Journal of Personality Assessment*, *90*, 335-347. doi: 10.1080/00223890802107925

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. doi:10.1016/j.paid.2017.11.031

Lewis, J. L. (2015). *Korn Ferry four dimensional executive assessment*. Retrieved from http://static.kornferry.com/media/sidebar_downloads/KF4D_Executive_Manual_FINAL.pdf

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*, 389-414. doi: 10.1177/0013164416646162

Lucía, B., Ponsoda, V., Abad, F. J., Morillo, D., Leenen, I., Rodríguez, S., & Aguado, D. A. (2014, July). A forced-choice test for assessing work-related competencies. In *Poster presented at the 9th international test commission conference*, San Sebastian, Spain. Retrieved from http://www.iic.uam.es/pdf/PosterITC2014Donosti.pdf

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*, 285-304. doi:10.1037/1082-989X.10.3.285

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*, 935-974.

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*, 222-248. doi:10.1177/1094428105275374

McCrae, R. R., & Costa, P. T. (1992). Discriminant validity of NEO-PIR facet scales. *Educational and Psychological Measurement*, *52*, 229-237.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*, 531-552. doi:10.1348/0963179042596504

Merk, J. (2016). *Die psychometrische güte des motivational value systems questionnaire: Untersuchungen zu objektivität, reliabilität und validität* [The psychometric quality of the motivational value system questionnaire: Research on objectivity, reliability and validity] (Unpublished doctoral dissertation). University of Regensburg, Bavaria, Germany. Retrieved from https://d-nb.info/1121302742/34

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P. M., Torre, J. de la, & 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*, 500-516. doi:10.1177/0146621616662226

Muthén, L., & Muthén, B. (2015). Mplus. *The comprehensive modelling program for applied researchers: User's guide*, *5*. Los Angeles, CA: Author.

Parvin, S., & Wang, P. (2014). Assessing best–worst scaling in consumer value research. In S. Rundle-Thiele, K. Kubacki, & D. Arli (Eds.), *Proceedings of Australian and New Zealand marketing academy conference 2014* (pp. 780-786). Brisbane, Queensland, Australia. Retrieved from http://opus.lib.uts.edu.au/bitstream/10453/30524/1/Parvin-Wang-Assessing BestWorst%20-%20ANZMAC%202014%20Upload.pdf

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, L. S. Wrightsman, J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www .R-project.org/

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, *25*, 189-197. doi: 10.2307/2136668

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

RStudio Team. (2018). RStudio: Integrated development for R. *Boston, MA: RStudio*. Retrieved from http://www.rstudio.com.

Rudmin, F. W., & Ahmadzadeh, V. (2001). Psychometric critique of acculturation psychology: The case of Iranian migrants in Norway. *Scandinavian Journal of Psychology*, *42*, 41-56. doi:10.1111/1467-9450.00213

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2018). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *25*, 1-13. doi:10.1177/1073191118762049

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L., & Maybey, W. (1992). *Occupational personality questionnaire manual*. Esher: SHL.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, *64*, 219-238. doi:10.1111/j.2044-8325.1991.tb00556.x

Schwarz, S. H., & Cieciuch, J. (2016). Implications of definitions of value. In F. Leong, D. Bartram, F. Cheung, K. Geisinger, & C. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 106-126). New York, NY: Oxford University Press. doi:10.1093/med:psych/9780199356942.003.0008

Stewart, G. L., Darnold, T. C., Zimmerman, R. D., Parks, L., & Dustin, S. L. (2010). Exploring how response distortion of personality measures affects individuals. *Personality and Individual Differences*, *49*, 622-628. doi:10.1016/j.paid.2010.05.035

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *79*, 273-286.

Thurstone, L. L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology*, *14*, 187-201. doi:10.1037/h0070025

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*, 600-613. doi:10.1177/0146621617703183

Waters, L. K. (1965). A note on the ''fakability'' of forced-choice scales. *Personnel Psychology*, *18*, 187-191. doi:10.1111/j.1744-6570.1965.tb00277.x

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349-363). New York, NY: Oxford University Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org/

Wickham, H. (2017). *Tidyverse: Easily install and load the ''tidyverse.''* Retrieved from https://CRAN.R-project.org/package=tidyverse

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 3-36.

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. *Implementing Reproducible Research*, *1*, 20.