Personality and Individual Differences 191 (2022) 111555

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods

Philseok Lee [a,*], Seang-Hwane Joo [b], Steven Zhou [a], Mina Son [a]

[a] Department of Psychology, George Mason University, United States of America
[b] Department of Educational Psychology, The University of Kansas, United States of America

ARTICLE INFO

Keywords:
Multidimensional forced-choice measure
Negatively keyed statements
Monte Carlo simulation
Partially ipsative scoring
Thurstonian Item Response Model
Personnel selection

ABSTRACT

Multidimensional forced choice (MFC) personality tests have recently come to light as important personnel assessments in industrial and organizational psychology. For developing MFC measures, researchers have recommended including heteropolar blocks (i.e., both negatively and positively keyed statements are mixed within a block) to improve the scoring estimation accuracy. However, very few studies have explored the impact of heteropolar blocks on psychometric properties within the MFC context. In this study, we 1) explored how heteropolar blocks influence reliability and validity of MFC tests through Monte Carlo simulations and 2) empirically demonstrated how MFC test designs associated with heteropolar blocks affect criterion-related validity using real examinees. Result shows the Thurstonian Item Response Theory (TIRT) scoring method and higher intrablock discrimination yielded better reliability and criterion-related validity. In addition, result suggests that one can achieve sufficient reliability (i.e., 0.87–0.90 on average) and validity (i.e., 0.40–0.45 on average) by using highly discriminating 20–40% heteropolar blocks with TIRT model. Our empirical demonstration showed that criterion-related validity results can be different depending on the test designs of heteropolar blocks. Practical implications and future research topics are discussed.

Historically, Likert-type scales have been widely used for measuring noncognitive constructs such as personality, attitude, and vocational interest in industrial and organizational psychology. Likert-type scales have commonly used a mix of positively and negatively keyed statements, and many studies have investigated the effect of negatively keyed statements (DeVellis, 2016; McLarnon et al., 2016; Merritt, 2012; Sliter & Zickar, 2014; Weijerts et al., 2013). Some advocates have suggested that the use of negatively keyed statements helps in reducing careless responses, because differently-keyed directional statements ask respondents to pay closer attention to the content (DeVellis, 2016; Weijerts et al., 2013). In contrast, other researchers have suggested that the negatively keyed statements adversely affect item discrimination (Sliter & Zickar, 2014), introduce extraneous variance (e.g., method effect) in responses (Biderman et al., 2011; DiStefano & Motl, 2006), and have a significant association with social desirability in high-stakes settings (McLarnon et al., 2016).

However, previous studies that investigated negatively keyed statements exclusively focused on Likert-type scales, and there has been little to no research in the context of multidimensional forced-choice (MFC) measures. For MFC measures, negatively keyed statements play a key role in establishing normative information in scoring methods, such as partially ipsative (PI) scoring from the classical test theory approach (Lee et al., 2018) or Thurstonian Item Response Theory (TIRT) scoring from the modern psychometric theory approach (Brown & Maydeu-Olivares, 2011). Although some general guidelines regarding MFC measure development are available in the literature (e.g., Brown & Maydeu-Olivares, 2011; Cao & Drasgow, 2019), there are no recommendations for mixing negatively and positively keyed statements (i.e., heteropolar block) in MFC measures. Particularly, little is known about how many negatively keyed statements should be included in MFC measures to ensure sufficient psychometric properties. Furthermore, recent personality research that used MFC measures presented somewhat inconsistent criterion-related validity evidence of MFC measures between PI scoring and TIRT scoring approaches. Given that they used

* Corresponding author at: Department of Psychology, George Mason University, 4400 University Drive, David King Hall Room 3056, Fairfax, VA 22030, United States of America.
E-mail address: plee27@gmu.edu (P. Lee).

https://doi.org/10.1016/j.paid.2022.111555
Received 28 September 2021; Received in revised form 2 December 2021; Accepted 6 February 2022

different MFC test designs associated with negatively keyed statements [e.g., Fisher et al., 2019 used 0% heteropolar block, while Walton et al., 2020 used 100% heteropolar blocks], these inconsistent findings may be due to varied test designs of heteropolar blocks. This issue also should be explored and our research aims to fill these knowledge gaps.

The main goal of this research is three-fold: 1) investigate how negatively keyed statements within a block influence reliability and criterion-related validity between the PI scoring and TIRT scoring methods; 2) empirically demonstrate how criterion-related validity results can be different depending on the negatively keyed statements; and 3) provide an optimal strategy of including negatively keyed statements to achieve the psychometric properties in the TIRT application. We conducted a Monte Carlo simulation study under various conditions including sample sizes, block discriminations, proportions of negatively keyed statements, latent trait correlations, and scoring methods. Furthermore, using real examinees, we demonstrated criterion-related validities of two MFC measures that used different proportions of negatively keyed statements.

1. Likert-type measures and MFC measures

Likert-type scales ask respondents to undertake an evaluation of every single statement, then indicate their level of agreement with a set of options, such as "I strongly disagree (1)" to "I strongly agree (5)". Despite the widespread use of Likert-type scales, they have been greeted with criticism due to their vulnerability to various forms of response biases, including rater errors (e.g., halo, leniency), cultural-specific biases (e.g., central tendency, acquiescence, extremity), as well as faking responses (Borman et al., 2001; Ferrando et al., 2011; He & van de Vijver, 2013; Meade, 2004).

In a bid to address the response biases associated with Likert-type scales, researchers have suggested MFC measures as an alternative (Wetzel & Greiff, 2018). MFC measures present two (i.e., pair), three (i.e., triplet), or four (i.e., tetrad) statements within an item block (hereafter, referred to as "block"). Within blocks, each statement represents different latent traits and is matched based on a similar level of social desirability or item extremity. Then, respondents are forced to select a statement that is "most like me", or rank the statements from "most like me" to "least like me" in each block. By matching social desirability of the statements within a block, it makes it more difficult for respondents to discern the most optimal or desirable answers, thus helping to reduce response biases (Cao & Drasgow, 2019; Wetzel & Greiff, 2018).

1.1. Increasing interests of MFC measures

Particularly, a considerable amount of research has investigated whether the MFC personality measures reduce faking responses compared to the Likert-type personality measures. Although there is some negative evidence on the effectiveness of MFC measures (Heggestad et al., 2006), most research suggests that MFC measures successfully reduce score inflation and maintain validity under motivated test situations (e.g., Bartram, 2007; Bowen et al., 2002; Cao & Drasgow, 2019; Christiansen et al., 2005; Fisher et al., 2018; Hirsh et al., 2008; Jackson et al., 2000; Lee, Joo, & Lee, 2019; Lee, Joo, Stark, & Chernyshenko, 2019; O'Neill et al., 2017). With these merits, use of MFC measures have burgeoned in industrial and organizational psychology over the last decade (e.g., the Occupational Personality Questionnaire [Brown & Bartram, 2009]; the Global Personality Inventory–Adaptive [CEB, 2010]; the Tailored Adaptive Personality Assessment System [Stark et al., 2014]; the Emotional Intelligence Questionnaire [Anguiano-Carrasco et al., 2015]; the Adaptive Employee Personality Test [Adept-15; Aon, 2015]; the Maladaptive Personality Assessment [Guenole et al., 2018]; and CIVIC MFC Assessment [Ng et al., 2021]). Most recently, Robie et al. (2021) surveyed 78 practitioners in industrial and organizational settings and found 78.2% of them most frequently use single-statement Likert-type personality measures and 20.5% most frequently use MFC personality measures. In addition, 41.0% answered they use a combination of Likert-type and MFC personality measures in their organizational settings. These recent trends clearly show an increasing interest in MFC personality measures.

Although different MFC formats are used (e.g., pair, triplet, and tetrad) in research and applied settings, this study focuses on the triplet format, given its increasing interest (e.g., Bartram, 2007; Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Guenole et al., 2018; Lee et al., 2018; Lee, Joo, & Lee, 2019; Lee, Joo, Stark, & Chernyshenko, 2019; Lin & Brown, 2017; Ng et al., 2021). The preference for the triplet format may stem from the fact that it is more informative than the pair format (Joo et al., 2019; Lee, Joo, & Lee, 2019; Lee, Joo, Stark, & Chernyshenko, 2019) and less cognitively demanding than the tetrad format (Brown & Maydeu-Olivares, 2011). An example of the triplet format using rank responses is shown below.

Rank the three statements in each group from "most like me (1)" to "least like me (3)"	Rank
(A) I do things according to a plan.	3
(B) I make friends easily.	1
(C) I enjoy hearing new ideas	2

We note that the three statements within this item block measure conscientiousness (A statement), extraversion (B statement), and openness to experience (C statement).

2. Partially ipsative (PI) scoring methods for MFC measures

2.1. Traditional scoring issues of MFC measures

Historically, MFC measures have not been widely used in high-stakes settings due to the ipsativity problem in that it only allows for within-person comparisons (Hicks, 1970; Johnson et al., 1988; Meade, 2004). For instance, when one simply inverts ranks of statements within a block (e.g., 3-point, 2-point, and 1-point are assigned to 1st, 2nd, and 3rd ranked statements, respectively) and sums inverted-rank for each block for scoring, the sum scores across MFC blocks would always be the same (i.e., 6 points) across individual examinees, making it impossible to distinguish the total score differences between these individuals. This feature of the traditional scoring approach has precluded the widespread use of MFC measures in high-stakes settings where between-person comparisons are essential (Dilchert et al., 2006; Heggestad et al., 2006; McCloy et al., 2005).

2.2. Partially ipsative scoring methods

To alleviate the ipsativity problem associated with the traditional MFC scoring method, researchers have used a heuristic scoring approach, referred to as a *partially ipsative* (*PI*) scoring method (e.g., Heggestad et al., 2006; McCloy et al., 2005; White & Young, 1998). By taking steps to introduce variation in scale scores, PI scores can be obtained, for example, by adding negatively keyed statements in an MFC measure or by including distractor statements that are not scored. This was demonstrated in a method described by Lee et al. (2018). They assigned 2 points when a positively keyed statement was chosen as the first rank or a negatively keyed statement was chosen as the third rank, 0 points when a positively keyed statement was selected as the third rank or a negatively keyed statement was selected as the first rank, and finally 1 point for any statement ranked second. Then, the PI scores were obtained by aggregating recoded scores for each dimension.

The PI scoring method has been shown to be effective in predicting outcomes (Converse et al., 2008; Jackson et al., 2000; Lee et al., 2018; Salgado et al., 2015). Salgado et al. (2015) conducted a meta-analysis investigating the validity of forced-choice personality measures based on traditional scoring approaches. They showed that the PI scoring method yielded higher predictive validities compared to both the fully

ipsative scoring method and the Likert-type summed scoring method across all occupational groups. For example, the average validity of the PI scoring method was 0.22 for openness (vs. 0.07 and 0.05 from fully ipsative scoring method and Likert-type scale, respectively), 0.38 for conscientiousness (vs. 0.14 and 0.22), 0.12 for extraversion (vs. 0.12 and 0.12), 0.16 for agreeableness (vs. 0.03 and 0.08), and 0.20 for emotional stability (vs. 0.07 and 0.11).

Based on the findings, Salgado et al. (2015) suggested that the PI scoring method "can be used for making personnel decisions because their validity is similar to, or even greater than, other well-known procedures (e.g., structured interviews, assessment centers, situational judgment tests)" (p. 820). More recently, Lee et al. (2018) compared three different scoring methods (i.e., PI method, an analogous PI method using the graded item response theory, and the TIRT method) for MFC triplet data, and they found the PI scoring method was as effective as the more complex TIRT-based scoring method for MFC measures. They suggested that the PI scoring method can be "a good alternative for organizations that cannot obtain the appropriate sample size or for which simpler scoring methods are needed for investigating validity evidence" (p. 232).

3. IRT-based scoring approach for MFC measures

Despite the wide use of the PI scoring method in the literature, it has been criticized by modern psychometric researchers claiming that the PI scoring method does not follow a psychometric model that assumes the comparative judgment process for ranking responses (Stark et al., 2012). The PI scoring method transforms MFC ranking data into scale scores by disassembling, recoding, and regrouping the ranking responses; thus, there is no psychometric connection between latent trait scores and the comparative judgment process of evaluating the statements within MFC blocks. In other words, it would be difficult to claim that the PI scores truly capture respondents' latent traits underlying the measured constructs by MFC measures. The recent development of MFC item response theory (IRT) models has addressed the ipsativity problem and made it possible to obtain normative scores by connecting latent variable modeling and the comparative judgment process. Among several MFC IRT models, the TIRT model has been the most widely used in applied research due to the fast and easy implementation via publicly available software, such as the Mplus program (Muthén & Muthén, 1998–2017).

3.1. A brief introduction to TIRT model

In TIRT modeling, rank response data are transformed into multiple binary outcomes. For instance, for a triplet MFC block, if the first statement is preferred to the second statement, the binary outcome is coded as 1, and 0 otherwise. Similarly, if the first statement is preferred to the third statement, the binary outcome is coded as 1, and 0 otherwise. If the second statement is preferred to the third statement, the binary outcome is scored as 1, and 0 otherwise. Therefore, the ranking response [1, 2, 3] is transformed into three binary outcomes, [1], [1], [1]. Then, the transformed binary outcomes are modeled and analyzed with a two-dimensional standard normal ogive IRT model. The conditional probability of selecting statement s to statement t is expressed as follows:

$$P(y = 1|\eta_A, \eta_B) = \Phi\left(\frac{-\gamma_{s,t} + \lambda_s\eta_A - \lambda_t\eta_B}{\sqrt{\psi_s^2 + \psi_t^2}}\right), \quad (1)$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function evaluating at x; $\gamma_{s,\,t} = (\mu_s - \mu_t)$ is the threshold parameter representing the mean difference between statements s and t; η_A and η_B are the latent trait attributes A and B for examinees; λ_s and λ_t are the factor loadings of statements s and t on the latent trait attributes η_A and η_B; and ${\psi_s}^2$ and ${\psi_t}^2$ are the unique variance of the two statements. The TIRT model has been well described in Brown and Maydeu-Olivares (2011).

4. Potential issues surrounding negatively keyed statements in MFC measures

Brown and Maydeu-Olivares (2011) suggested that MFC measures should be created by mixing positively and negatively keyed statements within a block, called a heteropolar block, to improve parameter estimation accuracy in the TIRT model application. Following their suggestion, recent empirical studies developed MFC measures by mixing negatively keyed statements within blocks (e.g., Lee et al., 2018, 2021; Ng et al., 2021; Walton et al., 2020; Wetzel & Frick, 2019). Although the recommendation of the inclusion of negatively keyed statements may bolster the estimation accuracy of MFC measures, researchers have raised concerns about the negatively keyed statements in MFC measures based on practical and psychometric reasons (e.g., Bürkner et al., 2019; Fisher et al., 2019; Lin & Brown, 2017; Ng et al., 2021).

First, the inclusion of many heteropolar blocks can make MFC measures more easily fakable because negatively keyed statements would not be interpreted as desirable as positively keyed statements. Respondents may still be able to fake their rank responses by avoiding a negatively keyed statement and choosing a positively keyed statement as a higher rank (Bürkner et al., 2019; Ng et al., 2021). This phenomenon would be more pronounced in high-stakes settings; for example, Donovan et al. (2003) found that job applicants tend to fake their responses by downplaying negative attributes (e.g., over 60% of applicants) rather than by exaggerating positive attributes (e.g., over 30% of applicants). Consequently, the inclusion of many negatively keyed blocks would result in differential functioning of MFC blocks between high-stakes settings and low-stakes settings (Lee & Joo, 2021).

Second, the inclusion of many heteropolar blocks can make MFC measures more cognitively demanding. The mixing of positively and negatively keyed statements within a block could yield a disruption in the item-response process because respondents have to frequently shift their cognitive processing of the content depending on the keyed directions (e.g., McLarnon & Carswell, 2013; Merritt, 2012; Roszkowski & Soven, 2010). Consequently, it could make comparative judgment more complex in high-stakes settings, because the disruption in the item-response process can interact with situational perceptions (e.g., goal of getting hired or promoted). According to information processing theory, because respondents have limited cognitive resources, those with a higher level of cognitive ability are able to better navigate the complex cognitive processing of item response (Pelled, 1996). Therefore, cognitive/linguistic processing demands inherent in such MFC blocks may lead to item bias and/or adverse impact against a selective population in the personnel selection context.

Third, to date, many studies have reported that positively and negatively keyed statements do not measure the same underlying construct and introduce method effects related to different wording of the statements, which may cause extraneous variance that could lower the construct validity and utility of a pre-employment assessment tool (Biderman et al., 2011; DiStefano & Motl, 2006; Horan et al., 2003). This also applies to MFC measures; if MFC blocks used in high-stakes testing are contaminated by variance attributable to negatively keyed statements, the psychometric properties of MFC measures could be compromised. Bürkner et al. (2019) recently found that including negatively keyed statements can result in methodological variance (e.g., method effect) that influences the covariance structure of MFC blocks. Consequently, the construct validity of MFC measures would be attenuated. Recently, Ng et al. (2021) stated that "including a negatively keyed statement in MFC format item blocks create problems by (a) making the measure more fakable and (b) may contribute in part to lower validity via methodological variance" (p. 12).

Considering these potential issues surrounding negatively keyed statements in MFC measures, it would be beneficial to use a small number of negatively keyed statements to avoid the aforementioned negative impacts. However, this situation creates a dilemma, because there is a trade-off between improvement of scoring accuracy and

adverse impacts involving negatively keyed statements. Unfortunately, there is no clear guideline concerning the use of negatively keyed statements in MFC personality measures for researchers and practitioners. Therefore, our research aims to fill this gap.

5. Study 1: Monte Carlo simulation study

5.1. Simulation design

To investigate the impact of negatively keyed statements in MFC scoring reliability and validity, we conducted a Monte Carlo simulation study. For the simulation study, the following simulation factors were considered.

1. Sample size: Two sample sizes were considered based on previous simulation and empirical studies using the TIRT model: (a) 500 and (b) 1000. Generally, simulation studies have been conducted based on large sample sizes (e.g., 1000 or 2000; Brown & Maydeu-Olivares, 2011, 2012; Morillo et al., 2016), and empirical studies have used relatively smaller sample sizes [e.g., Anguiano-Carrasco et al., 2015, $n = 486$; Lee et al., 2018, $n = 417$; Morillo et al., 2016, $n = 392$; Guenole et al., 2018, $n = 420$; Wetzel & Frick, 2019, $n = 593$; Ng et al., 2021, $n = 798$].
2. Intrablock discrimination: We created low and high discrimination conditions. For the low discrimination condition, factor loadings were randomly drawn from U(0.3, 0.8) and for the high discrimination condition, factor loadings were randomly drawn from U(0.8, 1.3). These two conditions were created based on previous research (Brown & Maydeu-Olivares, 2012; Bürkner et al., 2019; Lee et al., 2021). Bürkner et al. (2019) sampled factor loadings from U(0.3, 0.7) for the small discrimination condition and U(0.65, 0.95) for the large discrimination condition. Brown and Maydeu-Olivares (2012) created factor loadings ranging from 0.8 to 1.3 for the triplet MFC test. In addition, Lee et al. (2021) sampled factor loadings from U (0.8, 1.3) to create high discrimination conditions for the triplet MFC test.
3. Proportion of heteropolar blocks: We varied the proportions of the heteropolar blocks: 0%, 20%, 40%, 60%, 80%, and 100%. Note that the heteropolar block indicates the MFC blocks that include both positively and negatively keyed statements. These conditions reflect the various MFC designs in research settings. For example, Fisher et al. (2019) used 0%, Lee et al. (2018) utilized 70%, Wetzel and Frick (2019) included 95%, and Walton et al. (2020) used 100% heteropolar blocks in the 20-triplet MFC test. In our simulation, to create heteropolar blocks, we added one negatively keyed statement within a block. Thus, the 0% condition included no negatively keyed statement. The 20%, 40%, 60%, 80%, and 100% conditions included 4, 8, 12, 16, and 20 negatively keyed statements in the 20-triplet MFC test, respectively.
4. Scoring methods: Two scoring methods were used: PI and TIRT.
5. Latent trait ($\boldsymbol{\theta}$) intercorrelations: To mimic various intercorrelations among dimensions, three intercorrelation conditions were selected: (a) 0, (b) 0.3, and (c) 0.6.
6. Criterion-related validity: One outcome variable (Y) was created by correlating with the generated $\boldsymbol{\theta}$ scores. The correlation was set to 0.5 for the simulation purpose.
7. MFC format: This simulation used the triplet format. Among the various MFC test formats (e.g., pair, triplet, and tetrad), much of MFC research is focused on triplets with rank option (Lee et al., 2021)
8. Test dimension: Recent empirical studies using the TIRT model generally used 20-triplet MFC measures consisting of five to six dimensions (e.g., Fisher et al., 2019; Guenole et al., 2018; Lee et al., 2018; Walton et al., 2020; Watrin et al., 2019; Wetzel & Frick, 2019). We designed 20-triplet MFC tests measuring five dimensions to mimic the Big Five personality tests. 20-triplet MFC tests were made using 60 statements.

The generating parameters of loadings and thresholds adapted from Brown and Maydeu-Olivares (2012) can be found in the Appendix A. The total number of simulation conditions was 144 (i.e., $2 \times 2 \times 6 \times 2 \times 3$), and 100 replications were performed for each condition.

5.2. Simulation procedure

1. Vectors of five latent trait scores ($\boldsymbol{\theta}_d$, $d = 1, 2, \ldots 5$) were randomly sampled from a multivariate normal distribution, $MVN(\mathbf{0}, \boldsymbol{\Sigma})$, with the covariances among dimensions set to 0, 0.3 or 0.6 depending on the conditions. 60 uniquenesses (ε_s) corresponding to 60 statements were also randomly sampled from a univariate standard normal distribution as measurement errors. A criterion outcome variable was generated by correlating with $\boldsymbol{\theta}_d$ by 0.5.
2. Binary outcomes were generated across different simulation conditions. For example, to generate a pairwise binary outcome comparing the first and second statements ($y_{1,\ 2}$), a latent propensity score, $y_{1,\ 2}^*$, was first computed using the following equation:

$$y_{1,2}^* = -\tau_1 + (\lambda_1\theta_1 - \lambda_2\theta_2) + (\varepsilon_1 - \varepsilon_2). \quad (2)$$

Once $y_{1,\ 2}^*$ is generated, the binary outcome $y_{1,\ 2}$ was generated by dichotomizing $y_{1,\ 2}^*$ such that if $y_{1,\ 2}^*$ is positive, $y_{1,\ 2}$ is generated as 1 and $y_{1,\ 2}$ is generated as 0, otherwise. Binary outcomes of $y_{1,\ 3}$ and $y_{2,\ 3}$ were also generated through the same process.

3. The binary outcomes (e.g., $y_{1,\ 2}$, $y_{1,\ 3}$, and $y_{2,\ 3}$) were transformed into a ranking response. For example, if $y_{1,\ 2} = 1$, $y_{1,\ 3} = 1$, and $y_{2,\ 3} = 1$, then the ranking response is recoded as $1 > 2 > 3$, indicating the first statement is the first rank, the second statement is the second rank, and the third statement is the third rank.
4. The transformed ranking data were scored using the PI scoring method.
5. The TIRT model was fitted to the binary outcomes. The model was estimated using the mean-and-variance-adjusted unweighted least squares (ULSMV) estimator.

We conducted all the data generation and model estimations using Mplus 8 (Muthén & Muthén, 1998–2017) scripts via SAS PROC IML (SAS Institute, 2010).

5.3. Evaluation criteria

To evaluate the psychometric properties of MFC triplet measures and compare the PI and TIRT methods, we evaluated the scores with three criteria.

1. Reliability: Pearson correlations between the generating $\boldsymbol{\theta}$ parameters and estimated scores (i.e., $\widehat{\theta}$ parameters from the TIRT method and scale scores from the PI method, respectively) were computed. The correlations were then first squared and averaged across replications. Note that this measure ($r^2_{\widehat{\theta},\theta}$) is known as empirical reliability and has been reported from previous MFC studies (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Lin, 2021). To summarize the reliabilities from five dimensions, we took an average of the reliabilities from five dimensions. As for the general reliability measures, a value closer to 1.0 indicates better recovery.
2. Criterion-related validity: Criterion-related validity was evaluated based on correlations between the MFC scores (obtained from the PI and the TIRT scoring methods) and the generated outcome variable. The criterion-related validities were first averaged across replications and then averaged across dimensions to summarize the results. Given that the generated (true) correlation was 0.5, estimated criterion-related validities closer to 0.5 indicate better recovery.

3. Root Mean Squared Error (RMSE): Although RMSE is another way of operationalizing reliability, we added RMSE as an outcome to support the interpretation of the TIRT scoring accuracy. RMSE was computed for the estimated latent trait score parameters ($\widehat{\theta}$) obtained from the TIRT scoring method as follows:

$$RMSE(\widehat{\boldsymbol{\theta}}) = \sqrt{\frac{\sum (\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta})^2}{R}}, \quad (3)$$

where R is the total number of replications. RMSE for the five dimensions were computed individually and then averaged. A smaller RMSE indicates better recovery of latent trait $\widehat{\theta}$.

6. Study 1 results

Table 1 shows the detailed results of reliability and criterion-related validity in each simulation condition. To buttress the interpretation of the results, an ANOVA was conducted (see Table 2). Omega-square (ω^2) was used to examine the effect sizes, with values of 0.01, 0.06, and 0.14 representing small, medium, and large effects, respectively (Cohen, 1988). In addition, Figs. 1 and 2 show the general pattern of the simulation results.

6.1. Reliability

Table 2 shows the ANOVA results for the main effects and interactions that accounted for at least 1% of the variance in reliability recovery. All the variable factors were statistically significant ($p < 0.05$) with large effects observed for the intrablock discrimination ($\omega^2 = 0.302$), scoring method ($\omega^2 = 0.286$), and percent of heteropolar blocks ($\omega^2 = 0.227$). The ANOVA results also showed a significant interaction effect between theta correlation and percent of heteropolar blocks ($\omega^2 = 0.062$), and between theta correlation and scoring method ($\omega^2 = 0.053$).

Fig. 1 shows the graphical patterns of the reliability results across conditions. Generally, reliability improved as the proportions of heteropolar blocks increased for both the PI and TIRT scoring methods (see Fig. 1). In addition, the TIRT scoring method consistently outperformed the PI scoring method across conditions. For example, as shown in Table 1, under the condition of i) 40% heteropolar blocks, ii) a sample

Table 1
Results of reliability and criterion-related validity for Study 1.

Sample size	Factor loading	Theta corr	Scoring method	Evaluation criterion	% of heteropolar block					
					0%	20%	40%	60%	80%	100%
500	Low	0	TIRT	Reliability	0.70	0.73	0.75	0.76	0.76	0.76
				Validity	0.17	0.28	0.36	0.40	0.42	0.45
			Partially Ipsative	Reliability	0.65	0.68	0.68	0.69	0.69	0.68
				Validity	0.04	0.09	0.17	0.23	0.29	0.35
		0.3	TIRT	Reliability	0.66	0.72	0.76	0.77	0.78	0.78
				Validity	0.24	0.35	0.42	0.45	0.47	0.48
			Partially Ipsative	Reliability	0.54	0.60	0.63	0.66	0.68	0.69
				Validity	0.00	0.09	0.18	0.24	0.29	0.34
		0.6	TIRT	Reliability	0.65	0.74	0.79	0.81	0.82	0.82
				Validity	0.33	0.41	0.46	0.47	0.48	0.50
			Partially Ipsative	Reliability	0.36	0.47	0.55	0.61	0.65	0.69
				Validity	0.00	0.10	0.19	0.25	0.29	0.34
	High	0	TIRT	Reliability	0.80	0.87	0.89	0.90	0.90	0.90
				Validity	0.12	0.36	0.43	0.45	0.47	0.48
			Partially Ipsative	Reliability	0.77	0.81	0.83	0.84	0.84	0.84
				Validity	0.02	0.11	0.21	0.29	0.36	0.43
		0.3	TIRT	Reliability	0.75	0.87	0.89	0.90	0.90	0.90
				Validity	0.19	0.40	0.46	0.48	0.48	0.50
			Partially Ipsative	Reliability	0.66	0.74	0.79	0.82	0.84	0.84
				Validity	0.00	0.11	0.21	0.29	0.35	0.41
		0.6	TIRT	Reliability	0.71	0.87	0.90	0.91	0.91	0.91
				Validity	0.29	0.44	0.47	0.48	0.49	0.50
			Partially Ipsative	Reliability	0.47	0.61	0.71	0.77	0.81	0.84
				Validity	0.00	0.12	0.22	0.30	0.35	0.40
1000	Low	0	TIRT	Reliability	0.71	0.74	0.76	0.77	0.77	0.77
				Validity	0.19	0.29	0.37	0.40	0.43	0.46
			Partially Ipsative	Reliability	0.65	0.68	0.67	0.69	0.69	0.69
				Validity	0.04	0.09	0.17	0.24	0.29	0.35
		0.3	TIRT	Reliability	0.68	0.73	0.76	0.78	0.78	0.79
				Validity	0.26	0.36	0.42	0.45	0.47	0.49
			Partially Ipsative	Reliability	0.54	0.60	0.63	0.66	0.68	0.69
				Validity	0.00	0.09	0.18	0.24	0.29	0.35
		0.6	TIRT	Reliability	0.67	0.75	0.80	0.81	0.82	0.83
				Validity	0.35	0.42	0.46	0.48	0.49	0.50
			Partially Ipsative	Reliability	0.36	0.47	0.55	0.61	0.65	0.69
				Validity	0.00	0.10	0.19	0.25	0.29	0.34
	High	0	TIRT	Reliability	0.81	0.87	0.89	0.90	0.90	0.90
				Validity	0.14	0.36	0.43	0.46	0.47	0.48
			Partially Ipsative	Reliability	0.77	0.81	0.83	0.84	0.84	0.84
				Validity	0.02	0.11	0.21	0.29	0.36	0.43
		0.3	TIRT	Reliability	0.76	0.87	0.89	0.90	0.90	0.90
				Validity	0.21	0.41	0.46	0.48	0.49	0.50
			Partially Ipsative	Reliability	0.66	0.74	0.79	0.82	0.84	0.84
				Validity	0.00	0.11	0.21	0.29	0.35	0.42
		0.6	TIRT	Reliability	0.72	0.87	0.90	0.91	0.91	0.91
				Validity	0.31	0.44	0.47	0.49	0.49	0.50
			Partially Ipsative	Reliability	0.47	0.61	0.71	0.77	0.81	0.84
				Validity	0.00	0.12	0.22	0.30	0.35	0.40

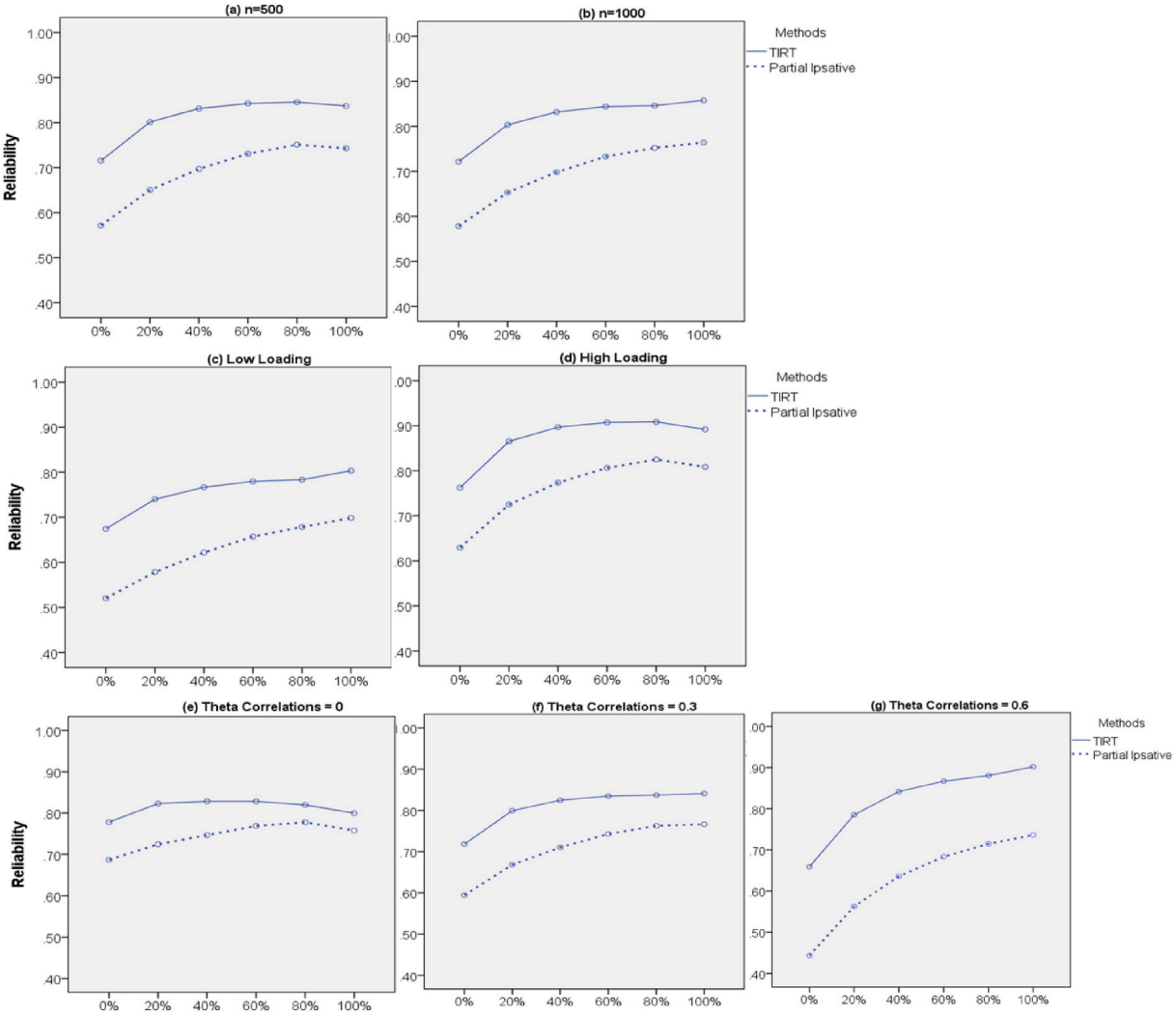

Fig. 1. Graphical presentation of reliability results for Study 1.

size of 1000, iii) high discrimination, and iv) 0.6 of theta correlation, the TIRT scoring method yielded a reliability of 0.90, while the PI scoring method yielded a reliability of 0.71. More importantly, the TIRT scoring method showed much improvement in reliability recovery as the proportion of heteropolar blocks increased from 0% to 20%. It showed a rather gradual increase after 20%. The pattern was consistent across the other simulation conditions. As shown in Fig. 1(c) and (d), higher reliabilities were found for the high intrablock discrimination condition compared with the low intrablock discrimination condition. In addition, across all theta correlation conditions, the outperformance of the TIRT method was more pronounced as the theta correlation increased.

6.2. Criterion-related validity

Table 2 also displays the ANOVA results for criterion-related validity. All the main effects were statistically significant ($p < 0.001$), with medium effects obtained for percent of heteropolar blocks ($\omega^2 = 0.486$) and scoring methods ($\omega^2 = 0.434$). Significant interaction effects between scoring methods and percent of heteropolar blocks ($\omega^2 = 0.033$) and between scoring methods and theta correlations were found ($\omega^2 = 0.010$).

Fig. 2(a) and (b) show that criterion-related validity estimates significantly improved from 0% to 20% of heteropolar block conditions, but it showed a rather gradual improvement after 20%. Similar patterns were observed across the sample size, intrablock discrimination, and theta correlation conditions. Notably, the TIRT method yielded better criterion-related validity estimates than the PI method. Under the condition of i) 40% heteropolar blocks, ii) sample size of 1000, iii) high discrimination, and iv) 0.6 of theta correlation, the TIRT method yielded the validity of 0.47, whereas the PI method yielded the validity of 0.22 under the same conditions. Under the condition of 0% heteropolar blocks, the TIRT method still yielded substantially better validity recovery than the PI method. For example, under the condition of i) 0% heteropolar blocks, ii) sample size of 1000, iii) high intrablock discrimination, and iv) 0.6 theta correlation, the TIRT method yielded the validity of 0.31, while the PI method yielded the validity of 0. Importantly, the simulation results show that, in the TIRT application, 20–40% heteropolar blocks consisting of highly discriminating statements can achieve sufficient (or acceptable) reliability (i.e., 0.87–0.90 on average) and validity (0.40–0.45 on average) recoveries.

6.3. RMSE

Table 2 shows the ANOVA results for main effects and interactions on the RMSE of $\widehat{\theta}$ recovery for the TIRT scoring method. All of the main effects were statistically significant ($p < 0.001$) with large effects observed for loading ($\omega^2 = 0.603$) and % of heteropolar blocks ($\omega^2 = 0.328$). A small but significant interaction effect between loading and % of heteropolar blocks was found ($\omega^2 = 0.028$).

Consistent with the reliability results, the RMSE of TIRT latent trait

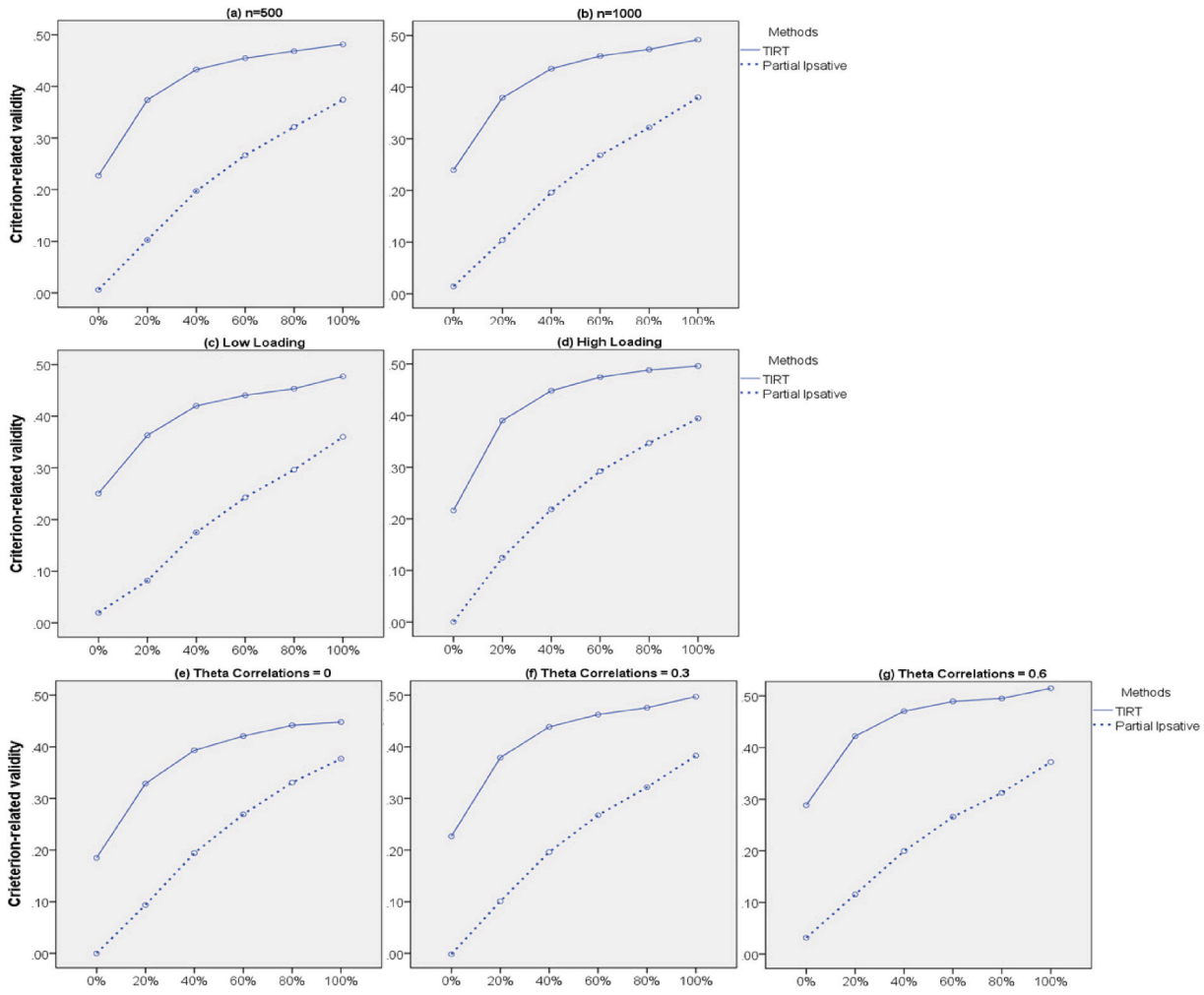

Fig. 2. Graphical presentation of criterion-related validity results for Study 1.

estimates significantly decreased from 0% to 20% of heteropolar blocks, but the reduction became decreased after 20%. This pattern was consistent across conditions. Interestingly, the sample size did not impact the RMSE. This may be because the latent trait estimation was largely influenced by the number of items, not samples size, as we fixed the number of MFC blocks to 20. As shown in Fig. 3(b), RMSE was found to be much lower for the high intrablock discrimination condition than for the low intrablock discrimination condition. Lastly, Fig. 3(c) shows

Table 2
ANOVA results of main effects and interactions on outcomes for Study 1.

Outcome	Source	df_B	F	p	ω^2
Reliability	Loading (L)	1	580.75	<0.001	0.302
	Scoring method (SM)	1	550.77	<0.001	0.286
	Heteropolar block % (HB)	5	88.05	<0.001	0.227
	Theta correlation (TC)	2	34.86	<0.001	0.035
	Sample size (S)	1	1.16	>0.05	0.001
	TC*HB	10	12.83	<0.001	0.062
	TC*SM	2	51.45	<0.001	0.053
Criterion-related validity	Heteropolar blocks % (HB)	5	714.25	<0.001	0.486
	Scoring methods (SM)	1	3184.41	<0.001	0.434
	Theta correlation (TC)	2	46.32	<0.001	0.012
	Loading (L)	1	54.44	<0.001	0.007
	Sample size (S)	1	1.80	>0.05	0.000
	SM*HB	5	49.32	<0.001	0.033
	SM*TC	2	35.68	<0.001	0.010
RMSE of TIRT model	Loading (L)	1	30,659.58	<0.001	0.603
	Heteropolar block % (HB)	5	3335.30	<0.001	0.328
	Sample size (S)	1	73.67	<0.001	0.010
	Theta correlation (TC)	2	157.90	<0.001	0.006
	L*HB	5	286.83	<0.001	0.028

Note. Only interaction effects that accounted for at least 1% of the variance in power are included. ω^2 = proportion of variance accounted for by the independent variables. df_B = degrees of freedom between.

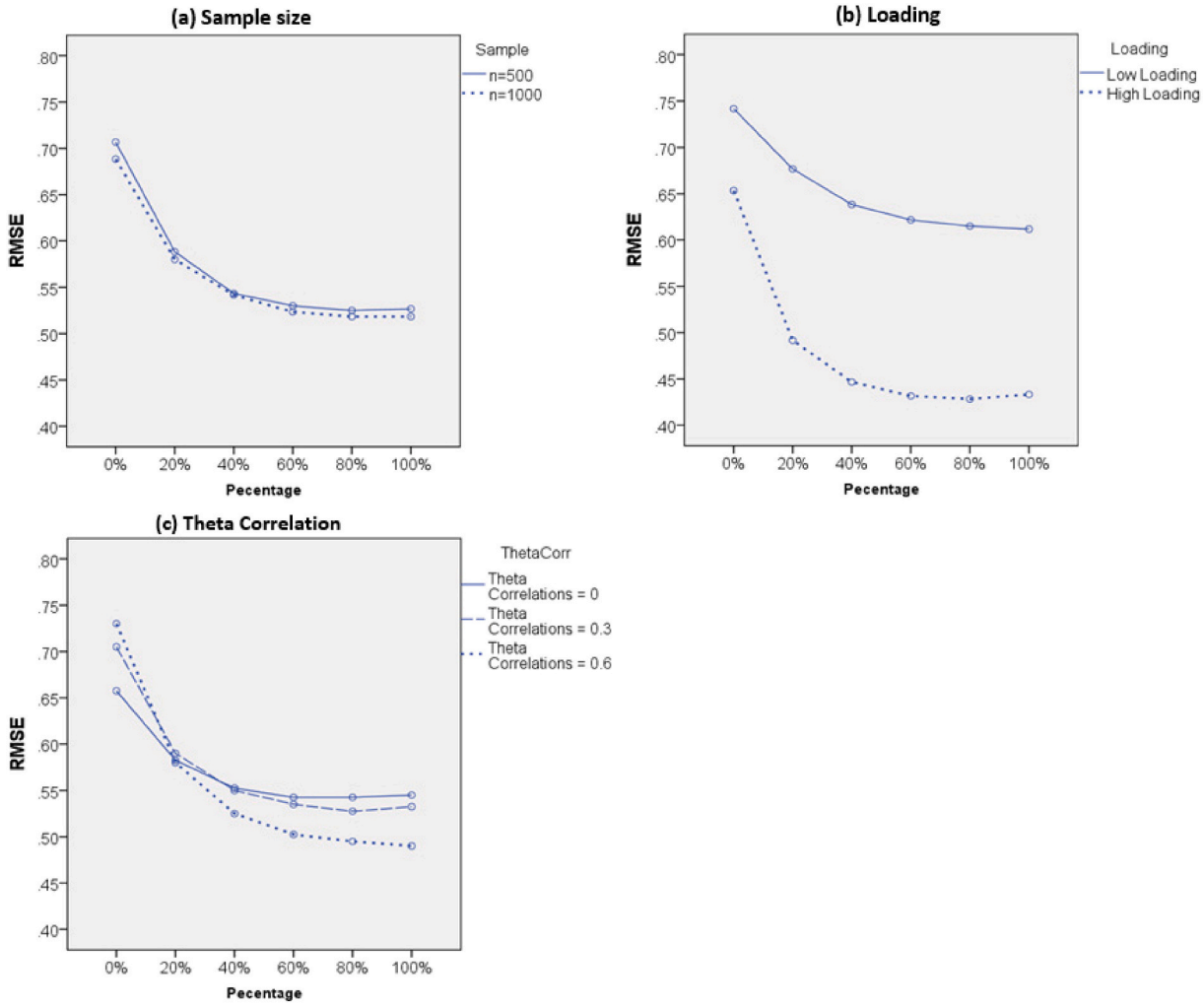

Fig. 3. Graphical presentation of RMSE results of TIRT model for Study 1.

that RMSEs decreased as % of heteropolar blocks increased for both theta correlation conditions.

7. Study 2: empirical study for criterion-related validity

There have been mixed research findings of criterion-related validities between two MFC scoring methods (i.e., PI and TIRT). One set of the literature showed that the TIRT scoring method yields similar or better criterion-related validities than the PI scoring method (e.g., Lee et al., 2018; Walton et al., 2020; Wetzel & Frick, 2019). In contrast, another study showed that the PI scoring method yielded better criterion-related validity evidence than the TIRT scoring method (Fisher et al., 2019). Fisher et al. (2019) suggested that "TIRT scoring should not be blindly implemented to replace CTT scoring (i.e., PI) on existing MFC personality assessments in practice" (p. 55).

These mixed results may stem from different strategies of including negatively keyed statements in MFC measures. When the PI scoring method is used, the inclusion of a large number of heteropolar blocks in MFC measures would improve the criterion-related validity by introducing enough variability in scale scores. By contrast, the inclusion of a small number of heteropolar blocks to maximize fake-resistance of MFC measure may degrade the criterion-related validity by reducing the variability of scale scores. For example, if a negatively keyed statement is included in the block (e.g., [A. I complete tasks successfully; B. I warm up quickly to others; C. ***I have no imagination***]), the PI scoring of the ranking response of [1, 2, 3] would be 2-, 1-, and 2-point for each statement. Therefore, a larger variability of scale scores can be obtained as compared to an MFC triplet with all positively keyed statements (e.g., 2-, 1-, and 0-point would be obtained for each statement).

We found that previous empirical studies differed appreciably in how negatively keyed statements were included in MFC measures. For example, Lee et al. (2018) developed a 20-triplet MFC test of Big Five personality with 8 positively and 4 negatively keyed statements per dimension. Specifically, 14 out of 20 blocks were heteropolar blocks. Wetzel and Frick (2019) also constructed a 20-triplet MFC measure of Big Five personality, and 19 out of 20 triplets were designed as heteropolar blocks. Walton et al. (2020) developed a 20-triplet MFC measure of Big Five personality and designed all 20 blocks as heteropolar blocks by including at least one negatively keyed statement in each block. In contrast, Fisher et al. (2019) developed a 20-triplet MFC measure of Big Five personality in a way that all blocks were equally keyed; all triplet blocks consisted of either three positively keyed statements or three negatively keyed statements to strictly match social desirability and reduce concerns about faking (Fisher, February 2020, email communication). Guenole et al. (2018) also developed a 20-triplet measure of maladaptive personality using only positively keyed statements (Guenole, June 2020, email communication). As such, test designs of MFC measures could yield different psychometric results such as criterion-related validity. In this study, we aim to empirically demonstrate the impact of heteropolar blocks on criterion-related validity, thus provide insights on the recent mixed research findings.

7.1. Sample and measure

Real examinee data were collected using Prolific, an online crowdsourcing website for survey sample participants. Following best practices in online survey data collection (e.g., Arthur et al., 2021; Kung et al., 2018), each of the two surveys had two attention check items; participants were unable to complete the survey if they missed either attention check item. Additionally, participants were unable to participate if they did not complete a Prolific pre-screening assessment indicating that they worked 31 or more hours per week on average. A total of 599 participants completed the first survey (heteropolar MFC); of these, 512 (85%) completed the second survey (equally-keyed MFC) one week later. Thus, the final sample was 512 full-time employees completed our study, with an average age of 30.07 (SD = 9.02) and a proportion of 52.73% males (n = 270). Each individual was compensated $5.50 for participating. Among them, 80.27% were white/Caucasian, 9.77% were Hispanic or Latino, 4.29% were Asian or Asian-American, 3.13% were Black or African-American, and the rest were other.

For the heteropolar MFC personality test, we used Brown and Maydeu-Olivares' (2011) 20-triplet Big Five personality measure. Each dimension (e.g., openness, conscientiousness, extraversion, agreeableness, and neuroticism) consists of 12 statements for a total of 60 statements. To facilitate the interpretation of keyed direction, we reworded the statements measuring neuroticism to statements measuring emotional stability. Therefore, a higher score for each dimension reflects a higher latent trait level for each personality dimension. In each dimension, eight statements were positively keyed and four statements were negatively keyed. Statements measuring different trait dimensions were mixed, and 15 out of 20 blocks were heteropolar blocks, including negatively keyed statements. To create the equally keyed MFC test, we also adapted Brown and Maydeu-Olivares' (2011) measure. We reworded the negatively keyed statements to positively keyed statements to make the equally keyed MFC test. The two MFC tests are shown in Appendix B.

Through the within-subject design, the study participants completed two MFC personality tests at two-time points, separated by one week. To mimic the high-stakes test environment, the following instruction was provided:

> Imagine the following situation. You've been selected by a large Fortune 500 company to try out a new employee recruitment process. If you perform well on the tests, you will be contacted by the company for future employment opportunities.
>
> (Adapted from Bernerth, 2005)

At Time 1, the participants completed the heteropolar MFC personality test and outcome variables consisting of (a) five items for life satisfaction (Diener et al., 1985) and (b) 12 items for positive and negative effects (Diener et al., 2010). All Likert-type items used a five-point scale. At Time 2, the participants completed the equally keyed MFC personality test.

7.2. Analytical strategy

We first transformed the ranking response data into PI data. Two points were assigned when a positively keyed statement was selected as the first rank or when a negatively keyed statement was selected as the third rank. By contrast, zero points were assigned when a positively keyed statement was chosen as the third rank or when a negatively keyed statement was chosen as the first rank. The second-ranked statement was scored as 1 point. To obtain the PI response data for each dimension, we aggregated the recoded scores by each dimension. Next, we scored the transformed binary outcomes of the MFC tests using the TIRT model with the ULSMV estimator. For this purpose, we used the Mplus 8.0 program (Muthén & Muthén, 1998–2017). To evaluate the criterion-related validity of the MFC personality tests, we correlated the MFC personality test scores obtained from the PI and the TIRT scoring methods with life satisfaction, positive affect, and negative affect scores.

8. Study 2 results

Table 3 presents the summary statistics for the MFC personality tests, as well as for the three criterion measures. We also examined the factor structures of the heteropolar and equally keyed MFC personality tests, and they were supported ($\chi^2(1660) = 2171.79$, RMSEA = 0.03 [95% CI: 0.022, 0.027], CFI = 0.90, TLI = 0.89 for the heteropolar MFC test; $\chi^2(1660) = 2429.45$, RMSEA = 0.03 [95% CI: 0.027, 0.033], CFI = 0.87, TLI = 0.86 for the equally keyed MFC test). Although our fit results were not great, the results were consistent with those of previous MFC studies that used the TIRT model (e.g., RMESEA = 0.03 from Brown & Maydeu-Olivares, 2011; RMSEA = 0.03, CFI = 0.85, and SRMR = 0.098 from Guenole et al., 2018; RMSEA = 0.04, CFI = 0.90, TLI = 0.88 from Morillo et al., 2016; RMSEA = 0.03, CFI = 0.89, and TLI = 0.89 from Lee, Joo, & Lee, 2019; Lee, Joo, Stark, & Chernyshenko, 2019).

Table 4 shows the results of the criterion-related validity between the two MFC tests across the PI and the TIRT scoring methods. Overall, much stronger criterion-related validities were found for the heteropolar MFC test than the equally-keyed MFC test. Moreover, the TIRT scoring yielded stronger criterion-related validities than the PI scoring. Specifically, for the TIRT scoring method, the heteropolar MFC test produced noticeably stronger criterion-related validity coefficients ($|r|$ ranging from 0.11 to 0.52, average $|r| = 0.26$), compared with the equally keyed MFC test ($|r|$ ranging from 0.00 to 0.32, average $|r| = 0.10$). Although the validity values of the heteropolar MFC test were somewhat smaller than previous meta-analytic findings (Steel et al., 2008), they still showed similar patterns. Steel et al. (2008) found significant relationships between Conscientiousness, Extraversion, Agreeableness, and Neuroticism of NEO-PI with life satisfaction [corrected $r = 0.27$ (95% CI = 0.19 to 0.36), 0.35 (0.31 to 0.39), 0.19 (0.15 to 0.23), and −0.44 (−0.40 to −0.48)], positive affect [corrected r = 0.31 (0.26 to 0.37), 0.53 (0.50 to 0.57), 0.15 (0.11 to 0.19), and −0.35 (−0.31 to −0.38)], and negative affect [corrected $r = -0.26$ (−21 to −30), −0.22 (−0.19 to

Table 3
Summary statistics of variables for Study 2.

MFC tests	Construct	No. of items/ blocks	M	SD	Reliability
Heteropolar MFC personality test with TIRT scoring	O	12	−0.01	0.89	0.88
	C	12	−0.02	0.89	0.89
	E	12	0.01	0.91	0.87
	A	12	−0.02	0.88	0.88
	S	12	−0.01	0.87	0.89
Equally keyed MFC personality test with TIRT scoring	O	12	−0.01	0.88	0.76
	C	12	−0.01	0.91	0.75
	E	12	0.01	0.88	0.76
	A	12	−0.02	0.88	0.74
	S	12	0.00	0.85	0.76
Heteropolar MFC personality test with PI scoring	O	12	14.76	4.40	0.76
	C	12	15.02	4.20	0.70
	E	12	10.27	4.93	0.66
	A	12	15.22	4.29	0.70
	S	12	12.34	4.08	0.69
Equally keyed MFC personality test with PI scoring	O	12	15.13	4.40	0.72
	C	12	12.96	4.31	0.64
	E	12	8.28	4.91	0.76
	A	12	13.88	4.12	0.65
	S	12	9.75	3.99	0.61
Life satisfaction	–	5	3.22	0.92	0.86
Positive affect	–	6	3.69	0.66	0.90
Negative affect	–	6	2.63	0.73	0.85

Notes: O = openness, C = conscientiousness, E = extraversion, A = agreeableness, S = emotional stability. Reliabilities of PI scoring were based on coefficient alpha. Reliabilities of TIRT scoring were based on empirical reliability ($r^2_{\hat{\theta},\theta}$). N = 512 for each variable.

Table 4
Criterion-related validity coefficients of MFC personality tests for Study 2.

Scoring method	Dimension	Heteropolar MFC test			Equally-keyed MFC test		
		LifeSat	AffPos	AffNeg	LifeSat	AffPos	AffNeg
PI scoring	O	0.18***	0.24***	−0.20***	0.00	0.00	0.03
	C	0.13**	0.07	−0.13**	−0.10*	−0.23***	0.10*
	E	0.24***	0.33***	−0.21***	0.04	0.12**	0.01
	A	0.14**	0.17***	−0.20***	0.02	0.05	0.06
	S	0.21***	0.36***	−0.49***	0.05	0.06	−0.21***
TIRT scoring	O	0.22***	0.29***	−0.28***	0.03	0.01	−0.02
	C	0.19***	0.18***	−0.26***	−0.04	−0.14**	0.02
	E	0.27***	0.36***	−0.28***	0.07	0.17***	−0.01
	A	0.18***	0.23***	−0.24***	0.06	0.06	0.00
	S	0.25***	0.38***	−0.52***	0.12**	0.15***	−0.32***

Notes: LifeSat = life satisfaction, AffPos = positive affect, AffNeg = negative affect, O = openness, C = conscientiousness, E = extraversion, A = agreeableness, S = emotional stability.
* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$

−0.26), −0.25 (−0.20 to −0.29), and 0.64 (0.60 to 0.67)].

A similar pattern was observed for the PI scoring method. The heteropolar MFC test produced stronger criterion-related validity coefficients for life satisfaction, positive affect, and negative affect ($|r|$ ranging from 0.05 to 0.49, average $|r| = 0.20$), compared with the equally keyed MFC test ($|r|$ ranging from 0.00 to 0.23, average $|r| = 0.09$). Notably, we found that the validity evidence from the equally keyed MFC test was seriously distorted for both the PI and TIRT scoring methods. For example, correlations between agreeableness and a) life satisfaction, b) positive affect, and c) negative affect became almost zero values for the equally keyed MFC test across the PI and the TIRT scoring methods.

9. Discussion

Recently, the use of MFC personality tests has drawn much attention in personnel assessments (Cao & Drasgow, 2019; Wetzel et al., 2020). Although the test design associated with heteropolar blocks is important for MFC test application, very few studies have explored this topic. This research 1) explored extent to which heteropolar blocks influence the reliability and validity of MFC tests through Monte Carlo simulation and 2) empirically demonstrated how MFC test designs associated with heteropolar blocks influence criterion-related validity evidence of MFC personality assessment using real examinee data.

Our main findings are as follows. First, the simulation study indicated that the manner in which negatively keyed statements are included and MFC data are scored are important considerations in MFC test applications. Specifically, the TIRT scoring method and higher intrablock discrimination yielded better reliability and criterion-related validity. In addition, our simulation results suggest that one can achieve sufficient reliability and validity by using highly discriminating 20–40% heteropolar blocks and applying the TIRT model. Second, our empirical demonstration showed criterion-related validity results can be different depending on the test designs of heteropolar blocks. Specifically, stronger criterion-related validity was found when the heteropolar MFC test was used rather than the equally keyed MFC test. In addition, the TIRT scoring method consistently yielded better results than the PI scoring method in terms of criterion-validity.

9.1. Implications for personality MFC test design and application

Firstly, our study indicates that different scoring methods could produce substantially different psychometric results depending on the conditions used. For example, with regard to the criterion-related validity, the TIRT method substantially outperformed the PI method across all conditions. Our findings indicate that recent suggestions that the PI method can be effectively used for personnel selection based on criterion-related validity (e.g., Lee et al., 2018; Salgado et al., 2015; Salgado & Lado, 2018) should be reconsidered. In contrast to Fisher et al.'s (2019) arguments (i.e., "TIRT scoring should not be blindly implemented to replace CTT scoring (i.e., PI) on existing FC personality assessments in practice" [p. 55]), our results indicate that the TIRT-based scoring method is superior in all simulation conditions. Thus, their argument may need to be reconsidered. We believe that the general improvement of TIRT scoring over PI scoring is due to better psychometric estimation of ranking data and better removal of ipsativity. This is a significant result in favor of TIRT as it applies to MFC measures. Particularly, we recommend the use of the TIRT method for the high-stakes decision-making purpose because of the better psychometric evidence rather than PI method. Consequently, this could increase the quality of decision-making and utility of MFC measures in high-stakes settings.

Furthermore, as described previously, recent research has shown mixed results of criterion-related validity between PI and TIRT scoring methods (Fisher et al., 2019; Lee et al., 2018; Walton et al., 2020; Wetzel et al., 2016; Wetzel & Frick, 2019). Our findings revealed that the conflicting findings may stem from different proportions of heteropolar blocks included across studies. Thus, it is difficult to generalize that PI scoring and TIRT scoring methods provide similar or different criterion-related validity evidence without considering the proportions of heteropolar blocks. Existing meta-analysis concerning the validity and fake-resistance of MFC measures exclusively focused on the traditional scoring methods (partially ipsative or fully ipsative) rather than IRT methods (e.g., Cao & Drasgow, 2019; Salgado et al., 2015; Salgado & Lado, 2018). Proportions of negatively keyed statements were not considered as a possible moderator. Although there are still a small number of empirical studies using MFC-TIRT models, a new meta-analysis on criterion-related validity of MFC measures based on the TIRT method will be needed.

Recommendations for the development of MFC measures from the TIRT model often promote the inclusion of negatively keyed statements to ensure psychometric properties and validity. This situation has created a dilemma for researchers and practitioners, that is, choosing between improving psychometric properties (e.g., reliability and validity) and sacrificing fake-resistant properties. Ng et al. (2021) recently pointed out that "one cannot both create item blocks that contain negatively and positively keyed statements and also match them within item block on desirability because the negatively keyed statements are either obviously or at least apparently less desirable" (p. 227). They also suggested that including negatively keyed statements in MFC blocks may make the MFC test more fakable and cause psychometric issues. This emphasizes the importance of the block design of MFC tests. Our

simulation findings suggest a potential balancing point to resolving this dilemma: we found that inclusion of heteropolar blocks with highly discriminating statements in 20% to 40% of the measure would be enough to achieve acceptable psychometric properties (i.e., sufficient reliability and validity), which also may contribute to reducing any negative impacts stemming from negatively keyed statements.

Methodologically, measurement invariance is achieved "when the relations between observed test scores and the latent attribute measured by the test are identical across subgroups" (Drasgow, 1984, p. 134). Noninvariance at the item level is referred to as differential item functioning (DIF). Our suggestion would also reduce any possible DIF for MFC measures between high-stakes settings and low-stakes settings that may occur from heteropolar blocks (Lee & Joo, 2021). Furthermore, by reducing disruption in the item-response process from the differently keyed directions (that increases cognitive load), adverse impact against a selective group may be mitigated in a personnel selection context. Finally, one may also improve the test-retest reliability of MFC measures as well, by using a relatively small number of heteropolar blocks, which can reduce transient or situational influences of MFC responses.

9.2. Limitations and suggestions for future research

Although our research deepens the understanding of MFC applications, it still has several limitations that should be noted and addressed in future research. First, although a total of 144 simulation conditions were conducted, they still cannot capture all possibilities that could occur in real settings. For example, to create heteropolar blocks, we added only one negatively keyed statement with a block. However, a different number of negatively keyed statements still can be included (e.g., two negatively keyed statements and one positively keyed statement). Future research should investigate more diverse conditions of negatively keyed statements.

Second, this research did not explore organizational data in a true high-stakes setting. Although the real examinee data for the empirical demonstration was collected using instructions with a selection scenario, the data is still limited in its ability to represent real high-stakes settings. Future research should investigate the validity issues between two different scoring methods in a real personnel selection context. Also, our outcome variables for criterion-related validity were collected via self-reported measures. Future research should use other-reported data in the organizational settings (e.g., job performance).

Third, among many IRT models for MFC measures, this research only focused on the TIRT model, which is a dominance response model. However, ideal point models have recently been increasingly applied in areas such as personality (Chernyshenko et al., 2001; Stark et al., 2006), vocational interest (Tay et al., 2011), job satisfaction (Carter & Dalal, 2010), and future research should examine whether our findings based on the dominance-based MFC IRT model can also be generalized to ideal-point based MFC IRT models (e.g., Joo et al., 2021; Lee, Joo, & Lee, 2019; Lee, Joo, Stark, & Chernyshenko, 2019; Stark et al., 2005).

10. Conclusion

Although negatively keyed statements play a key role in the fake-resistance and psychometric properties of MFC measures, very few studies have investigated the effect of negatively keyed statements. Our research filled this research gap through Monte Carlo simulation and an empirical demonstration. We recommend that researchers and practitioners use the TIRT-based scoring method rather than the PI method for the high-stakes decision making and include 20–40% of highly discriminating heteropolar blocks to MFC measures to simultaneously ensure the fake-resistance and psychometric properties such as reliability and validity. We hope this research provides a solid foundation for personality MFC assessment and a springboard for future psychometric development efforts.

CRediT authorship contribution statement

Philseok Lee: Conceptualization, Writing Draft.
Seang-Hwane Joo: Methodology and Data Analytics.
Steven Zhou: Data Collection and Editing.
Mina Son: Data Collection and Validation.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendix A. Generating item parameters for the main simulation study

Block	Dimension	Low loading	High loading	Threshold	Negatively keyed statement					
					0%	20%	40%	60%	80%	100%
1	1	0.5	1	0.5		–	–	–	–	–
	2	0.3	0.8	−1.2						
	3	0.8	1.3	−1.7						
2	1	0.8	1.3	0.7						
	2	0.5	1	1		–	–	–	–	–
	4	0.3	0.8	0.3						
3	1	0.3	0.8	−0.7						
	2	0.8	1.3	−1.2						
	5	0.5	1	−0.5			–	–	–	–
4	1	0.8	1.3	0.7			–	–	–	–
	3	0.3	0.8	1.2						
	4	0.5	1	0.5						
5	1	0.5	1	0.5				–	–	–
	3	0.3	0.8	−1.2						
	5	0.8	1.3	−1.7						
6	1	0.8	1.3	0.7						
	4	0.5	1	1			–	–	–	–
	5	0.3	0.8	0.3						
7	2	0.3	0.8	−0.7			–	–	–	–
	3	0.8	1.3	−1.2						
	4	0.5	1	−0.5						
8	2	0.8	1.3	0.7						
	3	0.3	0.8	1.2		–	–	–	–	–

(continued on next page)

(*continued*)

Block	Dimension	Low loading	High loading	Threshold	Negatively keyed statement					
					0%	20%	40%	60%	80%	100%
	5	0.5	1	0.5						
9	2	0.5	1	0.5						
	4	0.3	0.8	−1.2		–		–	–	–
	5	0.8	1.3	−1.7						
10	3	0.8	1.3	0.7			–	–	–	–
	4	0.5	1	1						
	5	0.3	0.8	0.3						
11	1	0.3	0.8	−0.7					–	–
	2	0.8	1.3	−1.2						
	3	0.5	1	−0.5						
12	1	0.8	1.3	0.7						
	2	0.3	0.8	1.2				–	–	–
	4	0.5	1	0.5						
13	1	0.5	1	0.5						
	2	0.3	0.8	−1.2						
	5	0.8	1.3	−1.7				–	–	–
14	1	0.8	1.3	0.7						
	3	0.5	1	1					–	–
	4	0.3	0.8	0.3						
15	1	0.3	0.8	−0.7						
	3	0.8	1.3	−1.2						
	5	0.5	1	−0.5					–	–
16	1	0.8	1.3	0.7						
	4	0.3	0.8	1.2					–	–
	5	0.5	1	0.5						
17	2	0.5	1	0.5						–
	3	0.3	0.8	−1.2						
	4	0.8	1.3	−1.7						
18	2	0.8	1.3	0.7						
	3	0.5	1	1						–
	5	0.3	0.8	0.3						
19	2	0.3	0.8	−0.7						
	4	0.8	1.3	−1.2						–
	5	0.5	1	−0.5						
20	3	0.8	1.3	0.7						
	4	0.3	0.8	1.2						
	5	0.5	1	0.5						–

Appendix B

Appendix B.1
Heteropolar MFC test.

Block	Dimension	Keyed	Statements
1	5	+	I am relaxed most of the time.
	3	+	I start conversations.
	1	+	I catch on to things quickly.
2	4	+	I show my gratitude.
	2	+	I do things according to a plan.
	5	+	I am not easily bothered by things.
3	1	−	I have difficulty understanding abstract ideas.
	3	+	I am the life of the party.
	4	+	I inquire about others' well-being.
4	2	+	I like order.
	1	+	I am good at many things.
	5	−	I get upset easily.
5	4	+	I sympathize with others' feelings.
	5	−	I worry about things.
	3	+	I feel at ease with people.
6	1	+	I love to think up new ways of doing things.
	3	−	I am quiet around strangers.
	2	−	I often forget to put things back in their proper place.
7	3	−	I keep in the background.
	5	−	I have frequent mood swings.
	4	+	I feel others' emotions.
8	2	+	I follow a schedule.
	1	+	I am full of ideas.
	3	−	I don't talk a lot.
9	1	+	I love to read challenging material.
	5	−	I get overwhelmed by emotions.
	4	−	I am not interested in other people's problems.
10	2	−	I waste my time.

(*continued on next page*)

Appendix B.1 (*continued*)

Block	Dimension	Keyed	Statements
	5	−	I get irritated easily.
	3	+	I talk to a lot of different people at parties.
11	3	+	I feel comfortable around people.
	4	+	I love to help others.
	2	+	I get jobs done right away.
12	5	+	I seldom feel blue.
	4	+	I know how to comfort others.
	1	−	I avoid difficult reading material.
13	3	−	I find it difficult to approach others.
	5	−	I panic easily.
	2	−	I neglect my duties.
14	4	+	I make time for others.
	2	+	I am always prepared.
	1	+	I can handle a lot of information.
15	3	+	I make friends easily.
	1	+	I have excellent ideas.
	5	−	I get stressed out easily.
16	2	+	I make plans and stick to them.
	5	+	I rarely get irritated.
	4	−	I am indifferent to the feelings of others.
17	2	−	I leave a mess in my room.
	4	+	I make people feel at ease.
	1	+	I am quick to understand things.
18	4	−	I feel little concern for others.
	3	+	I don't mind being the center of attention.
	1	−	I lack imagination.
19	1	−	I have difficulty imagining things.
	2	+	I like to tidy up.
	5	−	I often feel blue.
20	2	+	I love order and regularity.
	4	−	I am not really interested in others.
	3	+	I am skilled in handling social situations.

Note. Dimension 1 = Openness, 2 = Conscientiousness, 3 = Extroversion, 4 = Agreeableness, 5 = Emotional Stability.

Appendix B.2
Equally-keyed MFC test.

Block	Dimension	Keyed	Statements
1	5	+	I am relaxed most of the time.
	3	+	I start conversations.
	1	+	I catch on to things quickly.
2	4	+	I show my gratitude.
	2	+	I do things according to a plan.
	5	+	I am not easily bothered by things.
3	1	+	I easily understand abstract ideas.
	3	+	I am the life of the party.
	4	+	I inquire about others' well-being.
4	2	+	I like order.
	1	+	I am good at many things.
	5	+	I do not get upset easily.
5	4	+	I sympathize with others' feelings.
	5	+	I do not worry about things.
	3	+	I feel at ease with people.
6	1	+	I love to think up new ways of doing things.
	3	+	I am not quiet around strangers.
	2	+	I do not forget to put things back in their proper place.
7	3	+	I do not keep in the background.
	5	+	I do not have frequent mood swings.
	4	+	I feel others' emotions.
8	2	+	I follow a schedule.
	1	+	I am full of ideas.
	3	+	I talk a lot.
9	1	+	I love to read challenging material.
	5	+	I do not get overwhelmed by emotions.
	4	+	I am interested in other people's problems.
10	2	+	I rarely waste my time.
	5	+	I do not get irritated easily.
	3	+	I talk to a lot of different people at parties.
11	3	+	I feel comfortable around people.
	4	+	I love to help others.
	2	+	I get jobs done right away.
12	5	+	I seldom feel blue.
	4	+	I know how to comfort others.
	1	+	I enjoy difficult reading material.

(*continued on next page*)

Appendix B.2 (*continued*)

Block	Dimension	Keyed	Statements
13	3	+	I find it easy to approach others.
	5	+	I do not panic easily.
	2	+	I do not neglect my duties.
14	4	+	I make time for others.
	2	+	I am always prepared.
	1	+	I can handle a lot of information.
15	3	+	I make friends easily.
	1	+	I have excellent ideas.
	5	+	I do not get stressed out easily.
16	2	+	I make plans and stick to them.
	5	+	I often get irritated.
	4	+	I care deeply about the feelings of others.
17	2	+	I rarely leave a mess in my room.
	4	+	I make people feel at ease.
	1	+	I am quick to understand things.
18	4	+	I feel much concern for others.
	3	+	I don't mind being the center of attention.
	1	+	I have lots of imagination.
19	1	+	I have no difficulty imagining things.
	2	+	I like to tidy up.
	5	+	I rarely feel blue.
20	2	+	I love order and regularity.
	4	+	I am interested in others.
	3	+	I am skilled in handling social situations.

Note. Dimension 1 = Openness, 2 = Conscientiousness, 3 = Extroversion, 4 = Agreeableness, 5 = Emotional Stability.

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*, 83–97.

Aon, H. (2015). *2015 Trends in global employee engagement report*. Lincolnshire, IL: Aon Corp.

Arthur, W., Jr., Hagen, E., & George, F., Jr. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior, 8*, 105–137.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*, 263–272.

Bernerth, J. B. (2005). Perceptions of justice in employment selection decisions: The role of applicant gender. *International Journal of Selection and Assessment, 13*, 206–212.

Biderman, M. D., Nguyen, N. T., Cunningham, C. J., & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the big five. *Journal of Research in Personality, 45*, 417–429.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965–973.

Bowen, C. C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis, 10*, 240–259.

Brown, A., & Bartram, D. (2009). *Development and psychometric properties of OPQ32r (Supplement to the OPQ32 technical manual)*. Thames Ditton, UK: SHL Group Limited.

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

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

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of thurstonian IRT models. *Educational and Psychological Measurement, 79*, 827–854.

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*, 1347–1368.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49*, 743–748.

CEB. (2010). *Global personality inventory—Adaptive technical manual*. Thames Ditton, UK: CEB.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publisher.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality tests and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155–169.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research, 97*, 143–156.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science, 48*, 209–225.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*, 440–464.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81–106.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*, 134–135.

Ferrando, P. J., Anguiano-Carrasco, C., & Chico, E. (2011). The impact of acquiescence on forced-choice responses: A model-based analysis. *Psicológica, 32*, 87–105.

Fisher, P. A., Robie, C., Christiansen, N. D., & Komar, S. (2018). The impact of psychopathy and warnings on faking behavior: A multisaturation perspective. *Personality and Individual Differences, 127*, 39–43.

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*, 49–61.

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*, 513–526.

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55*, 794–800.

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*, 9–24.

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

Hirsh, J. B., Morisano, D., & Peterson, J. B. (2008). Delay discounting: Interactions between personality and cognitive ability. *Journal of Research in Personality, 42*, 1646–1650.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*, 435–455.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162.

Joo, S. H., Lee, P., & Stark, S. (2019). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods, 52*, 761–772.

Joo, S. H., Lee, P., & Stark, S. (2021). Modeling multidimensional forced choice measures with the zinnes and griggs pairwise preference item response theory model. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2021.1960142

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology, 67*(2), 264–283.

Lee, P., & Joo, S. H. (2021). A new investigation of fake-resistance of multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions, 7*, 31–48.

Lee, P., Joo, S. H., & Lee, S. (2019). Examining stability of personality profile solutions between likert-type and multidimensional forced choice measure. *Personality and Individual Differences, 142*, 13–20.

Lee, P., Joo, S. H., & Stark, S. (2021). Detecting DIF in multidimensional forced-choice measures using the thurstonian item response theory model. *Organizational Research Methods, 24*, 739–771.

Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement, 43*, 226–240.

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

Lin, Y. (2021). Reliability estimates for IRT-based forced-choice assessment scores. *Organizational Research Methods*, 1–16.

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

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222–248.

McLarnon, M. J., & Carswell, J. J. (2013). The personality differentiation by intelligence hypothesis: A measurement invariance investigation. *Personality and Individual Differences, 54*, 557–561.

McLarnon, M. J., Goffin, R. D., Schneider, T. J., & Johnston, N. G. (2016). To be or not to be: Exploring the nature of positively and negatively keyed personality items in high-stakes testing. *Journal of Personality Assessment, 98*, 480–490.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–551.

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) affective commitment scale: Effects of negatively worded items. *Journal of Business and Psychology, 27*, 421–436.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and markov chain Monte Carlo estimation. *Applied Psychological Measurement, 40*, 500–516.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide*. Retrieved from (8th ed.). Los Angeles, CA: Muthén & Muthén www.statmodel.com.

Ng, V., Lee, P., Ho, 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*, 224–237.

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., & Carswell, J. J. (2017). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences, 115*, 120–127.

Pelled, L. H. (1996). Demographic diversity, conflict, and work group outcomes: An intervening process theory. *Organization Science, 7*, 615–631.

Robie, C., Risavy, S. D., Jacobs, R. R., Christiansen, N. D., König, C. J., & Speer, A. B. (2021). An updated survey of beliefs and practices related to faking in individual assessments. *International Journal of Selection and Assessment*, 1–7.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*, 113–130.

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*, 797–834.

Salgado, J. F., & Lado, M. (2018). Faking resistance of a quasi-ipsative forced-choice personality inventory without algebraic dependence. *Journal of Work and Organizational Psychology, 34*, 213–216.

SAS Institute. (2010). *SAS 9.3 user's guide*. Cary, NC: Author.

Sliter, K. A., & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement, 74*, 214–226.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184–203.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2012). Constructing fake-resistant personality tests using item response theory: High stakes personality testing with multidimensional pairwise preferences. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessments* (pp. 214–239). NY: New York: Oxford University Press.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*, 153–164.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25–39.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*, 138–161.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement, 35*, 280–295.

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*, 706–718.

Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus likert responses on an occupational big five questionnaire. *Journal of Individual Differences, 40*, 134–148.

Weijerts, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*, 320–334.

Wetzel, E., & Frick, S. (2019). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment, 32*, 239–253.

Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales. *European Journal of Psychological Assessment, 36*, 511–515.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. *European Journal of Psychological Assessment, 34*, 1–5.

Wetzel, E., Roberts, B. W., Fraley, R. C., & Brown, A. (2016). Equivalence of narcissistic personality inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality, 61*, 87–98.

White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM). Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA*.