Contributions to Constructing Forced-choice Questionnaires Using the Thurstonian IRT Model

Luning Sun[1], Zijie Qin[2], Shan Wang[2], Xuetao Tian[2], and Fang Luo[2]

[1]The Psychometrics Centre, University of Cambridge

[2]Faculty of Psychology, Beijing Normal University

Author Note

This work was supported by the National Natural Science Foundation of China under Grant U1911201. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Xuetao Tian, Faculty of Psychology, Beijing Normal University, Beijing, China. Email: xttian@bnu.edu.cn.

Acknowledgements

We are very grateful for the generous help we received from the Editor-in-Chief, Professor Alberto Maydeu-Olivares with the manuscript. We would also like to thank the anonymous reviewers as well as Dr Joe Watson and Dr Chia-Wen Chen for their comments on prior versions of this manuscript.

Contributions to constructing forced-choice questionnaires using the Thurstonian IRT model

Abstract

Forced-choice questionnaires involve presenting items in blocks and asking respondents to provide a full or partial ranking of the items within each block. To prevent involuntary or voluntary response distortions, blocks are usually formed of items that possess similar levels of desirability. Assembling forced-choice blocks is not a trivial process, because in addition to desirability, both the direction and magnitude of relationships between items and the traits being measured (i.e., factor loadings) need to be carefully considered. Based on simulations and empirical studies using item pairs, we provide recommendations on how to construct item pairs matched by desirability. When all pairs contain items keyed in the same direction, score reliability is improved by maximising within-block loading differences. Higher reliability is obtained when even a small number of pairs consist of unequally keyed items.

Key words: Forced-choice questionnaire, Thurstonian IRT model, social desirability

Introduction

Self-report questionnaires are widely used for the measurement of psychological constructs such as personality, attitudes, and values (Paulhus & Vazire, 2007). The most common format in self-report questionnaires involves the use of a rating scale (Moors, 2009). For instance, participants are asked to rate their agreement to a statement on a scale from "completely disagree" to "completely agree" (Likert, 1932). Likert-type rating scales have been criticised for their susceptibility to response biases, which pose a threat to the validity of the measurement (for a review, see Wetzel, Böhnke, & Brown, 2016).

Response bias is defined as "a systematic tendency to respond inaccurately to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991). Researchers and practitioners are mostly concerned with two classes of response biases: response style and response set (Paulhus, 1991). Response styles reflect a consistent preference for certain response categories over others. Common examples include extreme response style, midpoint response style, acquiescence response style, and disacquiescence response style (Van Vaerenbergh & Thomas, 2013). There are different approaches to measuring and controlling response styles. For example, acquiescence can be modelled in a confirmatory factor analysis fashion (Billiet & McClendon, 2000), or using a random intercept item factor analysis model (Maydeu-Olivares & Coffman, 2006). Multidimensional item response models have been developed to model extreme response style (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011). In these models, an additional method factor is specified to capture (and

remove) the response bias.

Response styles imply a bias in a particular direction that is independent of the content of the test items, hence the bias is often assumed to be uniform (Cheung & Chan, 2002). In contrast, response sets are generally related to the content and reflect a conscious or unconscious attempt to create a certain impression. The most frequently studied response set is socially desirable responding (SDR), which is characterised by a "tendency to give positive self-descriptions" (Paulhus, 2002). Paulhus (1984) identified two components in SDR: an involuntary component, or self-deception, and a voluntary component, or impression management, which is also referred to as faking. Self-deception may be present in surveys and research settings, where the responses have no consequences for the respondent. When the responses do have consequences for the respondent, faking becomes more prominent. For example, job applicants are inclined to provide responses they believe may increase their chances of success and are reluctant to reveal their shortcomings to potential employers (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Goffin & Christiansen, 2003; Griffith, Chmielowski, & Yoshita, 2007). It is estimated that between 30% to 50% of the applicants would elevate their personality assessment scores in the selection process (Griffith & Converse, 2011).

SDR can be measured by specifically designed social desirability scales, such as the Edwards Social Desirability Scale (Edwards, 1957), the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), and the Balanced Inventory of Desirable Responding (Paulhus, 1984). These scales examine the extent to which a respondent is

susceptible to SDR and allow for statistical control of the response bias. Several modelling approaches are also available. For example, Neill and Jackson (1970) and Ferrando (2005) developed factor analytic procedures to capture social desirability. Böckenholt (2012, 2014) proposed an item response model which can distinguish response processes that are related to the construct being measured from those related to SDR. Pavlov, Maydeu-Olivares, and Fairchild (2019) suggested that unlike uniform biases, where the effect is additive, faking leads to a non-uniform bias, and a multiplicative effect, involving the construct being measured, the respondent's willingness to fake, and their interaction is needed to capture it.

Forced-choice Format

Forced-choice questionnaires (FCQs) have been introduced to reduce response biases and increase measurement validity (Bartram, 2007). In this format, items are presented in blocks, and respondents are asked to either provide a full or partial ranking of the items within each block (e.g., "Which item describes your behaviours or attitudes the most, and which item describes them the least?"). Consider an FCQ consisting of blocks of 3 or more items using most/least like me instructions. When responses are scored using traditional methods, for example, by assigning a score of 2 to the most-like-me item, a score of 1 to the item(s) not selected, and a score of 0 to the least-like-me item, trait scores are ipsative (Meade, 2004). Ipsative scores distort construct validity, criterion-related validity, and reliability estimates (Brown & Maydeu-Olivares, 2013). As a result, scoring FCQs using traditional methods is inadvisable (Meade, 2004).

FCQs can be scored using a statistical model that captures the response process

to FCQs. It usually involves an item response theory (IRT) model, as the data arising from these questionnaires is discrete. Among the IRT models that have been suggested (for a review, see Brown, 2016a), the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011) has been widely used in research and practice. One of its advantages over other models is its flexibility. The Thurstonian IRT model can be applied to questionnaires of different formats. It can also be extended to compositional questionnaire data (Brown, 2016b). The Thurstonian IRT model can be estimated by commonly used software such as Mplus (Brown & Maydeu-Olivares, 2012), and specific packages have also been developed (e.g., the thurstonianIRT R package, Bürkner, 2019), which greatly facilitate its application.

There is consensus in the literature that the use of FCQs reduces uniform response biases (Cheung & Chan, 2002). However, no such consensus exists regarding their effectiveness in reducing non-uniform response biases. For instance, Heggestad, Morrison, Reeve, and McCloy (2006) reported that FCQs are just as affected by faking as questionnaires consisting of items using Likert-type rating scales. In many other studies, FCQs proved resistant to faking (e.g., Bartram, 2007; Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen, & Hunt, 2002; Viswesvaran, Deller, & Ones, 2007). Cao and Drasgow (2019) conducted a meta-analysis and suggested that FCQs, compared to the single-statement format, “reduce faking on personality measures in real-life selection process”. In particular, FCQs with blocks constructed in a way that matched items with similar desirability were consistently found to be more faking resistant than traditional questionnaires in which

items were not presented in blocks.

In any case, we observe a growing interest in using FCQs in applications. One reason is that practitioners (and their clients) appreciate their higher face validity, as it appears that FCQs are harder to fake. Since the appearance of suitable statistical models for this kind of data, such as the Thurstonian IRT model , many FCQs have been developed and employed in practice (e.g., Anguiano-Carrasco, MacCann, Geiger, Seybert, & Roberts, 2015; Guenole, Brown, & Cooper, 2018; Joubert, Inceoglu, Bartram, Dowdeswell, & Lin, 2015; Merk, Schlotz & Falter, 2017; Ng et al., 2020; SHL., 2013).

When researchers and practitioners are tasked to construct an FCQ using the Thurstonian IRT model, they are confronted with the challenge of assembling items into blocks, aiming to maximise both the reliability of the trait scores and the resistance to faking (Bürkner, Schulte, & Holling, 2019). Previous research has provided general guidelines to maximise the reliability of trait scores (Brown & Maydeu-Olivares, 2011). Further, numerous studies have identified that faking can be reduced by matching the items within each block based on their desirability, which obstructs respondents from endorsing only socially favourable response options (e.g., Cao & Drasgow, 2019; Christiansen et al, 2005; Jackson et al., 2000; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). However, not all FCQs match items on their desirability when assembling the blocks. For example, Anguiano-Carrasco et al. (2015) developed a forced-choice measure of typical performance emotional intelligence, where 24 statements were randomly grouped into 8 triplets without balancing their desirability.

There has yet to be any research which establishes a method through which these two objectives can be simultaneously pursued.

The objective of this article is to provide practical guidance on assembling forced-choice blocks to simultaneously maximise faking resistance and score reliability under the Thurstonian IRT model. Below we explain why this is challenging, followed by two solutions that we propose to address the problem. Then we will present the results of two simulation studies designed to investigate the effectiveness of the solutions proposed, and one empirical study where we evaluate the feasibility of implementing these two solutions in practice. We conclude our article by summarising the findings, discussing their implications, and making recommendations for constructing faking resistant FCQs using the Thurstonian IRT model.

The Problem

Based on Thurstone's law of comparative judgement, Brown and Maydeu-Olivares developed a Thurstonian IRT model (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). The model is a reparameterisation of the Thurstonian factor model, an extension of the ordinal factor analysis for rating scale items to accommodate items presented in blocks. The Thurstonian factor model is a second-order factor model, providing factor scores for the traits (the second-order factors) as well as for the items (the first-order factors), which are of no interest in applications. Therefore, the model is reparameterised into a first-order model, namely the Thurstonian IRT model, which provides scores only for the traits. Specifically, in a block containing two items (item i measuring trait η_a and item k measuring trait η_b) with factor loadings λ_i and λ_k,

respectively, when item i is preferred over item *k*, the probability of the observed comparative response y_l is,

$$\Pr(y_l = 1|\eta_a, \eta_b) = \Phi\left(\frac{-\gamma_l + \lambda_i\eta_a - \lambda_k\eta_b}{\sqrt{\Psi_i^2 + \Psi_k^2}}\right)$$

where Φ(x) denotes the cumulative standard normal distribution function evaluated at x, γ_l is the threshold for binary outcome y_l, and Ψ_i^2 and Ψ_k^2 are the uniquenesses of the latent response variables. The parameters in the Thurstonian IRT model can be estimated using limited information methods such as unweighted least squares (ULS) and diagonally weighted least squares (DWLS) (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). Once the model parameters are estimated, the respondents' traits η_a and η_b can be yielded following a Bayes modal procedure, e.g., maximum a posteriori (MAP).

It is notable that desirability is not explicitly mentioned in the recommendations offered by Brown and Maydeu-Olivares (2011) when proposing their model. According to their recommendations, to ensure accurate estimation of the latent traits, half of the forced-choice blocks should be unequally keyed (i.e., combining both positively and negatively keyed items). Following the suggestion, for instance, Walton, Cherkasova, and Roberts (2020) constructed three FCQs, all of which were formed of blocks of three adjectives or statements, mixing one negatively keyed with two positively keyed. As positively and negatively keyed items usually differ substantially in their desirability levels, item blocks combining both can often be faked (Bürkner et al., 2019; Wang, Qiu, Chen, Ro, & Jin, 2017). For example, in a block consisting of three adjectives, *sympathetic*, *organised*, and *shy*, a respondent who is motivated to fake can easily do

so by choosing *sympathetic* over *shy* and *organised* over *shy*.

When the underlying constructs are formulated with congruent valence (all positive or all negative), items keyed in the same direction are more likely to have similar social desirability (e.g., Fisher, Robie, Christiansen, Speer, & Schneider, 2019). Hence, to match items on their social desirability, equally keyed blocks are preferred when constructing FCQs. Unfortunately, when modelling an FCQ comprised of equally keyed blocks only, results may be inaccurate, and item parameter estimates may be biased (Brown & Maydeu-Olivares, 2011; Wang et al., 2017). For instance, the simulation results by Bürkner et al. (2019) reveal that the Thurstonian IRT model fails to yield sufficiently accurate trait scores and inter-trait correlations in FCQs measuring up to five traits containing equally keyed blocks only. This poses a serious challenge for constructing FCQs under the Thurstonian IRT model. On the one hand, at the risk of faking resistance, the Thurstonian IRT model proves effective in scoring an FCQ with half of the blocks being unequally keyed; on the other hand, to effectively reduce faking, for FCQs with only equally keyed blocks, the Thurstonian IRT model is questioned as a proper scoring method.

Potential Solutions to Mitigate the Problem

To mitigate this problem, we propose two potential solutions. First, instead of having half of the blocks being unequally keyed, we suggest that a smaller number of unequally keyed blocks be included in an FCQ, so that it can be adequately analysed by the Thurstonian IRT model. This has never been examined in previous simulation studies. Bürkner et al. (2019) and Fisher et al. (2019) argued that it is difficult or even

impossible to build blocks from items that are matched on social desirability and, at the same time, unequally keyed. However, this is not consistent with empirical findings. For instance, Ng et al. (2021) developed a forced-choice extant character scale, where each triplet block combined positive and negative statements of similar social desirability. Using items from the International Personality Item Pool (Goldberg, 1999), Wetzel and Frick (2020) constructed 20 forced-choice triplets that were matched on social desirability, four of which were unequally keyed. Lee, Lee, and Stark. (2018) also found an overlap in social desirability among the positive and negative personality items., allowing the creation of unequally keyed blocks using items matched on social desirability.

Second, we suggest that, using items that are matched on social desirability, equally keyed blocks should be optimally designed to feed the most information to the Thurstonian IRT model. Specifically, with a given item bank, we can build item blocks in a way that maximises within-block loading differences and prioritises items with high factor loadings, so as to achieve a higher reliability in the estimation (Bürkner, 2022).

Extensive simulations are performed to evaluate these solutions. In Simulation 1, by varying the number of unequally keyed blocks in an FCQ, we aim to understand their influence on the measurement precision of the FCQ under the Thurstonian IRT model. In Simulation 2, where the FCQ contains equally keyed blocks only, we are interested in the effects of the within-block loading difference and the absolute factor loading, and experiment with different block assembling approaches.

To better inform real-world practice, we take into account the requirement of matched social desirability when assembling forced-choice blocks. Therefore, we introduce a social desirability index (SDI) for each item and ensure that only items with comparable SDIs are allowed to be grouped together. As items keyed in the same direction are more likely to have similar social desirability, we assume that the SDIs are positively correlated with the factor loadings. Notably, a correlation of zero would isolate the SDIs from the factor loadings and result in similar conditions as those in Brown and Maydeu-Olivares (2011). This assumption is further evaluated in the subsequent empirical study, where the Thurstonian IRT model is applied to three forced-choice personality measures that are constructed for a high-stakes assessment.

Simulation 1

Since we are not interested in the effect of block size, for the ease with the simulations, we decide to adopt the format of item pairs. We fix the number of traits to five, which is a common number of factors in personality assessment. To make the interpretation of the results simpler, all traits are positively defined, hence, positively correlated.

Design

In Simulation 1, we examined a total of 72 conditions by crossing the following factors: a) correlation between the factor loadings and the SDIs (0, 0.3, 0.6); b) keyed direction of forced-choice blocks (all being equally keyed, 1/6 being unequally keyed, 1/3 being unequally keyed); c) number of blocks (Short: 30 item pairs with 12 items per trait; Long: 60 item pairs with 24 items per trait); and d) inter-trait correlation (uniform correlations of 0, 0.3, and 0.6, and one real-world correlation matrix from the NEO-PI-

R in McCrae & Costa, 1992). Each condition was replicated (up to) 100 trials using different item banks and response data.

Below, we describe the simulation procedure in detail.

Generation of the Item Banks

Considering the relatively strict requirement on matched social desirability for item pairing and also the routine practice in test development, we generated twice the number of items needed in each FCQ. For short FCQs with 30 item pairs, each item bank consisted of 120 items, 60 being positively keyed and 60 being negatively keyed. For long FCQs with 60 item pairs, each item bank consisted of 240 items, 120 being positively keyed and 120 being negatively keyed. An equal number of items were assigned to each of the five traits. Their item parameters, including standardised factor loadings, SDIs, and intercepts, were simulated as described below.

1) Standardised factor loadings followed a uniform distribution from 0.45 to 0.95 (-0.95 to -0.45 for negative items), which covers a broad range that is commonly observed in items available from the International Personality Item Pool (Goldberg, 1999).

2) SDIs were simultaneously generated as the factor loadings. A social desirability scale similar to the one used in Wetzel and Frick (2020) was created, ranging from 1 to 5, where 3 was the mean value. In a positively defined trait, an item with a positive factor loading should indicate a socially desirable behaviour, hence an above-average SDI. SDIs for the positive items were sampled from a U(3, 5), and those for the negative items from a U(1, 3).

3) Intercepts were drawn from a U(-1, 1).

Two hundred sets of item parameters were generated (100 for short FCQs and 100 for long FCQs). The correlation between the factor loadings and the SDIs was manipulated by adjusting the matching relationship between them. With three correlation levels (0, 0.3, and 0.6), a total of 600 item banks were created.

Construction of the Item Pairs

Three FCQs were constructed out of each item bank, one with equally keyed item pairs only, one with 1/6 pairs being unequally keyed, and one with 1/3 pairs being unequally keyed. All item pairs were multidimensional (i.e., the two items were from different traits). Each item could only appear in one item pair. There were equal numbers of item pairs measuring each trait, and equal numbers of item pairs for all possible trait combinations.

To match the SDIs within each item pair, we constrained the difference in the SDIs to not exceed 0.5. As there was no overlap in the SDIs among the positive and negative items, it was impossible to construct a large number of unequally keyed item pairs with a sufficiently small within-pair SDI difference in some item banks. Specifically, in 90 “short” item banks and 39 “long” item banks, we were not able to construct FCQs with 1/3 item pairs being unequally keyed, resulting in a lack of 129 FCQs (21.5%) for these conditions. In contrast, we managed to construct FCQs with 1/6 item pairs being unequally keyed in all item banks. Across all final FCQs, the mean of the average within-pair SDI difference was 0.25 (minimum 0.15, maximum 0.33).

Simulation of the Response Data

Four response data sets were simulated for each FCQ, corresponding to the four levels of inter-trait correlations. In the first one, we set a correlation of zero among the traits. In the following two, all inter-trait correlations were fixed to a uniform value of either 0.3 or 0.6, as representatives of commonly reported low and moderate correlations among personality traits, respectively. In the last one, we took a real-world correlation matrix from the NEO-PI-R (McCrae & Costa, 1992; see Table 1), following the example of Bürkner et al. (2019). To avoid confusion in the interpretation, particularly due to the existence of both positively and negatively keyed items, we renamed the Neuroticism factor Emotional Stability and reversed the negative correlations. In this case, the average inter-trait correlation was 0.19.

Table 1

Inter-trait correlation matrix from the NEO-PI-R (McCrae & Costa, 1992)

	NEO-PI-R			
	ES	E	C	A
E	0.21			
C	0.53	0.27		
A	0.25	0.00	0.24	
O	0.00	0.40	0.00	0.00

Note. E = extraversion; C = conscientiousness; A = agreeableness; O = openness to experience; ES = emotional stability.

In each data set, a sample of 1,000 observations (i.e., subjects) was simulated. For each observation, 30 or 60 binary outcome variables were generated to indicate the responses to the forced-choice blocks. All latent traits were normally distributed with a variance of 1. The error variance of each item was set to 0.5 (and the error variance of the item pair was 1).

The response data were analysed using the Thurstonian IRT model with Mplus 8.3 (Muthén & Muthén, 1998-2019). All models in the present article employed the

estimator ULSMV. The latent traits were estimated using the MAP method.

Results

Below, we report the simulation results of the convergence rate, the model fit, and the recovery of the inter-trait correlations, the item parameters, and the latent trait scores.

Convergence Rate

The convergence rate[1] was 100% for all conditions with unequally keyed item pairs (Table 2). It became more of an issue when an FCQ was short and had only equally keyed item pairs, as on average 20% of the trials did not converge in these conditions.

[1] In a very small number of trials where the Thurstonian IRT model was able to converge in Mplus, more than half of the factor loadings were not significant, suggesting local minima that failed to recover the item parameters. Such trials were also considered as non-convergence in our results.

Table 2

Simulation results of the convergence rate and the recovery of the latent trait scores under different conditions in Simulation 1

Conditions			all equally keyed			1/6 being unequally keyed			1/3 being unequally keyed		
BpT	S-F Corr	I-T Corr	Conv	m(rel)	m(RMSE)	Conv	m(rel)	m(RMSE)	Conv	m(rel)	m(RMSE)
12	0	0	0.76	0.57	0.65	1.00	0.67	0.58	1.00	0.69	0.56
12	0	0.3	0.85	0.40	0.78	1.00	0.65	0.59	1.00	0.69	0.55
12	0	0.6	0.92	0.28	0.85	1.00	0.66	0.59	1.00	0.72	0.53
12	0	NEO	0.89	0.47	0.73	1.00	0.66	0.58	1.00	0.70	0.55
12	0.3	0	0.80	0.56	0.67	1.00	0.66	0.58	1.00	0.68	0.56
12	0.3	0.3	0.86	0.39	0.79	1.00	0.63	0.60	1.00	0.68	0.56
12	0.3	0.6	0.80	0.28	0.86	1.00	0.64	0.60	1.00	0.71	0.54
12	0.3	NEO	0.84	0.47	0.73	1.00	0.65	0.59	1.00	0.69	0.56
12	0.6	0	0.67	0.55	0.67	1.00	0.65	0.59	1.00	0.67	0.57
12	0.6	0.3	0.81	0.36	0.81	1.00	0.62	0.62	1.00	0.67	0.57
12	0.6	0.6	0.71	0.22	0.91	1.00	0.62	0.62	1.00	0.70	0.55
12	0.6	NEO	0.68	0.43	0.77	1.00	0.63	0.60	1.00	0.68	0.57
24	0	0	1.00	0.69	0.55	1.00	0.80	0.45	1.00	0.82	0.43
24	0	0.3	1.00	0.55	0.67	1.00	0.79	0.46	1.00	0.82	0.42
24	0	0.6	1.00	0.47	0.73	1.00	0.79	0.46	1.00	0.83	0.41
24	0	NEO	1.00	0.60	0.63	1.00	0.79	0.46	1.00	0.82	0.42
24	0.3	0	0.97	0.69	0.56	1.00	0.79	0.46	1.00	0.81	0.43
24	0.3	0.3	1.00	0.54	0.68	1.00	0.78	0.47	1.00	0.82	0.43
24	0.3	0.6	1.00	0.45	0.75	1.00	0.78	0.47	1.00	0.83	0.42
24	0.3	NEO	1.00	0.60	0.64	1.00	0.79	0.46	1.00	0.82	0.43
24	0.6	0	0.94	0.68	0.57	1.00	0.78	0.47	1.00	0.81	0.44
24	0.6	0.3	0.98	0.50	0.71	1.00	0.77	0.49	1.00	0.81	0.44

24	0.6	0.6	0.96	0.37	0.80	1.00	0.77	0.49	1.00	0.82	0.43
24	0.6	NEO	0.98	0.57	0.66	1.00	0.77	0.48	1.00	0.81	0.44

Note. BpT = blocks per trait; S-F Corr = correlation between the SDIs and the factor loadings; I-T Corr: inter-trait correlation; Conv = convergence rate; m(rel) = mean of the reliability; m(RMSE) = mean of the RMSE; all equally keyed = FCQs with equally keyed blocks only; 1/6 being unequally keyed = FCQs with 1/6 blocks being unequally keyed; 1/3 being unequally keyed = FCQs with 1/3 blocks being unequally keyed.

Model Fit

More than 97% of the models that successfully converged exhibited non-significant *p*-values for the chi-square test. Across all conditions, the maximum RMSEA was 0.016, and the minimum CFI was 0.955.

Recovery of the Inter-trait Correlations and Other Item Parameters

Table 3 presents the average relative biases (ARBs) of the factor loadings and the thresholds,[2] as well as the biases of the inter-trait correlations. Under most conditions with equally keyed item pairs only, the ARBs of the factor loadings averaged across all trails were between -10% and 2% (mostly negative). The inclusion of unequally keyed item pairs brought considerable improvement. Except for one condition, all ARBs of the factor loadings were below 2%, regardless of the percentage of the unequally keyed item pairs in the FCQs. The keyed direction of the item pairs seemed to have little to no influence on the ARBs of the thresholds, which were stably below 3%. Meanwhile, the estimated inter-trait correlations showed a negative bias in conditions with equally keyed item pairs only, whereas this bias was effectively eliminated when unequally keyed item pairs were introduced. Our results are largely consistent with the findings of Bürkner et al. (2019), although our simulations employ a smaller percentage of unequally keyed blocks, which is relatively easy to achieve in reality.

[2] When calculating the ARBs of the factor loadings and the thresholds, we excluded certain outliers, which were defined as absolute values of above five for factor loadings and absolute values of above ten for thresholds.

Table 3

Simulation results of the recovery of the factor loadings, the intercepts, and the inter-trait correlations under different conditions in Simulation 1

Conditions			all equally keyed			1/6 being unequally keyed			1/3 being unequally keyed		
BpT	S-F Corr	I-T Corr	ARB(λ)	ARB(γ)	corr-bias	ARB(λ)	ARB(γ)	corr-bias	ARB(λ)	ARB(γ)	corr-bias
12	0	0	-0.48%	1.35%	-0.04	1.03%	-0.64%	-0.01	1.00%	0.86%	0.00
12	0	0.3	-7.14%	1.57%	-0.16	1.30%	0.29%	0.00	1.08%	1.33%	0.00
12	0	0.6	-3.49%	1.43%	-0.10	2.27%	0.40%	0.00	2.06%	0.92%	0.01
12	0	NEO	-4.93%	2.62%	-0.13	1.14%	0.33%	-0.01	1.11%	1.08%	0.00
12	0.3	0	0.81%	2.61%	-0.03	0.88%	-0.70%	-0.01	1.24%	2.42%	0.00
12	0.3	0.3	-7.62%	0.18%	-0.16	1.14%	0.05%	0.00	1.15%	1.55%	0.00
12	0.3	0.6	-5.85%	1.76%	-0.11	1.63%	-0.33%	0.00	1.65%	0.88%	0.00
12	0.3	NEO	-3.74%	-0.83%	-0.11	1.01%	0.17%	-0.01	1.25%	1.85%	0.01
12	0.6	0	1.83%	0.85%	-0.02	1.05%	1.45%	-0.01	1.25%	-2.40%	0.00
12	0.6	0.3	-6.05%	1.23%	-0.16	1.49%	1.11%	0.00	1.11%	-1.78%	0.00
12	0.6	0.6	-15.25%	0.86%	-0.22	1.36%	1.02%	-0.01	1.54%	-0.85%	0.00
12	0.6	NEO	-3.49%	-0.48%	-0.10	1.00%	1.19%	-0.01	1.11%	-0.52%	0.00
24	0	0	-2.21%	2.47%	-0.07	0.65%	-0.69%	-0.01	0.89%	2.51%	0.00
24	0	0.3	-7.57%	1.94%	-0.14	0.69%	-0.45%	0.00	1.10%	2.74%	0.00
24	0	0.6	-7.62%	1.53%	-0.09	0.80%	-0.19%	0.00	1.55%	2.42%	0.00
24	0	NEO	-4.99%	1.49%	-0.11	0.40%	-0.37%	-0.01	1.06%	2.52%	0.00
24	0.3	0	-1.93%	1.05%	-0.06	0.20%	-0.01%	-0.01	0.88%	1.19%	0.00
24	0.3	0.3	-7.63%	1.32%	-0.14	0.40%	-0.06%	0.00	1.18%	1.73%	0.01
24	0.3	0.6	-7.37%	1.44%	-0.09	0.96%	0.33%	0.00	1.50%	1.46%	0.00
24	0.3	NEO	-4.51%	0.88%	-0.10	0.21%	0.41%	-0.01	0.85%	1.24%	0.00
24	0.6	0	-3.04%	-0.64%	-0.09	0.40%	2.27%	-0.01	0.90%	2.04%	0.00
24	0.6	0.3	-9.03%	-0.63%	-0.19	0.53%	2.03%	-0.01	1.16%	2.74%	0.00

24	0.6	0.6	-14.40%	-1.08%	-0.20	1.27%	1.85%	0.00	1.57%	2.62%	0.01
24	0.6	NEO	-7.86%	-0.61%	-0.17	0.37%	1.73%	-0.01	0.86%	2.25%	0.00

Note. BpT = blocks per trait; S-F Corr = correlation between the SDIs and the factor loadings; I-T Corr: inter-trait correlation; ARB(λ) = average relative bias of the factor loadings; ARB(γ) = average relative bias of the intercepts; corr-bias = bias of the inter-trait correlations; all equally keyed = FCQs with equally keyed blocks only; 1/6 being unequally keyed = FCQs with 1/6 blocks being unequally keyed; 1/3 being unequally keyed = FCQs with 1/3 blocks being unequally keyed.

Recovery of the Latent Trait Scores

We computed the squared correlation between the true latent trait scores and the factor scores estimated by the Thurstonian IRT model using MAP, which served as an indicator for the reliability of the MAP scores (Brown & Maydeu-Olivares, 2011). Our findings are in line with the results of Brown and Maydeu-Olivares (2011) and Bürkner et al. (2019). Specifically, as shown in Table 2 and Figure 1, long FCQs with more item pairs yielded higher reliabilities than short ones; FCQs containing unequally keyed item pairs demonstrated greater performance at recovering the latent trait scores than those with equally keyed item pairs only. Focusing on the long FCQs, the average reliability was 0.56 with equally keyed item pairs only, and this was increased to 0.78 and 0.82, respectively, when 1/6 and 1/3 item pairs were unequally keyed.

The RMSE of the trait scores pointed in the same direction. RMSEs higher than 0.5 were observed in short FCQs, as well as in long FCQs with equally keyed item pairs only. The average RMSE was reduced to 0.47 in long FCQs with 1/6 item pairs being unequally keyed, and further to 0.43 in those with 1/3 item pairs being unequally keyed. Considering the difficulty in constructing unequally keyed item pairs that were matched on the SDIs, the additional benefit of having 1/3 instead of 1/6 item pairs being unequally keyed seemed rather limited.

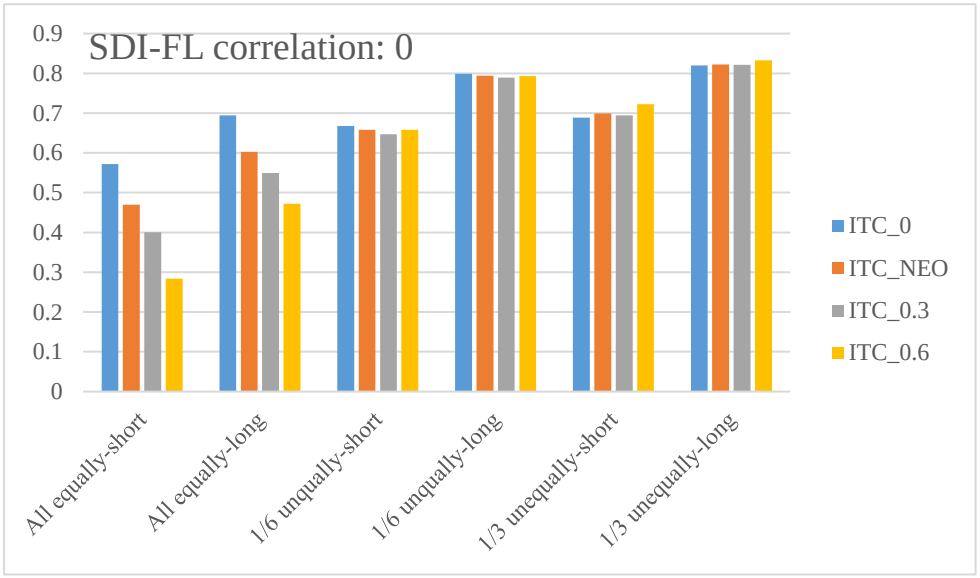
Consistent with the previous literature, the effect of the inter-trait correlation on the recovery of the latent trait scores depended on the keyed direction of the item pairs, suggesting an interaction between the two factors. For FCQs with equally keyed item pairs only, a lower inter-trait correlation led to an improvement in the accuracy of the

trait scores. The highest reliability and the lowest RMSE were always observed in conditions where the traits were not correlated with each other. Nonetheless, when unequally keyed item pairs were included in the FCQs, the influence of the inter-trait correlation became negligible.

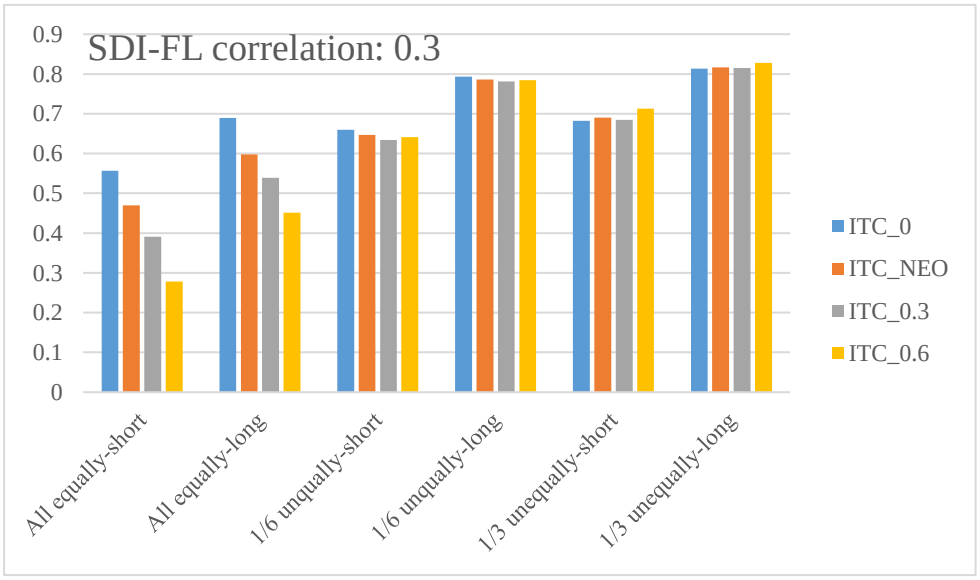
Although similar patterns were observed across different correlations between the factor loadings and the SDIs, a subtle but consistent decrease in the accuracy of the trait scores was coupled with a higher correlation between the factor loadings and the SDIs. We hypothesise that this effect came through the within-pair loading differences. When two items were paired up, a constraint was posed on the SDIs to ensure matched social desirability. With a higher correlation between the factor loadings and the SDIs, the discrepancy in the factor loadings between the items would be reduced, resulting in a loss of information in the Thurstonian IRT model. We will investigate this hypothesis in Simulation 2.

Figure 1

Estimation reliability of the latent trait scores under different conditions in Simulation 1: a. with a correlation of zero between the SDIs and the factor loadings; b. with a correlation of 0.3 between the SDIs and the factor loadings; and c. with a correlation of 0.6 between the SDIs and the factor loadings



(a)



(b)

Simulation 2

Design

In Simulation 2, we focus on FCQs with equally keyed item pairs only. By experimenting with different item pairing strategies, we aim to uncover the factors that influence the measurement precision of an FCQ under the Thurstonian IRT model. A total of 144 conditions were investigated by fully crossing the five factors below: a) absolute values of the factor loadings (Low, where factor loadings were sampled uniformly between 0.45 and 0.75 for positively keyed items and between -0.75 and -

0.45 for negatively keyed items; High, where factor loadings were sampled uniformly between 0.65 and 0.95 for positively keyed items and between -0.95 and -0.65 for negatively keyed items; and Wide, where factor loadings were sampled uniformly between 0.45 and 0.95 for positively keyed items and between -0.95 and -0.45 for negatively keyed items); b) within-pair loading differences (Maximum, where items from different traits with similar SDIs were paired up in a way that maximised the within-pair loading differences;[3] and Random, where items from different traits with similar SDIs were paired up randomly[4]); c) correlation between the factor loadings and the SDIs (0, 0.3, 0.6); d) number of blocks (Short: 30 item pairs with 12 items per trait, Long: 60 item pairs with 24 items per trait); and e) inter-trait correlation (uniform correlations of 0, 0.3, and 0.6, and one real-world correlation matrix from the NEO-PI-R in McCrae & Costa, 1992).

Each condition was replicated in 100 trials using different item banks and response data. Following a similar procedure as in Simulation 1, we generated the item banks, constructed the FCQs and simulated the response data. Specifically, 1,800 item banks were generated, including 600 with Low factor loadings, 600 with High factor loadings, and 600 with Wide factor loadings. Out of each item bank, two FCQs were constructed, using Random pairing and Maximum pairing, respectively. Information about the factor

[3] Under Maximum conditions, an item was paired up with the one that had the most distant factor loading among all SDI-matched items from a different trait. Item pairs were constructed sequentially. For each trial, 100 test forms were assembled using random items as the starting point, and the test form with the highest average within-pair loading difference was used in the simulation.

[4] Under Random conditions, in each item bank, we selected the same set of items that were used under the Maximum condition so that there was no difference in the values of the factor loadings between the corresponding trials in the two conditions.

loadings and the SDIs aggregated across all trials under each condition is shown in Table 4. As expected, the Low conditions showed the lowest average absolute factor loadings, the High conditions showed the highest average absolute factor loadings, whereas the average absolute factor loadings of the Wide conditions lay in the middle. With regard to the within-pair loading difference, the Wide condition showed the largest difference, whereas the differences in the Low and High conditions were comparable. In comparison to the Random pairing, the Maximum pairing effectively increased the within-pair loading difference by at least 50%. Furthermore, slightly larger differences in the SDIs were observed under the Maximum pairing conditions, particularly when the correlation between the factor loadings and the SDIs was higher. All standard deviations (SDs) across the 100 trials in each condition were minimum (not above 0.03), suggesting little simulation error.

Table 4

Information about the factor loadings and the SDIs of the item pairs under different conditions in Simulation 2

Condition			Random pairing						Maximum pairing					
BpT	S-F Corr	Range	m(abs λ)	SD(abs λ)	m(diff λ)	SD(diff λ)	m(diff SDI)	SD(diff SDI)	m(abs λ)	SD(abs λ)	m(diff λ)	SD(diff λ)	m(diff SDI)	SD(diff SDI)
12	0	(0.45, 0.75)	0.60	0.01	0.13	0.02	0.23	0.02	0.60	0.01	0.20	0.01	0.23	0.03
12	0	(0.65, 0.95)	0.80	0.01	0.13	0.02	0.23	0.03	0.80	0.01	0.20	0.01	0.23	0.03
12	0	(0.45, 0.95)	0.70	0.01	0.21	0.03	0.23	0.03	0.70	0.01	0.33	0.01	0.23	0.02
12	0.3	(0.45, 0.75)	0.60	0.01	0.12	0.01	0.23	0.02	0.60	0.01	0.19	0.01	0.24	0.03
12	0.3	(0.65, 0.95)	0.80	0.01	0.12	0.02	0.23	0.03	0.80	0.01	0.19	0.01	0.24	0.03
12	0.3	(0.45, 0.95)	0.70	0.01	0.20	0.03	0.23	0.03	0.70	0.01	0.32	0.01	0.24	0.02
12	0.6	(0.45, 0.75)	0.60	0.01	0.11	0.01	0.24	0.02	0.60	0.01	0.17	0.01	0.27	0.02
12	0.6	(0.65, 0.95)	0.80	0.01	0.10	0.01	0.23	0.03	0.80	0.01	0.17	0.01	0.27	0.02
12	0.6	(0.45, 0.95)	0.70	0.01	0.18	0.02	0.23	0.03	0.70	0.01	0.28	0.01	0.27	0.03
24	0	(0.45, 0.75)	0.60	0.00	0.12	0.01	0.23	0.02	0.60	0.00	0.20	0.01	0.24	0.02
24	0	(0.65, 0.95)	0.80	0.00	0.12	0.01	0.23	0.02	0.80	0.00	0.20	0.01	0.23	0.02
24	0	(0.45, 0.95)	0.70	0.01	0.21	0.02	0.23	0.02	0.70	0.01	0.33	0.01	0.23	0.02
24	0.3	(0.45, 0.75)	0.60	0.00	0.12	0.01	0.23	0.02	0.60	0.00	0.19	0.01	0.25	0.02
24	0.3	(0.65, 0.95)	0.80	0.00	0.12	0.01	0.23	0.02	0.80	0.00	0.19	0.01	0.25	0.02
24	0.3	(0.45, 0.95)	0.70	0.01	0.20	0.02	0.23	0.02	0.70	0.01	0.32	0.01	0.25	0.02
24	0.6	(0.45, 0.75)	0.60	0.00	0.10	0.01	0.23	0.02	0.60	0.00	0.17	0.01	0.27	0.02
24	0.6	(0.65, 0.95)	0.80	0.01	0.10	0.01	0.23	0.02	0.80	0.01	0.17	0.01	0.28	0.02
24	0.6	(0.45, 0.95)	0.70	0.01	0.17	0.01	0.23	0.02	0.70	0.01	0.28	0.01	0.28	0.02

Note. BpT = blocks per trait; S-F Corr = correlation between the SDIs and the factor loadings; Range: range of the factor loadings in the item bank; m(abs λ) = mean of the absolute values of the factor loadings; SD(abs λ) = standard deviation of the absolute values of the factor loadings; m(diff λ) = mean of the within-block loading differences; SD(diff λ) = standard deviation of the within-block loading differences; m(diff SDI) = mean of the within-block SDI differences; SD(diff SDI) = standard

deviation of the within-block SDI differences.

Results

Convergence Rate

Across different levels of inter-trait correlation, the convergence rate was comparable. For brevity, we present here only the results for conditions with the inter-trait correlation taken from the NEO-PI-R. As shown in Table 5, the Thurstonian IRT model in conditions where the factor loadings were sampled from a wide range was more likely to converge than in their counterparts. Having more item pairs and employing Maximum pairing led to an increase in the convergence rate.

Table 5

Simulation results of the convergence rate and the recovery of the latent trait scores under conditions with inter-trait correlations taken from the NEO-PI-R in Simulation 2

Condition			Random pairing			Maximum pairing		
BpT	S-F Corr	Range	Conv	m(rel)	m(RMSE)	Conv	m(rel)	m(RMSE)
12	0	(0.45, 0.75)	0.69	0.40	0.78	0.87	0.44	0.75
12	0	(0.65, 0.95)	0.58	0.46	0.74	0.90	0.50	0.71
12	0	(0.45, 0.95)	0.96	0.51	0.70	1.00	0.56	0.66
12	0.3	(0.45, 0.75)	0.64	0.38	0.79	0.86	0.44	0.75
12	0.3	(0.65, 0.95)	0.57	0.45	0.74	0.90	0.49	0.72
12	0.3	(0.45, 0.95)	0.97	0.51	0.70	0.99	0.56	0.66
12	0.6	(0.45, 0.75)	0.58	0.38	0.79	0.79	0.43	0.76
12	0.6	(0.65, 0.95)	0.43	0.42	0.77	0.76	0.48	0.73
12	0.6	(0.45, 0.95)	0.85	0.48	0.72	1.00	0.54	0.68
24	0	(0.45, 0.75)	0.94	0.53	0.69	1.00	0.58	0.65
24	0	(0.65, 0.95)	0.93	0.56	0.67	1.00	0.62	0.62
24	0	(0.45, 0.95)	1.00	0.64	0.60	1.00	0.71	0.54
24	0.3	(0.45, 0.75)	0.83	0.53	0.70	0.97	0.58	0.65
24	0.3	(0.65, 0.95)	0.86	0.56	0.67	1.00	0.62	0.62
24	0.3	(0.45, 0.95)	1.00	0.63	0.60	1.00	0.70	0.55
24	0.6	(0.45, 0.75)	0.74	0.51	0.71	1.00	0.56	0.66
24	0.6	(0.65, 0.95)	0.76	0.54	0.69	0.99	0.60	0.64
24	0.6	(0.45, 0.95)	1.00	0.62	0.62	1.00	0.68	0.57

Note. BpT = blocks per trait; S-F Corr = correlation between the SDIs and the factor loadings; Range: range of the factor loadings in the item bank; Conv = convergence rate; m(rel) = mean of the reliability; m(RMSE) = mean of the RMSE.

Model Fit

Among all of the 12,579 trials that successfully converged, less than 1.8% showed a significant *p*-value for the chi-square test. Only three trials had CFIs below .95 (.933, .939 and .940, respectively), and the maximum RMSEA was .017.

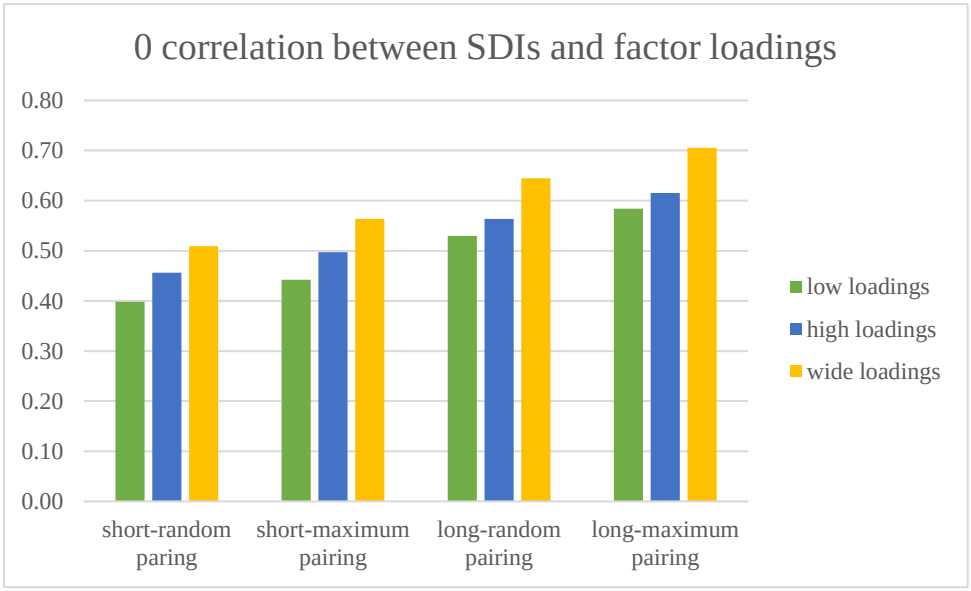
Recovery of the Latent Trait Scores

For the reason of readability, we report here only the results for selected simulation conditions. Detailed results for the complete simulations can be found online (Sun, 2022). The results about the reliability and the RMSE of the trait scores are very consistent, so we focus on the reliability here.

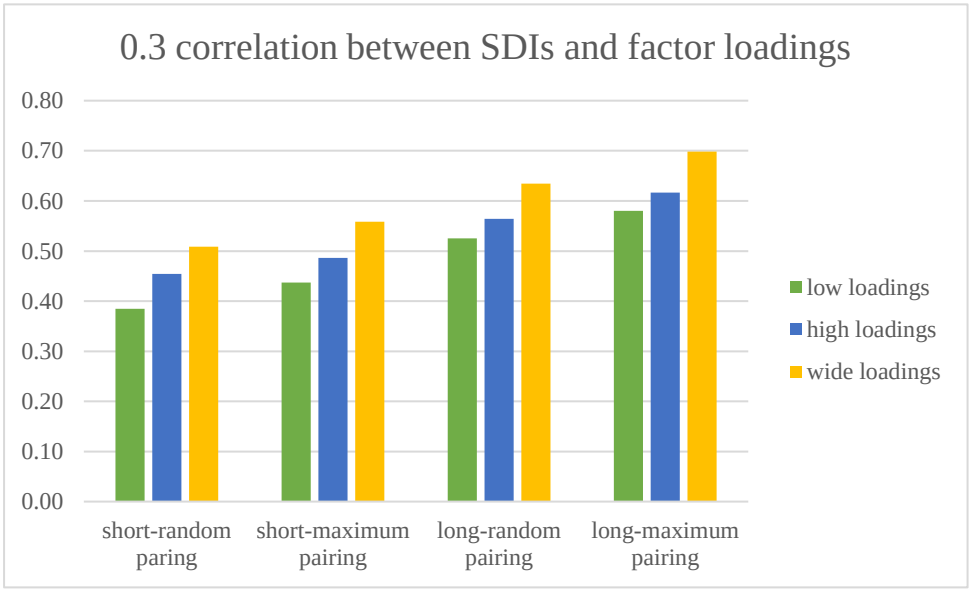
As illustrated in Figure 2, under the conditions using the NEO-PI-R correlation matrix, three factors effectively influenced the estimation reliability, i.e., the number of item pairs, the within-pair loading differences, and the absolute values of the factor loadings. Specifically, long FCQs with more item pairs obtained higher reliability. Maximum pairing led to an increase in the estimation reliability. As exactly the same set of items were used in the corresponding Random pairing and Maximum pairing trials, which ensured identical factor loadings between the two conditions, we could confidently attribute the benefit of Maximum pairing to the increased within-pair loading difference. For other simulation conditions held constant, the High loading condition outperformed the Low loading condition, while the Wide loading condition exhibited the highest reliability. Although the High condition had higher factor loadings, the Wide condition achieved more accurate recovery of the latent trait scores, as it allowed larger within-pair loading differences.

Figure 2

Estimation reliability of the latent trait scores under different conditions in Simulation 2: a. with a correlation of zero between the SDIs and the factor loadings; b. with a correlation of 0.3 between the SDIs and the factor loadings; and c. with a correlation of 0.6 between the SDIs and the factor loadings



(a)



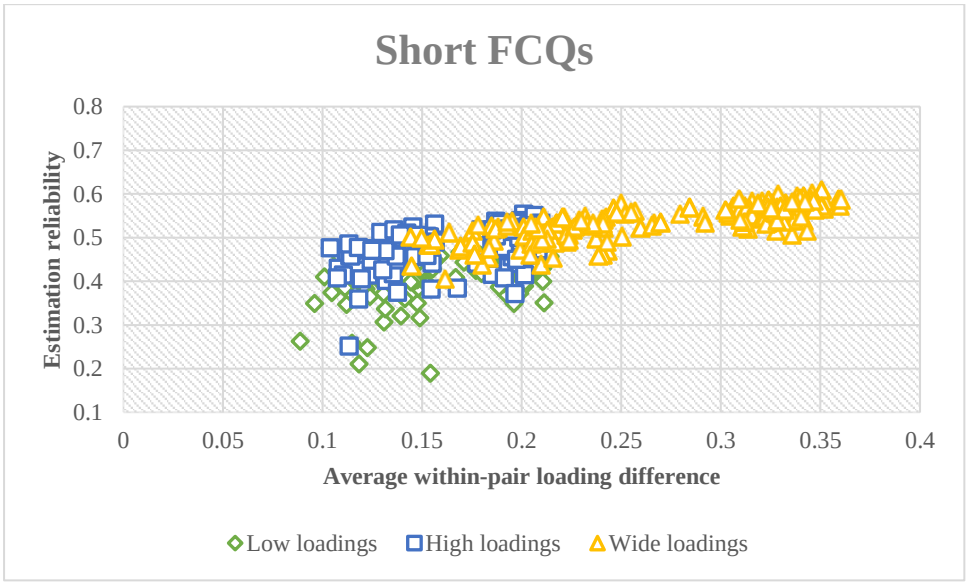
(b)

Figure 3, which depicts the simulation results of individual trials, clearly demonstrates the interplay of the absolute values of the factor loadings and the within-pair loading differences on the estimation reliability of the latent trait scores. Due to different pairing strategies, two clusters emerged within each factor loading condition. It seemed that with a larger within-pair loading difference, the Thurstonian IRT model

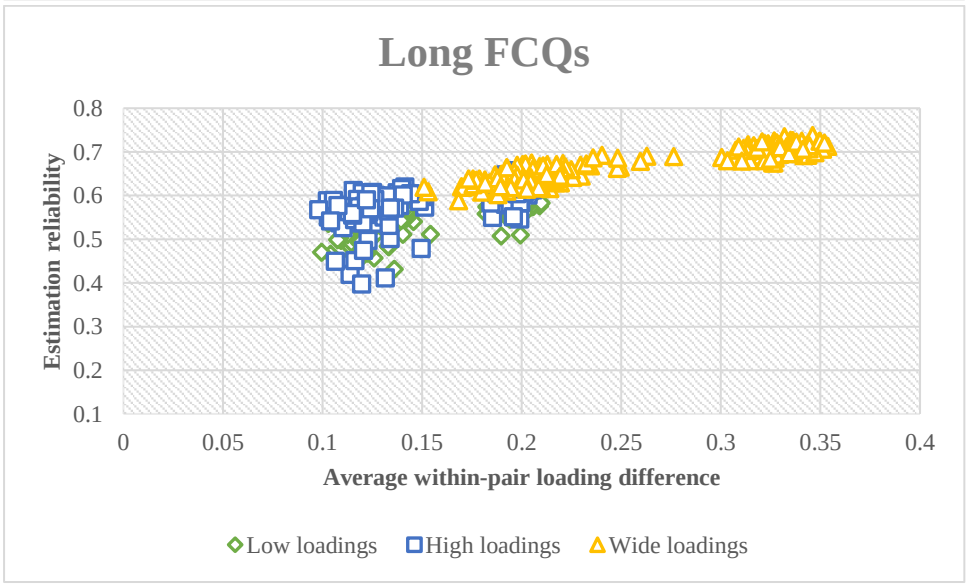
recovered the latent trait scores more accurately. When the within-pair loading difference was small, those with high factor loadings tended to yield greater reliability.

Figure 3

Estimation reliability of the latent trait scores for all trials in Simulation 2: a. in short FCQs with 30 item pairs each; and b. in long FCQs with 60 items pairs each



(a)



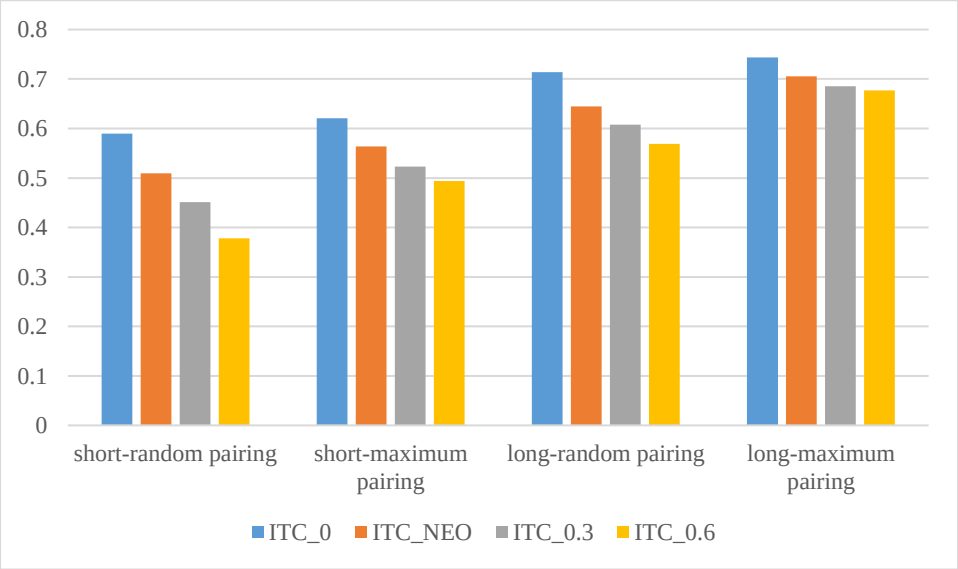
(b)

Similar to Simulation 1, we observed a slight decrease in the reliability when the correlation between the factor loadings and the SDIs was increased. Our hypothesis that this effect was mediated by the within-pair loading difference can now be confirmed by the findings above.

Lastly, we present the results with regard to the inter-trait correlation, taking the conditions with Wide loadings and zero correlation between the factor loadings and the SDIs as an example. Consistent with the conditions in Simulation 1, where only equally keyed item pairs were used, the higher the average correlation among the latent traits, the lower the estimation reliability. As shown in Figure 4, the effect seemed to weaken when joined with other factors (e.g., the number of item pairs). Notably, the reliability was elevated to around 0.7, when Maximum pairing was employed to construct long FCQs with items of Wide loadings. This was a remarkable improvement over the conditions with equally keyed item pairs only in Simulation 1.

Figure 4

Estimation reliability of the latent trait scores in conditions with Wide loadings and a correlation of zero between the factor loadings and the SDIs



Note. ITC: inter-trait correlation; short: FCQs with 30 item pairs; long: FCQs with 60 item pairs.

Empirical Study

The purpose of the empirical study is two-fold. First, we demonstrate how to use a given item bank to assemble forced-choice blocks that are matched on social desirability for a high-stakes assessment. Second, based on empirical results, we attempt to ascertain

the effects of factor loadings, within-block loading differences, and the inclusion of unequally keyed item pairs on the modelling of an FCQ with the Thurstonian IRT model. Below, we describe the procedure of test development and data collection. Then, we report the data analysis results using the Thurstonian IRT model, particularly the findings related to different item pairing approaches.

Developing the Forced-choice Personality Measures

We took a Chinese version of the 240-item NEO-PI-R (Dai, Yao, & Cai, 2004) as the initial item bank. The items were rated on their social desirability by 54 undergraduate students from Beijing (26 male and 28 female, ages ranging from 19 to 24, mean age 22 years). They were asked to rate how desirable they found each item on a 7-point Likert scale from "extremely undesirable" to "extremely desirable". The average rating was taken as the SDI for each item. The SD in the ratings ranged between 0.688 and 1.542.[5]

In order to calibrate the items, we recruited 741 participants online (275 male and 466 female, ages ranging from 13 to 62 with one missing, mean age 25 years), who were asked to indicate how much they agreed with the statements on a 5-point rating scale from "Strongly disagree" to "Strongly agree". They were told that detailed feedback on their personality profiles would be provided, which served as an incentive for honest responses. To examine the construct validity of the personality measure, confirmatory factor analysis was performed, where 20 positive items and 12[6] negative

[5] Unfortunately, the original file that contained all the rating data was missing. The SDs reported here were calculated based on the ratings of 43 participants.

[6] Due to a labelling mistake, one extra negative item was selected for the Emotional Stability domain.

items (the same ratio as in the initial item bank) with the highest factor loadings (in absolute values) were selected for each domain. Among these items, the lowest factor loadings ranged from 0.19 in Agreeableness and Openness to 0.30 in Extraversion. Two items were removed due to labelling mistakes, resulting in a refined item bank of 158 items (98 positive and 60 negative items). All items except three had factor loadings above .2 (or below -.2), and on average, the absolute factor loading was .46 (ranging from .19 to .78).

In the refined item bank, the average SDI was 4.43 (ranging from 2.60 to 5.86). The average SDI of the negative items was 3.33, whereas that of the positive items was 5.10.

It is worth mentioning that the maximum SDI among the negative items was 4.51, and the minimum SDI among the positive items was 3.63. This is consistent with Lee et al. (2018), who reported an overlap in the social desirability between negative and positive personality items.

Notably, a correlation of .36 was found between the factor loadings and the SDIs in the negative items and a correlation of .53 in the positive items. This finding provides empirical evidence for our hypothesis of the correlation between the SDIs and the factor loadings in the simulations.

Based on the 158 items in this refined item bank, three FCQs were created, each consisting of 60 multidimensional item pairs balanced across the five personality traits. All item pairs were matched on social desirability, as the difference in the SDIs between

the items in each pair was constrained to not exceed 0.42.[7]

The three FCQs differed in the approach to the construction of the item pairs. The first FCQ (referred to as "Mixed") included nine unequally keyed item pairs (15% of the total pairs), which were built from the negative and positive items with overlapping SDIs. The remaining 51 equally keyed item pairs were constructed following Maximum pairing. Under Maximum pairing, a loading difference above 0.1 was prespecified between the two items in each pair. One hundred test forms were generated, and the one with the highest average loading difference was chosen as the final one. Similar to the conditions in Simulation 2, the second FCQ (referred to as "Maximum") consisted of equally keyed item pairs only, and the items were confined to those with absolute factor loadings above 0.3.[8] All 60 item pairs were built following the Maximum pairing procedure as described above. As only equally keyed item pairs were included in the Maximum FCQ, its average loading difference was significantly lower than that of the Mixed FCQ (Maximum: 0.18, Mixed: 0.27, $t(71) = -3.65$, $p < .001$). The third FCQ (referred to as "Minimum") used exactly the same set of items as in the Maximum FCQ (to ensure comparable factor loadings), but the items were paired up randomly. Similarly, 100 test forms were generated, and the one with the lowest average loading difference was chosen (to maximise the difference in the loading differences between the two FCQs). The average loading difference of the Minimum

[7] The decision for this arbitrary value was partially informed by the conditions in the simulation studies, where the average within-pair difference in the SDIs on a 5-point scale was mostly below 0.3.

[8] One Openness item with factor loading of 0.29 was included, as there were not enough Openness items with absolute factor loadings above 0.3.

FCQ was 0.10, significantly lower than that of the Maximum FCQ ($t(109) = -6.96$, $p < .001$).

Among the three FCQs, there was no significant difference in the within-pair SDI difference (Mixed: 0.15; Maximum: 0.16; Minimum: 0.14; $F(2, 177) = 1.22$, $p = .30$). However, the three pairing approaches effectively resulted in a significant difference in the within-pair loading difference ($F(2, 177) = 33.37$, $p < .001$). It is also noted that the factor loadings (in absolute values) in the Mixed FCQ were significantly lower than those in the other two (Mixed: 0.45, Maximum/Minimum: 0.48, $t(233) = -2.04$, $p < .05$). Fourteen items with absolute factor loadings below 0.3 were used in the Mixed FCQ, whereas only one item with absolute factor loading below 0.3 was included in the other two. Additionally, due to the correlation between the SDIs and the factor loadings, the unequally keyed item pairs relied heavily on items with relatively low factor loadings (in absolute values), as the average factor loading (in absolute value) of the unequally keyed pairs was as low as 0.33.

Data collection

The FCQs together with a 140-item Likert-type personality questionnaire were embedded in the admission process for a second bachelor's degree at the Beijing Normal University. Participants were informed that their personality scores would be considered in the admission decision-making. Hence, this was presented as a high-stakes assessment. This study received ethical approval from the Institutional Review Board, Faculty of Psychology, Beijing Normal University (Reference numbers: 202012140058 and 202206030087), where the requirement for participants' consent

was waived.

Two samples took part in the assessment. Age and gender information was not collected due to the nature of the assessment. Sample 1 consisted of a total of 622 undergraduate students. They were allowed 35 minutes to complete the Maximum and the Minimum FCQs (the item pairs were mixed together randomly), as well as a Likert-type personality questionnaire. For the forced-choice measure, participants were told to select one of the two statements in each pair that most represented how they perceived themselves. To minimise the effect of the time constraint on the responses, participants with incomplete responses were excluded from the analysis (n = 211). One other participant was removed due to careless responses (straight-lining more than 100 questions). This resulted in a sample size of 410.

Sample 2 consisted of 654 undergraduate students, who were asked to complete the Mixed FCQ and the same Likert-type personality questionnaire within 35 minutes. After removing incomplete responses, 602 remained in the analysis.

Results

The Thurstonian IRT model was fitted to the response data of each FCQ following exactly the same procedure as reported in the simulation studies. Model fit was assessed via the chi-square test, the RMSEA and its 90% confidence interval, and the SRMR. As shown in Table 6, the RMSEAs for all three models were below .05, which has been suggested as the cut-off for a good fit (MacCallum, Browne, & Sugawara, 1996). The SRMRs were all slightly higher than .08, indicating a marginally acceptable model fit (Hu & Bentler, 1999). Although the Mixed FCQ appeared to show the poorest model

fit, the overall fit did not differ much across the three models and was within the acceptable range. We subsequently examined the modification indices (MIs) using a cut-off of 10 and found seven instances of cross-loadings for the Maximum model (average MI: 12.85), eight instances for the Minimum model (average MI: 18.34), and 29 instances for the Mixed model (average MI: 19.49), suggesting that the specification of certain items particularly in the Mixed model might be problematic. This finding is further confirmed by the results below.

Table 6

Goodness of fit indices of the Thurstonian IRT models fitted to the response data of the three FCQs in the empirical study

Model	Chi-square	Degrees of freedom	*p*-value	RMSEA	90% C.I.	SRMR
Maximum	1975.420	1640	0.000	0.022	0.018, 0.026	0.084
Minimum	2034.939	1640	0.000	0.024	0.021, 0.028	0.087
Mixed	2532.129	1640	0.000	0.030	0.028, 0.032	0.087

Note. RMSEA = The Root Mean Square Error of Approximation; C.I = Confidence Interval; SRMR = The Standardised Root Mean Square Residual.

Among the 120 factor loadings that were estimated in each model, 32 were not significant in the Minimum FCQ, 30 were not significant in the Mixed FCQ, and 25 were not significant in the Maximum FCQ. Focusing on the significant estimates of the factor loadings, we calculated their correlations with the factor loadings that were originally used when constructing the item pairs. The correlation coefficient was .89 in the Minimum FCQ, .85 in the Mixed FCQ, and .93 in the Maximum FCQ. It seems that the Maximum FCQ resulted in the best recovery of the factor loadings among the three models.

Table 7 shows the inter-trait correlation matrices for the three models. In the Maximum FCQ, all traits were positively correlated with each other. In the models of Minimum FCQ and Mixed FCQ, one of the traits was negatively correlated with the

others. Notably, in both models, factor loadings for more than half of the items in this problematic factor were not significant. It is likely that this factor was not properly captured by the Thurstonian IRT model.

Table 7

Inter-trait correlations estimated by the Thurstonian IRT models fitted to the response data of the three FCQs in the empirical study

Model	Correlation	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
Mixed	Openness to Experience	-0.151	-0.890***	-0.413***	-0.300**
	Conscientiousness		-0.031	0.532***	0.815***
	Extraversion			0.088	0.215*
	Agreeableness				0.513***
Maximum	Openness to Experience	0.602**	0.875***	0.890***	0.583**
	Conscientiousness		0.848***	0.802***	0.850***
	Extraversion			0.917***	0.830***
	Agreeableness				0.777***
Minimum	Openness to Experience	0.330**	0.267*	-0.523***	0.334**
	Conscientiousness		0.428***	-0.925**	0.822***
	Extraversion			-0.247	0.524***
	Agreeableness				-0.724***

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

We further calculated the empirical reliability (Lin, 2022; Maydeu-Olivares & Brown, 2010) of the factor scores for the three models (see Table 8). Most factors achieved above .7 reliability. The difference in the mean reliability among the three FCQs was very small.

Table 8

Reliability of the factor scores estimated by the Thurstonian IRT models fitted to the response data of the three FCQs in the empirical study

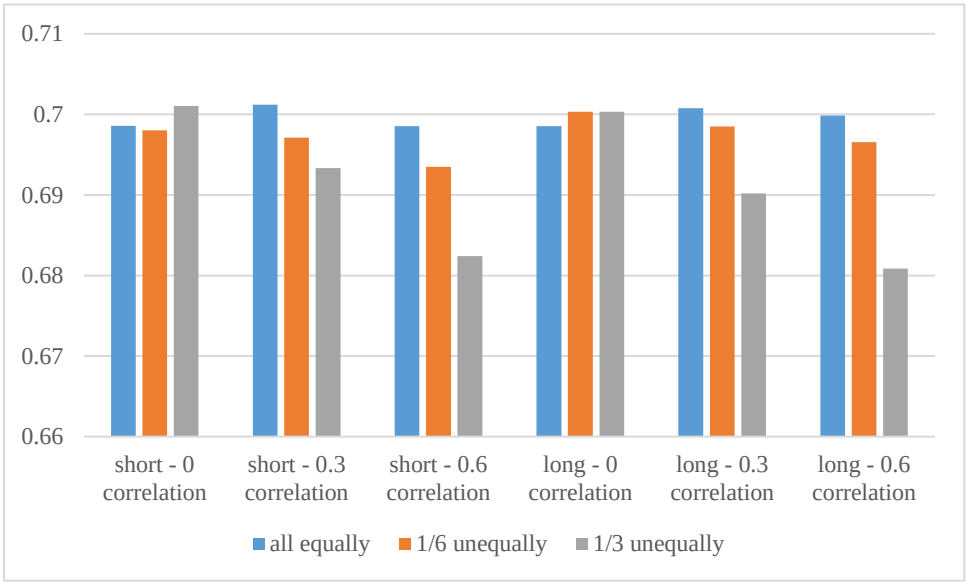
Model	Openness to Experience	Conscientiousness	Extraversion	Agreeableness	Emotional Stability	Mean
Maximum	0.68	0.75	0.75	0.74	0.75	0.73
Minimum	0.68	0.77	0.66	0.75	0.70	0.71
Mixed	0.71	0.76	0.71	0.69	0.74	0.72

Taken together, the results suggest that the Maximum paring outperformed the other two approaches in the current item bank. Compared with the Minimum FCQ, the Maximum FCQ had significantly higher within-pair loading differences. Since these two FCQs included exactly the same set of items, it is evident that the greater within-pair loading differences led to an improvement in the model fit, recovery of the factor loadings, inter-trait correlation structure, and score reliability of the Thurstonian IRT model.

In the meanwhile, although the Mixed FCQ had an even higher within-pair loading difference, the Thurstonian IRT model showed a poorer performance than that fitted to the Maximum FCQ. We believe that this finding was mostly driven by the lowered factor loadings in the Mixed FCQ, which included a total of 14 items with absolute factor loadings below 0.3. Due to the correlation between the SDIs and the factor loadings, only items with relatively low factor loadings could be selected to form unequally keyed pairs: 16 out of the 18 items in the unequally keyed pairs had absolute factor loadings below 0.4. This is consistent with the conditions in Simulation 1. As shown in Figure 5, when unequally keyed pairs were included into FCQs with a greater positive correlation between the SDIs and the factor loadings, the average factor loadings (in absolute values) were decreased. Unlike in the simulations where the absolute factor loadings were generally high (not below .45), the items in the empirical study were drawn from a lower range of factor loadings. It seems that the impact caused by the low factor loadings outweighed the benefits brought by the large within-pair loading differences in the Mixed FCQ.

Figure 5

Average factor loadings (in absolute values) of the FCQs in Simulation 1



Note. Short: FCQs with 30 item pairs; long: FCQs with 60 item pairs; 0/0.3/0.6 correlation: correlations of 0/0/3/0/6 between the SDIs and the factor loadings.

One might argue that the poor performance of the Mixed FCQ could be a result of lower robustness to SDR in the unequally keyed pairs, especially considering that the FCQ was placed in a high-stakes assessment. To rule out such a possibility, we collected another sample of 31 students who were comparable to the students that participated in the high-stakes assessment in the empirical study. They were asked to rate the relative difference in the social desirability between the two items in each item pair. As shown in Table 9, on average, more than 80% of the students detected slight or no difference in the social desirability between each pair of items. Notably, the percentage of perceiving slight or no difference within the unequally keyed pairs was actually higher than that within the equally keyed pairs. Hence, the unequally keyed pairs were no different than the equally keyed pairs in terms of the social desirability matching. The results also confirmed that our construction of the item pairs was effective at matching items on their social desirability.

Table 9

Distribution of ratings of the relative social desirability between each pair of items in the three FCQs (N = 31)

FCQ	Number of pairs	no difference	slight difference	significant difference
Maximum	60	30.8%	52.6%	16.6%
Minimum	60	30.9%	53.0%	16.1%
Mixed	60	32.4%	52.5%	15.1%
- equally key pairs	9	32.4%	52.0%	15.6%
- unequally keyed pairs	51	33.3%	54.2%	12.5%

Discussion

Previous literature has suggested that the forced-choice format eliminates uniform response bias (Brown & Maydeu-Olivares, 2018), while empirical findings on its effectiveness in reducing non-uniform bias, such as SDR have not been consistent. Nonetheless, there seems to be an agreement that it is essential for items in the forced-choice blocks to be matched on their social desirability (e.g., Brown et al., 2017; Cao & Drasgow, 2019). This poses a modelling challenge, as problems such as non-convergence and low score reliabilities have been reported when the Thurstonian IRT model is applied to FCQs exclusively containing equally keyed blocks (Bürkner et al., 2019; Wang et al., 2017). In this article, we provide two feasible solutions. First, we argue that item blocks can be matched on social desirability and, at the same time, unequally keyed. Our results of Simulation 1 indicate that having up to 1/3 item blocks being unequally keyed brings substantial improvement in the item parameter estimation accuracy and the trait score reliability under the Thurstonian IRT model. Second, we suggest that when assembling forced-choice blocks, the within-block loading differences should be maximised, and items with high factor loadings be prioritised. As demonstrated in Simulation 2, these two strategies prove beneficial to FCQs containing

only equally keyed blocks when scored by the Thurstonian IRT model.

Informed by the simulation results, an empirical study has been carried out, where three FCQs, using different item pairing approaches were built from an existing item bank, and applied to a high-stakes assessment. The results showed that the Maximum FCQ outperformed the other two. The Minimum FCQ, which used the same set of items as the Maximum FCQ, revealed the lowest average within-pair loading difference. Although the Mixed FCQ achieved the highest average within-pair loading difference, due to the inclusion of several items with absolute factor loadings below 0.3, it exhibited the lowest average factor loading (in absolute value). A clear message can be taken from the empirical study: when constructing an FCQ using the Thurstonian IRT model, it is preferable to maximise the within-block loading differences and prioritise items with high factor loadings. There is a need for caution when using items with extremely low factor loadings. Although they are often needed for the unequally keyed item pairs, they might undermine the modelling performance.

SDI and its Correlation with the Factor Loading

Both the simulations and the empirical study were designed to address the practical concerns about the SDR. Specifically, a new parameter, SDI, was introduced in the simulations to ensure matched social desirability in each block. It was also directly measured in the empirical study and guided the construction of the FCQs. Although the SDI is not explicitly included in the Thurstonian IRT model, it indirectly influences the modelling performance through its correlation with the factor loadings.

The assumption of a positive correlation between the SDIs and the factor

loadings originates from our observations through the daily practice of test development. It has also been implied in existing literature. For instance, studies on the Minnesota Multiphasic Personality Inventory (Edwards, 1957; Edwards & Edward, 1991) showed that the first principal component which influenced a majority of the scales was strongly related to social desirability. Bäckström (2007) also found a higher-order factor in a five-factor personality inventory that was correlated with self-deception and impression management, the two components of SDR. It is speculated that when the social desirability factor is not modelled in the factor structure, its variation is manifested through the correlation among content factors. Consequently, the factor loadings on the content factors are influenced by the hidden social desirability factor, resulting in a correlation between the factor loadings and the social desirability levels of the items. This assumption is confirmed by the observations in Lee et al. (2018) as well as in our empirical study, where a moderate correlation between the SDIs and the factor loadings was found in the Big Five personality items.

In practice, the correlation between the SDIs and the factor loadings creates two challenges for the FCQ construction. First, when building equally keyed item blocks, we struggle to find items with factor loadings that are wide apart; this is because when items are matched on social desirability, the correlation between the SDIs and the factor loadings to an extent limits the magnitude of the loading difference. Second, when building unequally keyed item blocks, it is more likely for items of low factor loadings to be matched on social desirability, undermining the modelling performance. Given these challenges, it is important for practitioners to carefully evaluate the items

available in the item bank before assembling the forced-choice blocks.

The Inclusion of Unequally Key Item Blocks in an FCQ

Consistent with previous studies (e.g., Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019), our simulations found a significant advantage in including unequally keyed item blocks in an FCQ. Instead of having half of the blocks being unequally keyed, with up to 1/3 of blocks being unequally keyed, almost all trials successfully converged. More importantly, we observed in the long FCQs remarkably high reliability (around 0.8) for the latent trait score estimation. This suggests that embedding a relatively small number of unequally keyed blocks in an FCQ can lead to a substantial improvement in the estimation reliability.

Nonetheless, in practice, we should remain cautious if we decide to include unequally keyed item blocks in an FCQ. Two questions could be raised to help with the decision-making. First, is it possible to assemble sufficient unequally keyed blocks in a given item bank? Several empirical studies have proved it feasible (e.g., Ng et al., 2020; Wetzel & Frick, 2020). In our simulations with twice the required number of items in the item bank, we failed at 20% of the trials where 1/3 of the blocks were unequally keyed. In the empirical study, we were able to construct nine unequally keyed pairs out of 158 items (with 60 negatively keyed items), which was 15% of the item pairs in the FCQ. Second, is there any potential disadvantage in including unequally keyed item blocks? In our empirical study, surprisingly, the Mixed FCQ turned out not to be the best-performing one. Although it had the largest within-pair loading differences, it suffered from a low average factor loading. Some items with very low

factor loadings were included, many of which were used to construct the unequally keyed pairs, resulting in a poor fit of the Thurstonian IRT model. This contrasts the results in the simulations, where almost perfect model fit was observed. The inclusion of items with low factor loadings might explain the discrepancy in the model fit between the simulations and the empirical study. The results also highlight the necessity of examining the feasibility of applying the simulation findings to the real world.

In an applied setting, it is common to prepare a large number of items when constructing an FCQ. With more items available, it is more likely to identify positive and negative items with comparable social desirability. These items are critical, as they allow the assembly of unequally keyed blocks without sacrificing the faking resistance. In the meantime, their factor loadings should be sufficiently high so as not to influence the modelling performance. Furthermore, it is recommended to include more neutrally framed items in the item bank, which are low on the social desirability scale (Bäckström, Björklund & Larsson, 2009; Heggestad et al., 2006).

Limitations and Future Research

Some limitations in our studies have been noticed. First, in the simulations, we fixed the number of traits to five based on the prototype of the Big Five personality traits. However, this is not always the case in practice. For instance, many personality tests used in occupational assessment measure more constructs (e.g., SHL, 2013). More efforts are warranted to investigate the FCQs that measure more than five traits. Second, we realised that the factor loadings in the simulations were slightly too high, although they were similar to the ranges reported in previous studies (e.g., Brown & Maydeu-

Olivares, 2011; Bürkner et al., 2019). These values were chosen in order to better demonstrate the effect of Maximum pairing. Further simulations using factor loadings comparable to those observed in the empirical study are needed to replicate the findings reported here. Third, we made the decision to use item pairs instead of item triplets in our simulations. Although we do not expect any major difference in the implications, it would be good if future simulations could explore forced-choice blocks of other sizes.

One limitation of the empirical study is that the time allowed for the Maximum and Random FCQs was rather limited. Around a third of the participants could not finish the test in time, which significantly reduced the sample size. Another limitation is that no data was collected under an honest condition. Without a comparison, we were unable to examine the accuracy in the recovery of the trait scores or demonstrate the empirical cost of different pairing approaches. Moreover, we realised that, although items were matched on their SDIs, slight differences in the social desirability were still detectable within the item pairs, and individuals might perceive the relative differences to a varying extent. As a direction for future research, the SDI could be potentially factored into the response process model. This will enable more meaningful simulations and allow a better understanding of how the relative social desirability among the items affects individual's responses to the forced-choice blocks. Another mitigation could be the use of the inter-item agreement index, as suggested by Pavlov, Shi, Maydeu-Olivares, and Fairchild (2021) when assembling the forced-choice blocks.

An anonymous reviewer suggested that automated test assembly algorithms should be considered for the optimal assembly of forced-choice blocks (e.g., Kreitchmann,

Abad, & Sorrel, 2022). For instance, mixed integer programming can provide an optimal solution that not only avoids the dependence on ordering that occurs when item blocks are assembled sequentially, but also optimises both within-block loading differences and average loadings together (e.g., Bürkner, 2022). We will explore these advanced methods in the future and particularly examine if our simpler methods perform similarly in practice. Meanwhile, we have made our Maximum pairing algorithm available online (see Supplementary Material) to help practitioners construct their forced-choice questionnaires.

Conclusion

The Thurstonian IRT model has greatly facilitated the use of the forced-choice format in research and practice. Although previous simulation-based studies have provided general guidelines for the FCQ construction (e.g., Brown & Maydeu-Olivares, 2011; Bürkner, 2022), we place particular emphasis on the many requirements often encountered in practice. One of them is to ensure that the resulting FCQ is faking resistant. This article strives to contribute to the practice of developing faking resistant FCQs under the Thurstonian IRT model. Based on a comprehensive review of existing simulations and empirical findings, we provide the suggestions below to the practitioners as guidance for the FCQ construction. As a starting point, we acknowledge that the practitioners often have limited flexibility. For instance, when they are tasked to develop an FCQ to measure certain traits, the number of traits and the inter-trait correlations are usually fixed.

1. A greater number of blocks should be used to increase measurement precision. Nonetheless, this depends on the availability of proper items, which we will discuss in detail below. Practical factors such as cognitive load and fatigue effect should also be considered when determining the format and the length of an FCQ to be created.
2. When a given item bank is available, it should be calibrated based on a sample comparable to the target respondents. It is recommended to remove items with low factor loadings (in absolute values). Additionally, a measure of desirability of the items as viewed in the context of the assessment should be collected.
3. When assembling forced-choice blocks, items within each block should be matched on their desirability. Furthermore, the within-block loading difference should be maximised, particularly in the case of equally keyed blocks. Based on our experience, and as suggested by the simulations of Bürkner (2022), a loading difference of 0.2 should be aimed for.
4. It is recommended to include some unequally keyed blocks (up to a third of the total number of forced-choice blocks if half is not possible) in the FCQ.
5. Holding the above conditions constant, it is optimal to prioritise items of high factor loadings. Items with absolute standardised factor loadings below 0.3 should be avoided.
6. These recommendations are easier to meet, the larger the initial item bank, from which only the most suitable items are selected. We recommend having

at least 50% more items in the initial item bank than items intended to include in the FCQ.

References

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*(1), 83-97. https://doi.org/10.1177/0734282914550387

Bäckström, M. (2007). Higher-order factors in a five-factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment*, *23*(2), 63-70. https://doi.org/10.1027/1015-5759.23.2.63

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335-344. https://doi.org/10.1016/j.jrp.2008.12.013

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

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

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

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

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

Bolt, D. M., & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale Measurement of Extreme Response Style. *Educational and Psychological Measurement*, *71*(5), 814–833. https://doi.org/10.1177/0013164410388411

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., & 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. (2018). Modelling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.) *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (pp. 523–569)*. Hoboken: Wiley. https://doi.org/10.1002/9781118489772.ch18

Bürkner, P. C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662. https://doi.org/10.21105/joss.01662

Bürkner, P. C. (2022). On the Information Obtainable from Comparative Judgments. *Psychometrika*, 1-34. https://doi.org/10.1007/s11336-022-09843-z

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827-854. https://doi.org/10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347-1368. https://doi.org/10.1037/apl0000414

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*(1), 55–77. https://doi.org/10.1207/S15328007SEM0901_4

Christiansen, N., Burns, G., & Montgomery, G. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance, 18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

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

Dai, X., Yao, S., & Cai, T. (2004). Reliability and Validity of the NEO-PI-R in Mainland China. *Chinese Mental Health Journal, 18*(3), 171–174.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden Press.

Edwards, L. K., & Edwards, A. L. (1991). A principal-components analysis of the Minnesota Multiphasic Personality Inventory factor scales. *Journal of Personality and Social Psychology*, *60*(5), 766-772. https://doi.org/10.1037/0022-3514.60.5.766

Ferrando, P. J. (2005). Factor Analytic Procedures for Assessing Social Desirability in Binary Items. *Multivariate Behavioral Research,* 40(3), 331-349. https://doi.org/10.1207/s15327906mbr4003_3

Fisher, P. A., Robie, C., Christiansen, N. D., Speer, A. B., & Schneider, L (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus Classical Test Theory scoring. *Personnel Assessment and Decisions, 5*(1), 49-61. https://doi.org/10.25035/pad.2019.01.003

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with

ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625–641. https://doi.org/10.1080/10705510903203573

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11*, 340–344. https://doi.org/10.1111/j.0965-075X.2003.00256.x

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Griffith, R. L., Chmielowski, T., Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, 36(3), 341–355. https://doi.org/10.1108/00483480710731310

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New Perspectives on Faking in Personality Assessment* (Vol. 1, pp.34-52). New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jackson, D., Wroblewski, V., & Ashton, M. (2000). The impact of faking on employment tests: does forced choice offer a solution? *Human Performance, 13*(4), 371–388. https://doi.org/10.1207/S15327043HUP1304_3

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*(1), 92-97. https://doi.org/10.1111/ijsa.12098

Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2021). A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods, 54*, 1476-1492. https://doi.org/10.3758/s13428-021-01677-4

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*(1), 229-235.

https://doi.org/10.1016/j.paid.2017.11.031

Likert, R. (1932). A technique for the measurement of attitudes, *Archives of Psychology*, 14, 1-55.

Lin, Y. (2022). Reliability estimates for IRT-based forced-choice assessment scores. Organizational Research Methods, *25*(3), 575-590. https://doi.org/10.1177/1094428121999086

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences*, *32*(2), 247-256. https://doi.org/10.1016/S0191-8869(01)00021-6

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

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

McCrae, R. R., & Costa, P. T. (1992). Discriminant validity of NEO-PIR facet scales. *Educational and Psychological Measurement, 52*(1), 229-237. https://doi.org/10.1177/001316449205200128

Meade, A. (2004). Psychometric problems and issues involved with creating and using

ipsative measures for selection. *Journal of Occupational and Organisational Psychology, 77*(4)*,* 531–552. https://doi.org/10.1348/0963179042596504

Merk, J., Schlotz, W., & Falter, T. (2017). The motivational value systems questionnaire (MVSQ): psychometric analysis using a forced choice Thurstonian IRT model. *Frontiers in Psychology*, *8*, 1626. https://doi.org/10.3389/fpsyg.2017.01626

Moors, G. (2009). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research*, *22*(1), 93-119. http://doi.org/10.1093/ijpor/edp036

Muthén, L.K., .& Muthén, B.O. (1998–2019). *Mplus 8.3 [computer software]*. Los Angeles: Authors.

Neill, J. A., & Jackson, D. N. (1970). An Evaluation of Item Selection Strategies in Personality Scale Construction. *Educational and Psychological Measurement*, *30*(3), 647–661. https://doi.org/10.1177/001316447003000312

Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment*, *103*(2), 224-237. https://doi.org/10.1080/00223891.2020.1739056

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R.

Shaver and L.S. Wrightsman (Eds), *Measures of Personality and Social Psychology Attitudes* (pp. 17-59). New York: Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410607454-10

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 224–239). New York: Guilford.

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of Applicant Faking on Forced-Choice and Likert Scores. *Organizational Research Methods*, *22*(3), 710–739. https://doi.org/10.1177/1094428117753683

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

SHL. (2013). *OPQ32r technical manual version 1.0*. Thames Ditton, UK: SHL Group.

Sun, L. (2022). Full results of the simulation studies. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.19222407.v1

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195-217.

http://doi.org/10.1093/ijpor/eds021

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*(3), 175–199. https://doi.org/10.1207/s15327043hup1903_1

Viswesvaran, C., Deller, J., & Ones, D.S. (2007). Personality measures in personnel selection: Some new contributions. *International Journal of Selection and Measurement, 15*(3), 354-358. https://doi.org/10.1111/j.1468-2389.2007.00394.x

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment, 27*(4), 706-718. https://doi.org/10.1177/1073191119843585

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*(8), 600-613. https://doi.org/10.1177/0146621617703183

Wetzel, E., Böhnke, J. R., Brown, A. (2016). Response biases. In Leong, F. T. L., Bartram, D., Cheung, F., Geisinger, K. F., Iliescu, D. (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 349–363). New York, NY: Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0024

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