Personnel Assessment and Decisions

Volume 7
Issue 1 *Special Issue: Understanding Effects of Impression Management on Assessment Outcomes*

Article 4

2021

A New Investigation of Fake Resistance of a Multidimensional Forced-Choice Measure: An Application of Differential Item/Test Functioning

Philseok Lee
George Mason University

Seang-Hwane Joo
University of Kansas

Recommended Citation

Lee, Philseok and Joo, Seang-Hwane (2021) "A New Investigation of Fake Resistance of a Multidimensional Forced-Choice Measure: An Application of Differential Item/Test Functioning," *Personnel Assessment and Decisions*: Vol. 7 : Iss. 1 , Article 4.
DOI: https://doi.org/10.25035/pad.2021.01.004
Available at: https://scholarworks.bgsu.edu/pad/vol7/iss1/4

A New Investigation of Fake Resistance of a Multidimensional Forced-Choice Measure: An Application of Differential Item/Test Functioning

Philseok Lee[1] and Seang-Hwane Joo[2]

1. George Mason University
2. University of Kansas

ABSTRACT

To address faking issues associated with Likert-type personality measures, multidimensional forced-choice (MFC) measures have recently come to light as important components of personnel assessment systems. Despite various efforts to investigate the fake resistance of MFC measures, previous research has mainly focused on the scale mean differences between honest and faking conditions. Given the recent psychometric advancements in MFC measures (e.g., Brown & Maydeu-Olivares, 2011; Stark et al., 2005; Lee et al., 2019; Joo et al., 2019), there is a need to investigate the fake resistance of MFC measures through a new methodological lens. This research investigates the fake resistance of MFC measures through recently proposed differential item functioning (DIF) and differential test functioning (DTF) methodologies for MFC measures (Lee, Joo, & Stark, 2020). Overall, our results show that MFC measures are more fake resistant than Likert-type measures at the item and test levels. However, MFC measures may still be susceptible to faking if MFC measures include many mixed blocks consisting of positively and negatively keyed statements within a block. It may be necessary for future research to find an optimal strategy to design mixed blocks in the MFC measures to satisfy the goals of validity and scoring accuracy. Practical implications and limitations are discussed in the paper.

KEYWORDS

faking, personality, multidimensional forced-choice measure, changing item paradigm, measurement invariance, differential item functioning, differential test functioning, high-stake setting

Historically, personality measures have been widely used for managerial and organizational decision making (Stark et al., 2012; Hough et al., 2015). Interest in the personality measures stems from research findings that personality predicts important job-related outcomes such as job performance (Barrick et al., 2001), training performance (Colquitt et al., 2000), and teamwork and team performance (Peeters et al., 2006). Additionally, the use of personality measures reduces adverse impact and provides incremental validity over cognitive ability tests in predicting job performance (Hough & Oswald, 2008).

Despite the popularity, there have been overwhelming concerns about faking (i.e., conscious attempts to make a positive impression) associated with Likert-type measures. Likert-type measures present multiple statements individually to respondents and ask them to indicate their level of agreement or disagreement according to a set of response categories (e.g., five option or seven option). However, in the high-stake settings, such as in personnel selections, respondents can easily fake their answers by simply choosing a more socially desirable response option. The resulting responses can distort test reliability and validity, change rankings of applicants, and reduce the utility of selection systems (e.g., Bott et al., 2007, Komar et al., 2008; Lee et al., 2017; Mueller-Hanson et al., 2003; Peeters & Lievens, 2005; Salgado, 2016).

To address faking issues associated with Likert-type personality measures, multidimensional forced-choice (MFC) measures have recently come to light as important components of personnel assessment systems (e.g., Anguiano-Carrasco et al., 2015; Brown & Maydeu-Olivares,

Corresponding author:
Philseok Lee
George Mason University, 4400 University Dr., Fairfax, VA 22030
Email: plee27@gmu.edu

2011; Guenole et al., 2018; Lee et al., 2018; Lee et al., 2020; Stark et al., 2012; Wetzel & Greiff, 2018). MFC measures present two (i.e., a pair), three (i.e., a triplet), or four (i.e., a quartet) statements representing different constructs within an item block, which forces respondents to either select a "most like me" statement or to rank statements from "most like me" to "least like me" in each block. Respondents may experience difficulty discerning the most desirable answers because statements within a block are matched based on a similar level of social desirability and/or item extremity. Therefore, faking responses can be reduced (Wetzel et al., 2020).

For the effectiveness of MFC measures, there have been somewhat mixed research findings. For example, Heggestad et al. (2006) discovered that MFC measures do not necessarily reduce faking in an individual-level analysis over Likert-type measures. More recently, Young (2018) identified that the pairwise preference MFC measure of a dark triad was not more fake resistant than a Likert-type measure. Additionally, Ng et al. (2021) similarly found that the triplet MFC measure of character did not reduce faking responses over a Likert-type measure. However, a multitude of studies provided more favorable results to MFC measures, showing that MFC measures successfully reduce test score inflation (e.g., Cao & Drasgow, 2019; Martin et al., 2002; Christiansen et al., 2005; Jackson et al., 2000; Trent et al., 2020; Vasilopoulos et al., 2006; Lee et al., 2019; Wetzel et al., 2020) and maintain validity in motivated testing situations (e.g., Bartram, 2007; Hirsh & Peterson, 2008; Lee et al., 2018; O'Neill et al., 2017; Zhang et al., 2020).

Investigating Fake Resistance for MFC Measures

Despite various efforts to investigate the fake resistance of MFC measures, prior research mainly focused on the scale mean differences between honest and faking conditions (e.g., Martin et al., 2002; Converse et al., 2008; Fisher et al., 2019; Jackson et al., 2000; O'Neill et al., 2017; Vasilopoulos et al., 2006). For example, Jackson et al. (2000) showed that the MFC measure is more effective in reducing faking than a Likert-type measure, as indicated by the mean differences (i.e., Cohen's *d*) between the honest and faking samples (i.e., 0.32 for the MFC measure vs. 0.95 for the used Likert-type measure). Further, Martin et al. (2002) conducted an analysis of variance to discover a significant interaction between test forms (MFC and Likert-type measures) and test conditions (honest and faking). The MFC measure yielded no differences in personality scores regardless of whether respondents were in the honest or the faking conditions. Alternatively, the Likert-type measure produced significant score inflation in the faking condition.

Nevertheless, previous studies do not provide an in-depth understanding of the response process of the two personality item formats between honest and motivated test conditions, as they exclusively focused on the composite scale-level scores. Given the recent advancements of item response theory (IRT) for MFC measures (e.g., Brown & Maydeu-Olivares, 2011; Stark et al., 2005; Lee et al., 2019; Joo et al., 2020), there is a current need to investigate the fake resistance of MFC measures through a new methodological lens. One approach is to apply differential item functioning (DIF) and differential test functioning (DTF) methodologies across different testing situations (Robie et al., 2001). DIF refers to a particular item that may have different response probabilities for different groups of people even though they have the same latent traits level (Camilli & Shepard, 1994), and DTF refers to the differences in the expected total test scores of the respondents with an equal level of latent traits (Drasgow & Hulin, 1990). Through DIF and DTF methodologies, it is possible to evaluate which personality measure (i.e., MFC or Likert-type measures) is more fake resistant at the item and test level across different testing conditions. Research suggests that the presence of DIF and DTF in personality measures can be interpreted as evidence of faking (Griffin et al., 2004; Stark et al., 2001; Zickar & Robie, 1999).

Faking the Response Process in MFC and Likert-Type Measures

To model the faking response process, Zickar and Robie (1999) proposed a *changing person paradigm* and a *changing items paradigm.* The former assumes respondents change the person's latent trait (i.e., theta shift) by the process of faking response. In contrast, the latter assumes that respondents perceive items differently, resulting in differences between item parameters. Although research has generally supported the changing person paradigm (Robie et al., 2001; Stark et al., 2004; Zickar & Robie, 1999), this study employs the changing items paradigm because DIF and DTF are related to the item- and test-level biases, and the changing items paradigm enable an evaluation of the differential nature of item responses between MFC and Likert-type measures under honest and faking conditions.

Zickar (2000) noted that changes in how items are perceived and interpreted might yield different consequences of choosing particular items. The respondents may experience a different decision-making process between MFC and Likert-type items due to the distinct cognitive processes of perceiving and deciding among different item responses. For Likert-type items, respondents are assumed to evaluate their absolute level of agreement or disagreement for each statement and indicate a response option that best fits their latent trait. In contrast, for MFC items, respondents are assumed to conduct comparative judgment among statements within a block and rank them according to their preference.

In MFC measures, ranking decision making involves a much more complicated interaction among statements within a block. Lin and Brown (2017) noted that item parameters (e.g., loadings and thresholds) for MFC measures could be affected by interactions of surrounding statements within a block, which is referred to as a contextual effect. Some statements become more socially desirable than other statements, depending on a combination of different traits within a block, leading to "desirability-induced response biases" (Lin & Brown, 2017, p. 409). The contextual effect of MFC measures would not only make DIF situations more complicated but also yield different natures of differential functioning compared to Likert-type measures. Therefore, it is not guaranteed that item parameters obtained from the single-statement Likert-type measure are still invariant when they are paired in MFC blocks. Besides, the measurement invariance of MFC measures between honest and faking test conditions should not be simply assumed. Nevertheless, previous research generally accepts the invariance assumption without testing measurement biases (Morillo et al., 2019; Pavlov et al., 2019). Considering that the main purpose of MFC measures is to reduce faking, it is particularly important to confirm the measurement invariance of the MFC measure between honest and faking conditions.

Recent Developments of the MFC DIF Method

Recently, Lee et al. (2020) proposed a new DIF detection method involving triplet MFC measures at the block-level based on the Thurstonian IRT (TIRT) model. Their work showed the efficacy of the proposed MFC DIF method through various Monte Carlo simulation conditions and an empirical demonstration. This MFC DIF method can be applied to test the fake-resistance of MFC measures compared to Likert-type measures through the within-subject experimental design (e.g., honest and faking conditions). However, DIF results based on chi-square significance statistics have been criticized due to the sensitivity to sample size and their minor practical implications (Drasgow et al., 2018; Meade et al., 2012; Stark et al., 2004). Nye and Drasgow (2011) suggested that the statistical significance DIF test "does not address the practical importance of observed differences between groups and does not provide users with information about the effects of nonequivalence on the organizational outcomes of an assessment" (p. 966). To better understand the size of DIF, Lee et al. (2020) proposed the DIF effect size of the MFC measure by adapting Nye's (2011) DIF effect size.

Furthermore, from a practical perspective, "DTF is the primary concern for organizations because selection decisions are based on total test scores rather than individual items" (Stark et al., 2004, p. 498). Lee et al. (2020) also proposed DTF effect sizes of MFC measures by adopting the method used by Stark et al. (2004). The measurement invariance of MFC measure can be evaluated at both the item and test level by applying these methods. If the MFC measure yields fewer DIF items and smaller DIF effect sizes as well as DTF effect sizes between honest and faking conditions, it could serve as further empirical evidence that the MFC measure may be more fake resistant than a Likert-type measure.

The Present Study

This study aims to (a) investigate the measurement equivalence of MFC and Likert-type personality measures between honest and faking conditions; (b) evaluate how DIF occurs differently between the two measures; and (c) determine which measure produces smaller DIF and DTF effect sizes. To achieve this, four research questions (RQs) are proposed:

RQ1: How many items/blocks exhibit DIF in MFC and Likert-type measures?

RQ2: How differently does DIF occur between MFC and Likert-type measures?

RQ3: How do DIF effect sizes differ across MFC and Likert-type measures?

RQ4: How do DTF effect sizes differ across MFC and Likert-type measures?

METHOD

Research Measure and Sample

This study uses the same Big Five personality MFC triplet measure and Likert-type measure as Lee et al. (2018). The measure comprises 12 statements per dimension, and positively and negatively keyed statements (e.g., 8 positively and 4 negatively keyed statements per dimension). These were mixed to enhance trait score estimation accuracy as recommended by Brown and Maydeu-Olivares (2011).

For data collection, the within-subject design was used. In Korea, 537 college students answered the 20-triplet MFC personality measure and the corresponding Likert-type measure (i.e., the same 60 statements in 20 triplets) under honest responding instructions. Two weeks later, 460 participants among them participated in the faking condition. Under the honest instruction, participants were notified that the results would be used only for research purposes and were requested to answer as honestly as possible. Under the faking instruction, respondents were requested to imagine that they were applying for their dream job in a personnel selection process (e.g., Mueller-Hanson et al., 2006). Four hundred seventeen students completely answered both conditions (50% male/female with an average age of 20.94 years), thereby creating the data analyzed in this study. Because two MFC blocks (all positively keyed) consistently yielded very large residual variances, which caused estimation problems for DIF analysis, they were removed. The remaining 18 blocks were used for subsequent MFC DIF analyses. The same single statements were used for Likert-type measures. The items for the MFC and Likert-type measures are presented in Tables 2 and 3.

Analytical Strategy

For the MFC DIF test, the TIRT model was applied (Brown & Maydeu-Olivares, 2011) as well as the TIRT DIF method (Lee et al., 2020). For the DIF test of the Likert-type measure, categorical MACS DIF method was applied.

In the TIRT model for triplet measure, rank response data were transformed into three sets of binary outcomes (i.e., comparison between the first and the second statements (y_{i1i2}); comparison between the second and the third statements (y_{i1i3}); comparison between the second and the third statements (y_{i2i3})). The transformed binary outcomes were then modeled and analyzed with a two-dimensional standard normal ogive IRT model, as described in detail by Brown and Maydeu-Olivares (2011).

In practical settings, it is generally impossible to know in advance which blocks are free from DIF and which are suitable anchors for free baseline DIF tests. Thus, a *sequential free baseline approach* was applied for TIRT DIF detection and categorical MACS DIF detection (Lee et al., 2020). The sequential free baseline approach has been found effective in detecting DIF with low Type I error and high power in simulation studies (Chun et al., 2016; Kim et al., 2016; Lopez et al., 2009; Meade & Wright, 2012). The Appendix further describes the details of the sequential free baseline approach for MFC DIF and categorical MACS DIF methods; and the Supplemental Materials present Mplus examples of the MFC DIF and categorical MACS DIF methods.

Last, the DIF effect size was calculated to further investigate the identified DIF items of the MFC and Likert-type measures by adapting Nye's (2011) method. Furthermore, the DTF effect sizes for the MFC and Likert-type personality measure were computed by adapting Stark et al.'s method (2004). The effect sizes can be interpreted as Cohen's *d* (0.2, 0.5, and 0.8 for small, medium, and large, respectively). The Appendix also shows the detailed description of DIF and DTF effect sizes.

A direct comparison of DIF results between MFC and Likert-type measures is difficult because MFC DIF is tested at the block level, whereas Likert-type DIF is tested at the single-statement level. Thus, the Likert-type measure was considered a baseline to evaluate how single-statement items in the Likert-type measure function differently when presented in the MFC measure. Also, this study more relied on describing how DIFs differently occurs and evaluating the DIF and DTF effect sizes rather than simply comparing the number of detected DIF items.

RESULTS

Table 1 presents descriptive statistics between Likert-type and MFC personality measures across honest and faking conditions. We note that MFC data were scored using the classical test scoring method (in Table 1). The classical test scoring for MFC measures is still being commonly used in research and practical settings (e.g., Bowen et al., 2002; Converse et al., 2008; Fisher et al., 2019; Heggestad et al., 2006; Jackson et al., 2000; Martin et al., 2002; O'Neill et al., 2018; Vasilopoulos et al., 2006). Although there are different approaches to obtain classical test scoring for MFC measures (Salgado & Lado, 2018), we chose the "inverse scoring" method. If a positively keyed statement is chosen as *most like me* or a negatively keyed statement is chosen as *least like me*, two points were assigned to the statement. In contrast, if a positively keyed statement was selected as *least like me* or a negatively keyed statement was selected as *most like me*, zero points were assigned. The second-ranked statements are scored as one point. Overall, smaller effect sizes (i.e., Cohen's *d*) were found for the MFC measure than the Likert-type measure across Big Five personality traits (*d* = 0.54 vs. 0.36 for agreeableness; *d* = 0.39 vs. -0.10 for openness; *d* = 1.05 vs. 0.92 for conscientiousness; *d* = 0.73 vs. 0.60 for extraversion; *d* = -0.68 vs. -0.59). A preliminary analysis also tested whether the same five personality constructs were measured between two different instruction conditions. The configural invariance was tested, and both measures satisfied the configural invariance between honest and faking conditions (RMSEA = .06, CFI = .89, and TLI = .88 for the Likert-type measure; RMSEA = .03, CFI = .91, and TLI = .91 for the MFC measure).

RQ1: How many items exhibit DIF in Likert and MFC measures?

Table 2 shows the DIF analysis results for the Likert-type measures. Based on the Bonferroni corrected alpha, 15 out of 54 items were classified as DIF items. More specifically, two items (items 3 and 25) were identified as DIF for conscientiousness; three items (items 5, 14, and 51) for extraversion; one item (item 28) for agreeableness; four items (items 18, 39, 40, and 46) for openness; and five items (items 16, 26, 42, 45, and 50) for neuroticism. Table 3 shows the DIF analysis results for the MFC measure using both a nominal alpha level and a Bonferroni-corrected alpha level. Interestingly, when single-statements were constructed as MFC blocks, only one (i.e., block 11) was flagged as DIF based on the Bonferroni-corrected alpha. In sum, 15 items were identified as DIF in the Likert-type measures, whereas only one MFC block was identified as DIF when formed as a triplet MFC block (RQ1).

RQ2. How differently does DIF occur between two measures?

Tables 2 and 3 show that fewer DIF items occurred when statements were formed as MFC blocks rather than when they were presented as a single statement in the Likert-type measure. As an example, items 16 and 18 in the Likert-type measure were detected as DIF items, but the MFC block 6 (corresponding with items 16, 17, and 18 in the Likert-type measure) was identified as a non-DIF block. Figure 1 shows the item characteristic curves (ICC) for items 16, 17, and 18 in the Likert-type measure across the honest and faking conditions. Items 16 and 18 were identified as DIF favoring for the faking condition and the DIF effect sizes were 0.30 (small to medium DIF) for item 16 and 1.46 (large DIF) for item 18. In contrast, when items 16, 17, and 18 were formed in an MFC triplet (block 6), the DIF effect was substantially reduced. Figure 2 shows the item response surfaces of three different binary outcomes (i.e., y_{i16i17}, y_{i16i18}, and y_{i17i18}) that yielded very similar curves for the triplet block. Importantly, the DIF effect sizes of three binary outcomes were negligible (0.09, 0.06, and 0.09, respectively), and the average block effect size was 0.08. Although item 18 (i.e., Am not interested in abstract ideas) in the Likert-type measure showed a very large DIF effect size (d_{DIF} = 1.46), the effect size of the binary outcome associated with this statement was substantially decreased in MFC block 6. That is, d_{DIF} decreased to 0.06 when the third statement (Am not interested in abstract ideas) was compared with the first statement (i.e., Fear for the worst) within the block. Also, d_{DIF} decreased to 0.09 when the third statement was compared with the second statement (i.e., Keep in the background). Similar patterns were also found in other cases.

Interestingly, we found non-DIF statements (e.g., items 31, 32, and 33) in the Likert-type measure became a DIF block (e.g., block 11) when they were formed as a block in the MFC measure. Figure 3 shows quite similar ICCs of items 31, 32, and 33 (in the Likert-type measure) with small effect sizes (d_{DIF} = 0.27, 0.12, and 0.15) between the honest and faking conditions. However, when the same statements were used in MFC block 11, binary outcomes of the block yielded significant DIF. The y_{i31i32} and y_{i31i33} in Figure 4 show very different item response surfaces. Particularly, the direction of loading in conscientiousness changed from the honest to the faking condition as they were compared with other statements measuring extraversion (y_{i31i32}) and agreeableness (y_{i31i33}). It may occur any unexpected interactions of surrounding statements within a block. We examined statement endorsement proportions of binary comparison outcomes and found that endorsement proportions of three statements (A. Waste my time; B. Find it difficult to approaches others; C. Trust what people say) were equally distributed in the honest condition (56.8% vs. 43.2% for the comparison between statements A and B; 51% vs. 49% for the comparison between statements A and C; 41% vs. 59% for the comparison between statements B and C). However, the endorsement proportions substantially changed when the positive statement was compared to negative statement within a block (42% vs. 58% for the comparison between statements A and B; 16% vs. 84% for the comparison between statements A and C; 19% vs. 81% for the comparison between statements B and C). We suspect "desirability-induced response biases" occurred in this case.

RQ3: How do DIF effect size differ across MFC and Likert-type measures?

Tables 2 and 3 generally show that larger DIF effect sizes were found in the Likert-type measures (*M* = 0.27, range = [0.00 – 1.46]) compared to the MFC measures (*M* = 0.18, range = [0.00 – 0.91]). Overall, this finding indicates that MFC measures can be a more fake-resistant assessment tool. However, interesting results were also found. When differently keyed statements were compared in a mixed block (i.e., block consisting of positively and negatively keyed statements), the corresponding pairwise comparison still yielded medium to large DIF effect sizes. For example, in the MFC block 3 (i.e., A. Panic easily; B. Do not enjoy going to art museums; C. Know how to captivate people), when the first statement A was compared with the second statement B, the DIF effect size was 0.16. However, when the first statement A and the second statement B were compared with the third statement C, the DIF effect

TABLE 1.

Descriptive Statistics for Likert-type and MFC Measures Across Honest and Fake-Good Conditions

Measure type	Trait	Honest condition group				Fake-good condition group				Cohen's *d*
		Minimum	Maximum	Mean	*SD*	Minimum	Maximum	Mean	*SD*	
Summed score for Likert-type measure	A	22.00	54.00	39.75	5.20	23.00	55.00	42.55	5.15	0.54
	O	17.00	59.00	42.54	7.35	24.00	60.00	45.29	6.68	0.39
	C	12.00	48.00	30.74	6.76	20.00	50.00	38.08	7.26	1.05
	E	11.00	52.00	33.68	7.96	14.00	55.00	39.20	7.20	0.73
	N	11.00	49.00	28.65	7.35	10.00	46.00	23.93	6.46	-0.68
Classical test scoring for MFC measure	A	4.00	22.00	14.40	3.56	4.00	22.00	15.60	3.06	0.36
	O	0.00	23.00	13.44	3.76	3.00	22.00	13.08	3.46	-0.10
	C	0.00	20.00	9.33	4.24	0.00	20.00	13.37	4.56	0.92
	E	0.00	22.00	11.72	5.33	1.00	22.00	14.69	4.43	0.60
	N	1.00	20.00	10.39	4.49	0.00	20.00	7.95	3.75	-0.59

Note. A = Agreeableness; O = Openness; C = Concientiousness; E = Extraversion; N = Neuroticism.

TABLE 2.
Categorical MACS DIF Analysis Results for the Likert-Type Measure

Item	Dim	Statement	Honest Condition					Faking Condition					DIF with .05 alpha	DIF with Bonferroni Corrected	Effect Size
			λ	γ_1	γ_2	γ_3	γ_4	λ	γ_1	γ_2	γ_3	γ_4			
1	A	Respect others.	1.00	-3.36	-2.61	-1.39	1.06	1.00	-2.98	-2.80	-1.78	0.85	Non-DIF	Non-DIF	0.19
2	O	Have a rich vocabulary.	1.00	-2.33	-0.98	0.12	1.68	1.44	-2.59	-1.33	-0.06	1.46	**DIF**	Non-DIF	0.29
3	C	Follow through with my plans.	1.00	-3.33	-0.72	0.71	3.27	0.72	-3.30	-0.95	0.69	2.72	**DIF**	**DIF**	0.42
4	O	Get excited by new ideas.	1.40	-2.89	-1.58	-0.34	0.99	1.84	-3.20	-1.98	-0.56	1.00	Non-DIF	Non-DIF	0.19
5	E	Warm up quickly to others.	1.00	-4.29	-1.46	0.76	3.57	0.62	-2.72	-1.42	0.56	2.57	**DIF**	**DIF**	0.52
6	C	Don't put my mind on the task at hand.	0.47	-2.57	-1.03	0.12	1.95	0.50	-2.28	-1.21	0.19	2.27	Non-DIF	Non-DIF	0.05
7	N	Panic easily.	1.00	-1.79	-0.36	0.49	2.22	1.00	-1.79	-0.36	0.49	2.22	Anchor	Anchor	0.00
8	O	Do not enjoy going to art museums.	1.92	-1.85	-0.82	0.09	1.16	1.92	-1.85	-0.82	0.09	1.16	Anchor	Anchor	0.00
9	E	Know how to captivate people.	0.38	-2.13	-0.62	0.59	2.79	0.42	-3.19	-1.02	0.39	2.56	**DIF**	Non-DIF	0.24
10	E	Am the life of the party.	0.46	-1.88	-0.45	0.96	2.66	0.55	-2.02	-0.40	1.15	3.42	Non-DIF	Non-DIF	0.06
11	A	Cut others to pieces.	1.16	-2.59	-1.36	-0.32	1.11	1.22	-3.00	-1.45	-0.32	1.25	Non-DIF	Non-DIF	0.02
12	N	Am filled with doubts about things.	0.30	-1.86	-0.94	-0.03	1.41	0.40	-1.99	-0.95	0.17	1.63	Non-DIF	Non-DIF	0.18
13	O	Look for a deeper meaning in things.	1.43	-2.26	-1.07	-0.09	1.25	2.02	-2.68	-1.36	-0.16	1.42	**DIF**	Non-DIF	0.15
14	E	Cheer people up.	0.36	-3.11	-1.75	-0.09	2.09	0.52	-3.53	-1.93	0.14	2.45	**DIF**	**DIF**	0.33
15	C	Carry out my plans.	0.82	-3.54	-1.57	0.18	2.34	0.96	-4.54	-1.72	0.56	3.40	**DIF**	Non-DIF	0.27
16	N	Fear for the worst.	0.36	-1.73	-0.75	-0.26	1.15	0.38	-2.25	-1.18	-0.53	1.03	**DIF**	**DIF**	0.30
17	E	Keep in the background.	0.32	-1.32	0.10	0.93	2.07	0.23	-1.43	0.05	0.99	2.10	**DIF**	Non-DIF	0.14
18	O	Am not interested in abstract ideas.	2.14	-1.30	0.32	1.40	2.59	2.08	-2.29	-1.30	-0.07	1.52	**DIF**	**DIF**	1.46
19	A	Have a good word for everyone.	1.17	-2.48	-1.13	0.03	2.12	0.88	-2.47	-1.03	0.05	2.02	Non-DIF	Non-DIF	0.15
20	C	Am exacting in my work.	0.50	-2.36	-0.52	0.46	2.02	0.52	-2.61	-0.80	0.52	2.21	Non-DIF	Non-DIF	0.16
21	O	Have a vivid imagination.	2.41	-3.14	-1.49	-0.23	1.38	2.60	-3.10	-1.67	-0.26	1.52	Non-DIF	Non-DIF	0.12
22	C	Do things according to a plan.	0.83	-3.11	-1.13	0.30	2.51	0.79	-3.96	-1.17	0.60	3.21	**DIF**	Non-DIF	0.29
23	A	Get back at others.	0.60	-1.82	-0.80	0.02	1.32	0.59	-2.05	-1.09	-0.25	1.21	Non-DIF	Non-DIF	0.22
24	E	Feel comfortable around people.	0.29	-2.24	-0.97	0.17	1.65	0.34	-2.27	-1.25	0.20	2.07	**DIF**	Non-DIF	0.10
25	C	Find it difficult to get down to work.	0.77	-2.68	-0.59	0.52	2.40	0.63	-2.50	-0.64	0.82	2.98	**DIF**	**DIF**	0.46
26	N	Am often down in the dumps	1.93	-1.76	0.37	1.49	4.04	1.92	-2.64	-0.29	1.14	3.80	**DIF**	**DIF**	0.54
27	O	Enjoy thinking about things.	2.26	-4.02	-2.13	-0.86	0.75	2.00	-2.99	-2.16	-0.77	1.13	Non-DIF	Non-DIF	0.26

Note. O indicates Openness, C indicates Concientiousness, E indicates Extraversion, A indicates Agreeablenss, and N indicates Neuroticism. Effect sizes are computed for single statements.

TABLE 2 (CONTINUED).

Categorical MACS DIF Analysis Results for the Likert-Type Measure

Item	Dim	Statement	Honest Condition					Faking Condition					DIF with .05 alpha	DIF with Bonferroni Corrected	Effect Size
			λ	γ_1	γ_2	γ_3	γ_4	λ	γ_1	γ_2	γ_3	γ_4			
28	A	Treat all people equally.	0.62	-2.64	-0.90	0.10	1.58	0.82	-2.65	-1.47	-0.29	1.63	**DIF**	**DIF**	0.34
29	E	Would describe my experiences as somewhat dull.	0.22	-1.70	-0.43	0.27	1.61	0.28	-2.09	-0.38	0.59	1.97	**DIF**	Non-DIF	0.11
30	N	Am not easily frustrated.	0.71	-1.40	0.15	1.30	2.38	0.83	-1.68	0.08	1.10	2.15	Non-DIF	Non-DIF	0.11
31	C	Waste my time.	0.50	-1.79	0.00	1.08	2.25	0.57	-2.29	-0.40	0.86	2.37	Non-DIF	Non-DIF	0.27
32	E	Find it difficult to approach others.	0.60	-2.64	-0.66	0.40	2.30	0.60	-3.14	-0.96	0.51	2.72	Non-DIF	Non-DIF	0.12
33	A	Trust what people say.	1.26	-2.75	-1.43	0.01	1.93	1.12	-2.90	-1.65	-0.06	2.07	Non-DIF	Non-DIF	0.15
34	A	Am concerned about others.	0.86	-2.43	-1.26	-0.32	1.55	1.01	-2.67	-1.88	-0.48	1.54	**DIF**	Non-DIF	0.29
35	O	Believe in the importance of art.	2.95	-3.04	-1.97	-0.66	1.15	3.40	-3.02	-1.83	-0.28	1.67	**DIF**	Non-DIF	0.34
36	N	Rarely lose my composure.	0.90	-1.59	-0.05	1.20	3.02	0.79	-2.09	-0.20	1.19	2.85	**DIF**	Non-DIF	0.33
37	A	Have a sharp tongue.	1.49	-3.28	-1.80	-0.69	0.86	1.49	-3.28	-1.80	-0.69	0.86	Anchor	Anchor	0.00
38	N	Get stressed out easily.	1.15	-2.33	-0.73	0.20	1.93	1.23	-2.81	-0.93	0.14	1.74	Non-DIF	Non-DIF	0.15
39	O	Carry the conversation to a higher level.	1.40	-1.81	-0.56	0.65	2.00	1.74	-2.03	-0.95	0.25	1.67	**DIF**	**DIF**	0.42
40	O	Am not interested in theoretical discussions.	1.17	-1.85	-0.76	0.02	1.29	1.91	-2.63	-1.43	-0.38	1.41	**DIF**	**DIF**	0.60
41	E	Talk to a lot of different people at parties.	0.66	-2.62	-0.70	0.97	3.06	0.66	-2.62	-0.70	0.97	3.06	Anchor	Anchor	0.00
42	N	Have frequent mood swings.	0.90	-1.55	-0.43	0.30	1.59	1.22	-2.29	-0.78	0.25	2.04	**DIF**	**DIF**	0.31
43	C	Finish what l start.	0.66	-3.28	-1.42	0.06	2.12	0.66	-3.28	-1.42	0.06	2.12	Anchor	Anchor	0.00
44	A	Contradict others.	1.11	-3.48	-2.38	-1.11	0.70	1.38	-3.70	-2.78	-1.50	0.73	Non-DIF	Non-DIF	0.13
45	N	Feel threatened easily.	1.11	-1.16	0.33	1.47	2.99	1.30	-2.00	0.05	1.40	3.02	**DIF**	**DIF**	0.32
46	O	Enjoy wild flights of fantasy.	2.70	-3.44	-1.87	-0.67	1.08	2.62	-2.57	-1.49	-0.30	1.61	**DIF**	**DIF**	0.41
47	C	Get chores done right away.	0.74	-2.37	-0.13	1.14	2.70	0.77	-2.79	-0.71	1.19	3.09	**DIF**	Non-DIF	0.18
48	A	Sympathize with others' feelings.	0.80	-2.64	-1.73	-0.70	1.05	1.03	-3.00	-2.22	-0.78	1.12	Non-DIF	Non-DIF	0.10
49	A	Believe that others have good intentions.	1.26	-2.41	-1.27	0.11	1.71	1.14	-2.57	-1.56	-0.03	1.68	Non-DIF	Non-DIF	0.16
50	N	Often feel blue.	1.90	-2.28	-0.17	1.17	3.65	2.69	-4.27	-0.78	1.21	4.49	**DIF**	**DIF**	0.73
51	E	Make friends easily.	0.90	-3.72	-1.46	0.86	3.63	0.68	-2.88	-1.07	0.97	3.51	**DIF**	**DIF**	0.47
52	O	Do not like art.	3.52	-3.63	-2.09	-1.05	0.93	3.83	-3.36	-2.22	-0.76	1.42	Non-DIF	Non-DIF	0.28
53	E	Start conversations.	0.56	-3.15	-1.40	0.27	2.85	0.62	-2.74	-1.14	0.69	3.30	Non-DIF	Non-DIF	0.26
54	C	Need a push to get started.	0.49	-2.01	-0.57	0.34	1.64	0.49	-2.16	-0.41	0.70	1.99	Non-DIF	Non-DIF	0.23

Note. O indicates Openness, C indicates Concientiousness, E indicates Extraversion, A indicates Agreeablenss, and N indicates Neuroticism. Effect sizes are computed for single statements.

TABLE 3.
TIRT DIF Analysis Results for the Triplet MFC Measure

Block	Dim	Statement	Honest Condition		Faking Condition		DIF with .05 alpha	DIF with Bonferroni Corrected	Effect Size of Pairwise Comparison	Block-level Effect Size
			λ	γ	λ	γ				
1	A	Respect others.	0.48	-2.00	0.48	-2.00	Anchor Block	Anchor Block	0.00	0.00
	O	Have a rich vocabulary.	0.99	-2.59	0.99	-2.59			0.00	
	C	Follow through with my plans.	1.85	-0.70	1.85	-0.70			0.00	
2	O	Get excited by new ideas.	0.52	-0.38	0.63	-0.49	Non-DIF	Non-DIF	0.13	0.14
	E	Warm up quickly to others.	0.95	-0.61	0.79	-0.83			0.15	
	C	Don't put my mind on the task at hand.	-0.53	-0.26	-0.39	-0.11			0.13	
3	N	Panic easily.	1.07	-0.22	1.11	0.10	Non-DIF	Non-DIF	0.16	0.29
	O	Do not enjoy going to art museums.	-0.54	0.11	-0.60	0.44			0.45	
	E	Know how to captivate people.	0.78	0.20	1.04	0.28			0.26	
4	E	Am the life of the party.	1.49	-0.62	1.80	-0.34	Non-DIF	Non-DIF	0.21	0.15
	A	Cut others to pieces.	-0.27	0.88	-0.14	1.13			0.19	
	N	Am filled with doubts about things.	0.20	1.47	0.48	1.58			0.06	
5	O	Look for a deeper meaning in things.	0.57	0.14	0.51	-0.05	Non-DIF	Non-DIF	0.11	0.09
	E	Cheer people up.	0.59	-0.24	0.72	-0.25			0.04	
	C	Carry out my plans.	0.76	-0.34	0.77	-0.16			0.11	
6	N	Fear for the worst.	0.32	0.11	0.21	0.19	Non-DIF	Non-DIF	0.09	0.08
	E	Keep in the background.	-0.59	-0.61	-0.73	-0.68			0.06	
	O	Am not interested in abstract ideas.	-0.32	-0.78	-0.28	-0.98			0.09	
7	A	Have a good word for everyone.	0.78	-0.71	0.23	-1.00	**DIF**	Non-DIF	0.23	0.26
	C	Am exacting in my work.	1.03	0.17	1.33	0.20			0.37	
	O	Have a vivid imagination.	0.81	0.86	1.13	1.00			0.19	
8	C	Do things according to a plan.	1.07	-1.58	1.07	-1.58	Anchor Block	Anchor Block	0.00	0.00
	A	Get back at others.	-0.87	0.13	-0.87	0.13			0.00	
	E	Feel comfortable around people.	0.91	1.72	0.91	1.72			0.00	
9	C	Find it difficult to get down to work.	-1.04	-1.04	-1.17	-1.37	Non-DIF	Non-DIF	0.06	0.32
	N	Am often down in the dumps	1.41	1.12	0.92	1.21			0.36	
	O	Enjoy thinking about things.	0.97	2.40	1.03	2.54			0.54	

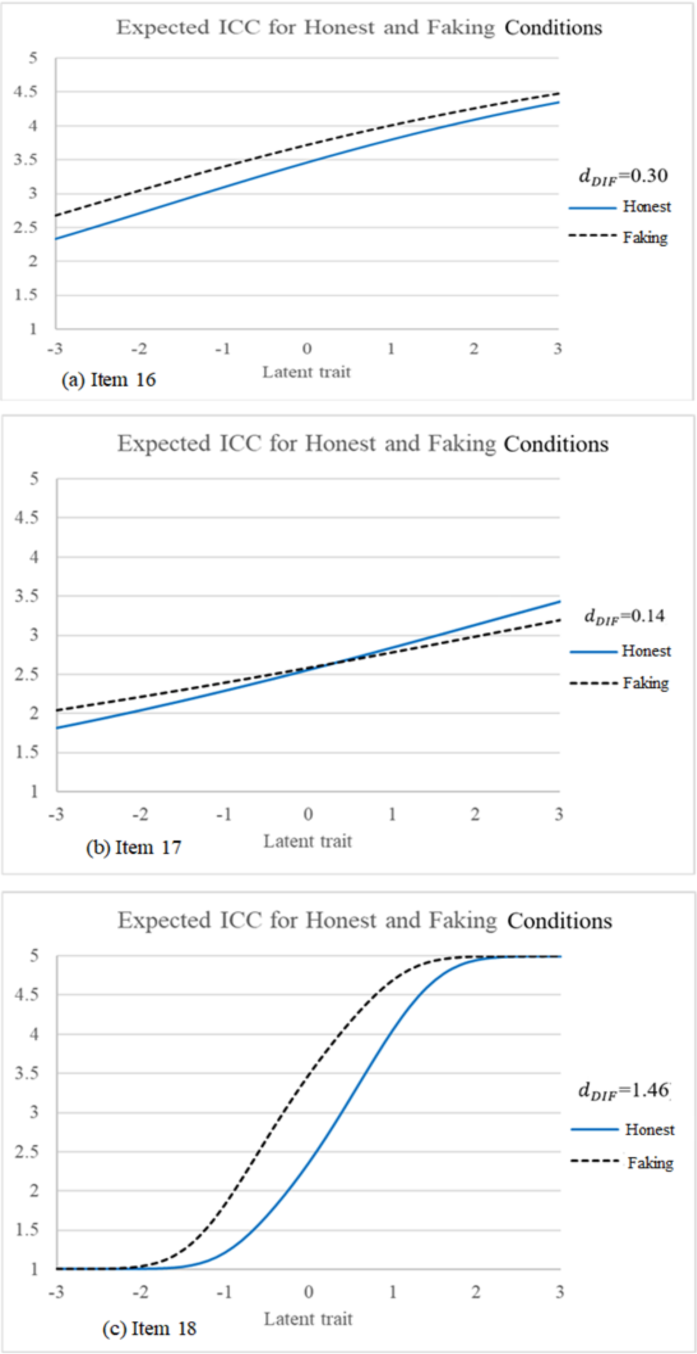
Note. O indicates Openness, C indicates Concientiousness, E indicates Extraversion, A indicates Agreeablenss, and N indicates Neuroticism. Effect sizes are computed for pairwise binary outcome.

TABLE 3 (CONTINUED).
TIRT DIF Analysis Results for the Triplet MFC Measure

Block	Dim	Statement	Honest Condition		Faking Condition		DIF with .05 alpha	DIF with Bonferroni Corrected	Effect Size of Pairwise Comparison	Block-level Effect Size
			λ	γ	λ	γ				
10	A	Treat all people equally.	0.27	-0.27	0.15	-0.25	Non-DIF	Non-DIF	0.08	0.10
	E	Would describe my experiences as somewhat dull.	-0.59	0.10	-0.79	0.06			0.08	
	N	Am not easily frustrated.	-0.59	0.36	-0.64	0.32			0.13	
11	C	Waste my time.	1.06	-0.38	-0.84	0.02	**DIF**	**DIF**	0.88	0.69
	E	Find it difficult to approach others.	-1.53	-0.06	-0.90	0.19			0.91	
	A	Trust what people say.	1.02	0.57	0.77	0.08			0.27	
12	A	Am concerned about others.	0.68	-0.07	0.68	-0.07	Anchor Block	Anchor Block	0.00	0.00
	O	Believe in the importance of art.	0.96	-0.06	0.96	-0.06			0.00	
	N	Rarely lose my composure.	-1.08	0.07	-1.08	0.07			0.00	
13	A	Have a sharp tongue.	-0.73	1.39	-0.84	0.98	Non-DIF	Non-DIF	0.24	0.17
	N	Get stressed out easily.	1.24	1.06	1.36	0.87			0.12	
	O	Carry the conversation to a higher level.	0.37	-0.43	0.42	-0.16			0.17	
14	O	Am not interested in theoretical discussions.	-0.54	0.03	-1.12	0.11	Non-DIF	Non-DIF	0.24	0.25
	E	Talk to a lot of different people at parties.	1.17	0.50	1.53	0.74			0.39	
	N	Have frequent mood swings.	1.35	0.37	1.39	0.45			0.11	
15	C	Finish what I start.	0.77	-1.57	0.73	-1.98	Non-DIF	Non-DIF	0.16	0.18
	A	Contradict others.	-0.45	-0.86	-0.45	-1.18			0.27	
	N	Feel threatened easily.	0.91	0.85	0.89	0.64			0.11	
16	O	Enjoy wild flights of fantasy.	1.25	-1.56	1.26	-1.45	Non-DIF	Non-DIF	0.16	0.13
	C	Get chores done right away.	1.55	0.02	1.76	-0.09			0.11	
	A	Sympathize with others' feelings.	0.87	1.45	0.74	1.09			0.12	
17	A	Believe that others have good intentions.	0.62	-0.72	0.58	-0.82	Non-DIF	Non-DIF	0.05	0.10
	N	Often feel blue.	1.28	-0.37	1.29	-0.32			0.05	
	E	Make friends easily.	1.12	0.26	0.82	0.15			0.19	
18	O	Do not like art.	-1.14	1.94	-1.11	1.41	Non-DIF	Non-DIF	0.27	0.34
	E	Start conversations.	1.67	1.01	1.71	1.31			0.27	
	C	Need a push to get started.	-1.47	-0.60	-1.33	0.23			0.46	

Note. O indicates Openness, C indicates Concientiousness, E indicates Extraversion, A indicates Agreeablenss, and N indicates Neuroticism. Effect sizes are computed for pairwise binary outcome.

FIGURE 1.
Expected Item Characteristic Curves for Item 16, 17, and 18 of Likert-Type Measures



Note. Item 16 is a DIF item; Item 17 is a non-DIF item; Item 18 is a DIF item.

sizes increased to 0.45 and 0.26, respectively. Similar patterns were found for MFC block 9 (A. Find it difficult to get down to work; B. Am often down in the dumps; C. Enjoy thinking about things). The pairwise comparisons yielded much larger DIF effect sizes when statement C was compared to statements A and B (d_{DIF}= 0.36 and 0.54) than when statement B was compared to just statement A (d_{DIF}= 0.06). We found five blocks yielded block-level DIF effect sizes ranging from 0.2 to 0.3, and one block yielded a medium effect size of 0.69, with all of them being mixed blocks. Overall, these results show the MFC measure generally yields smaller DIF effect sizes than the Likert-type measure. However, DIF still can occur when positively and negatively keyed statements are mixed in the same MFC block.

RQ4: How do DTF effect sizes differ across two measures?

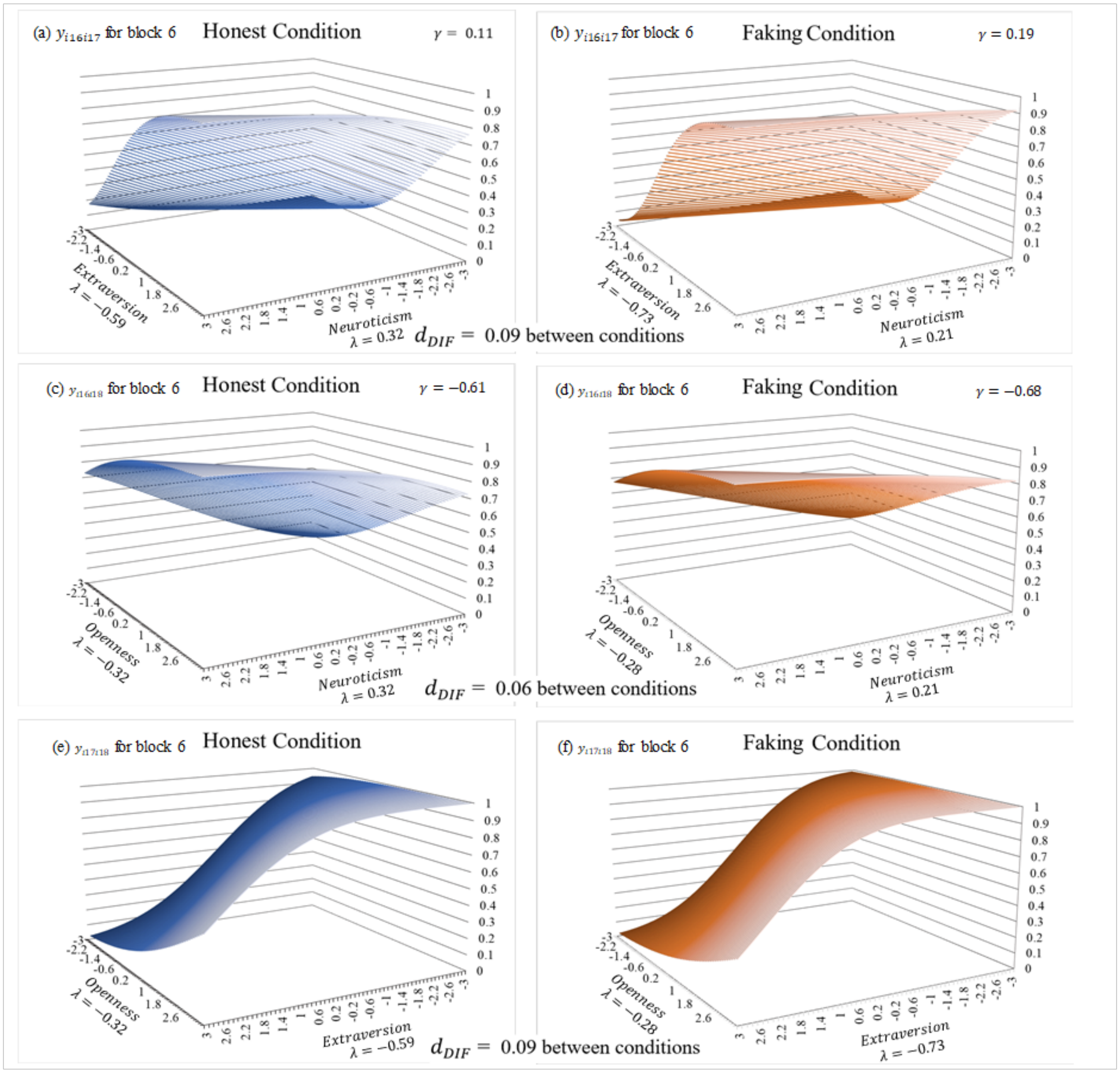
To examine the practical importance of measurement invariance at the test level, this study computed overall DTF effect sizes for MFC and Likert-type measures across five dimensions. d_{DTF} was -0.08 for the MFC measure, but d_{DTF} was -0.48 for the Likert-type measure. At the test level, MFC measures yielded a minimal test bias between test conditions, whereas the Likert-type measure produced a moderate level of test bias favoring in the faking condition.

DISCUSSION

This research employed the *changing items paradigm* to evaluate the differential nature of item responses between MFC and Likert-type measures under honest and faking conditions. The main findings are as follows. First, fewer DIF occurred when statements were presented as an MFC block compared to a single statement in the Likert-type measures. Based on the Bonferroni correction, only one MFC block was identified as DIF for the MFC measure, whereas 15 items (i.e., statements) were detected as DIF for the Likert-type measure (RQ1). Second, when single-statements in the Likert-type measure are used to make an MFC item, the same statements do not always show the same DIF results in both formats. Importantly, non-DIF items in the Likert-type measure also do not guarantee item invariance in the MFC measure between the honest and faking conditions (RQ2). Third, lower DIF effect sizes were generally found for the MFC measure than the Likert-type measure. However, pairwise comparisons involving positively and negatively keyed statements still present small to medium DIF effect sizes in MFC blocks (RQ3). Last, a much lower overall DTF effect size was found for the MFC measure than the Likert-type measure (RQ4). Taken together, the measurement invariance between test conditions can be better established in the MFC measure, which empirically supports that MFC measures could be more fake resistant than Likert-type measures.

FIGURE 2.

Item Response Surface for MFC Block 6 (Non-DIF)



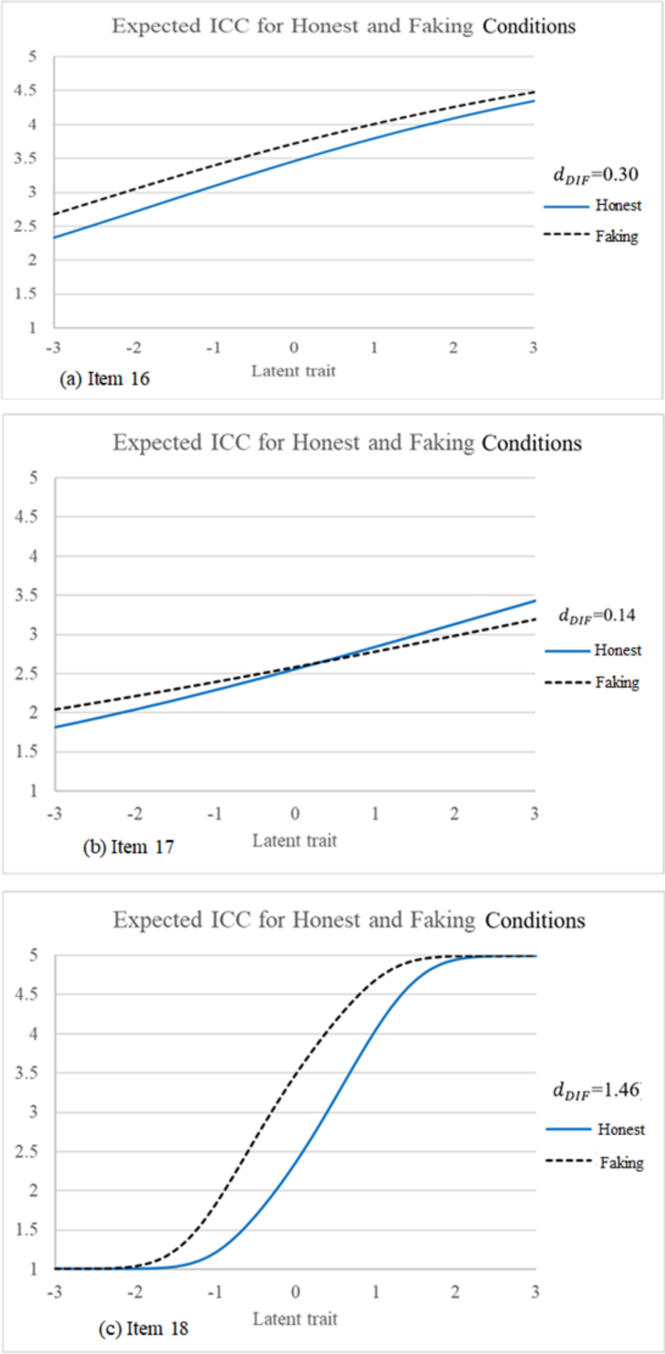
Note. Block 6 is a Non-DIF block. The d_{DIF} = 0.09, 0.06, and 0.09 for y_{i16i17}, y_{i16i18}, and y_{i17i18}, respectively. The horizontal axes represent the dimensions associated with the statements in the respective comparisons, and the vertical axis represents the probability of preferring the former statement to the latter in each instance. (a) and (b) are response surfaces for y_{i16i17} across honest and faking conditions; (c) and (d) are response surfaces for y_{i16i18} across honest and faking conditions; (e) and (f) are response surfaces for y_{i17i18} across honest and faking conditions. λ and γ represent factor loading and thresholds.

Contributions to Faking Research on MFC Measures

This research provides important contributions to the personality faking research on MFC measures. Previous studies on the fake resistance of MFC measures mainly relied on changing person paradigm by evaluating correlations of scorings or scale mean differences between honest and faking conditions. However, to establish a meaningful scoring comparison between the test conditions, it is essential that items or tests should provide an equivalent measurement across test conditions (Nye &

FIGURE 3.

Expected Item Characteristic Curves for Item 31, 32, and 33 of Likert-Type Measures



Note. Item 31, 32, 33 are non-DIF items.

Drasgow, 2011).

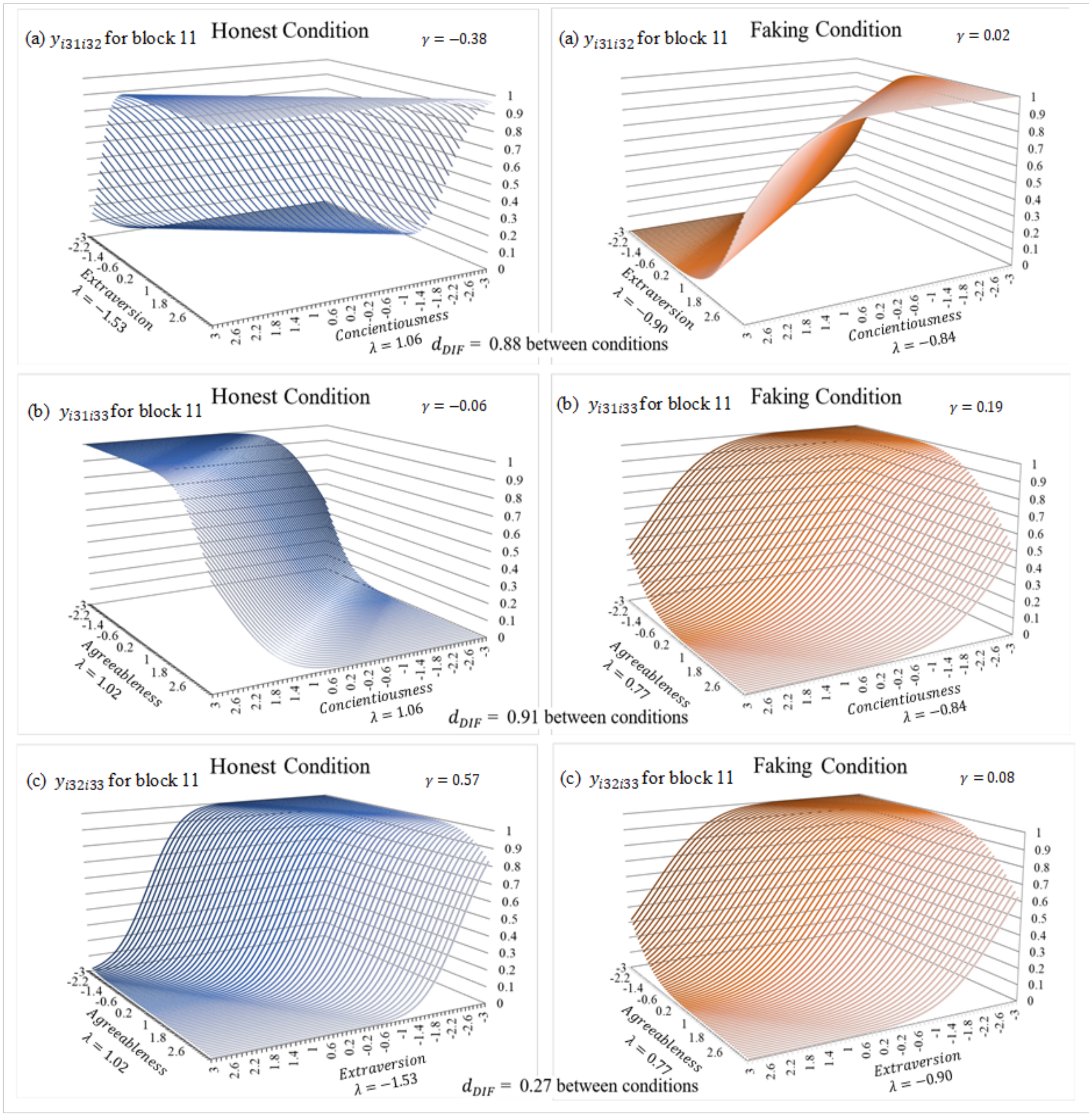
Recently, Pavlov et al. (2019) pointed out the measurement invariance issue of MFC measures in their research. They introduced a regression-based moderation framework to model faking effects and investigated the scorings from MFC and Likert-type measures. They first estimated item parameters of MFC measures from the honest sample, then scored latent traits of the faking sample using the item parameters obtained from the honest sample. To this end, they "assumed measurement invariance across experimental conditions to ensure comparability of scores" (Pavlov et al., 2019, p. 720). However, if the measurement invariance between honest and faking conditions is not satisfied, scores in the faking condition could be biased because the scores were obtained using variant item parameters from the honest sample. If that happens, research findings would not be tenable. In this vein, Pavlov et al. (2019) pointed out that "future studies are advised to more firmly establish the psychometric equivalence of the applied measures to optimize investigation of the forced-choice format as a faking mitigation strategy" (p. 732). The good news is that our results can be served as empirical evidence of measurement invariance between the test conditions and support previous faking research focusing on scoring comparison of MFC measures without testing item invariances (e.g., Pavlov et al., 2019).

Next, this research scored MFC response data using the TIRT model. Many studies examining the fake resistance of MFC measures generally relied on the classical scoring method (e.g., Martin et al., 2002; Converse et al., 2008; Fisher et al., 2019; Heggestad et al., 2006; Jackson et al., 2000; O'Neill et al., 2017; Vasilopoulos et al., 2006). Fisher and colleagues (2019) recently showed classical test scoring can be more valid than IRT-based scoring for MFC measures. Despite the wide use and interests of classical scoring in the organizational or research settings, this method has been criticized by applied psychometricians because it does not represent a comparative judgment process of selecting statements within a block (e.g., Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015; Stark et al., 2012). By applying a model-based MFC IRT method and a newly developed DIF method for MFC measures, this study was able to evaluate a more accurate response process in MFC data (e.g., binary paired comparison between statements in a block) and evaluated measurement invariance at both the item level and the test level.

Last, this study not only examined the differential functioning of MFC and Likert-type measures at the item level but also investigated DTF effect sizes of the two formats at the test level. From an organizational perspective, hiring decisions are generally made based on test scores rather than individual item scores (Stark et al., 2004). This study showed that there was little test-level bias for the MFC measure, but there was a moderate-level of test bias for the Likert-type measure. This result confirms that the MFC measure could be more effective to reduce faking at the test level than the Likert-type measure.

FIGURE 4.

Item Response Surface for MFC Block 11 (DIF) d_{DIF} = 0.27 Between Conditions



Note. Block 11 is a DIF block. The d_{DIF} = 0.88, 0.91, and 0.27 for y_{i31i32}, y_{i31i33}, and y_{i32i33}, respectively. The horizontal axes represent the dimensions associated with the statements in the respective comparisons, and the vertical axis represents the probability of preferring the former statement to the latter in each instance. (a) and (b) are response surfaces for y_{i31i32} across honest and faking conditions; (c) and (d) are response surfaces for y_{i31i33} across honest and faking conditions; (e) and (f) are response surfaces for y_{i32i33} across honest and faking conditions. λ and γ represent factor loading and thresholds.

Practical Implications

Our study provides important practical implications for the development of MFC measures. A common practice for constructing MFC measures begins with developing single statements item pools, evaluating item invariance of single statements (via DIF analysis for single-statement items), and removing any problematic DIF items from the item pools. Then, researchers and practitioners construct MFC item blocks by pairing non-DIF single-statements based on the social desirability. In this process, measurement invariance between single-statement items and MFC items is generally assumed without testing differential item functioning of MFC measures between different test conditions (Morillo et al., 2019). However, this research shows that a combination of non-DIF single statements in the item pool do not necessarily guarantee item invariance between single statements and MFC blocks. In the test development, we recommend researchers and practitioners conduct MFC DIF tests and ensure whether MFC blocks still achieve measurement invariance.

Although this research shows the MFC measure better holds measurement invariance than the Likert-type measure across the test conditions, it is important to note that DIF still can occur depending on the combination of statements in the MFC block. This is particularly pronounced when statements with a positive and a negative meaning are compared in the same MFC block. Thus, MFC measures may still be susceptible to faking if MFC measures include many mixed blocks consisting of positively and negatively keyed statements (within a block). We examined statement endorsement proportions within each MFC block in the honest condition to investigate whether self-enhancement bias could occur in honest MFC responses. We found almost 30% (i.e., 16 out of 36 binary outcomes) of pairwise comparison involved unequal endorsement (e.g., at least 10% difference) favoring more desirable items. For example, for block 8 (A: Do things according to a plan; B: Get back at others; C: Feel comfortable around people), a much lower endorsement proportion of the B statement was found when it was compared to the A statement (30% vs. 70%) and the C statement (27.6% vs. 72.4%). These findings indicate that participants even in the honest condition may tend to strongly avoid a statement apparently measuring negative personality traits. Thus, self-enhancement bias may still occur in the honest research context or low-stakes setting.

Following Brown and Maydeu-Olivares' (2011) suggestion, many studies developed MFC measures by mixing positively and negatively keyed statements to improve the accuracy of scoring in the TIRT model (e.g., Bürkner et al., 2019; Lee et al., 2018; Ng et al., 2021; Wetzel & Frick, 2020). Although the recommendation of including negatively keyed statements may improve the scoring accuracy of MFC measures, several researchers raised a question if a mixed block can harm the original purpose of MFC, which is faking resistance (e.g., Bürkner et al., 2019; Fisher et al., 2019; Lin & Brown, 2017; Ng et al., 2021; Wang et al., 2017). It may be necessary for future research to find an optimal strategy to design mixed blocks in the MFC measures to satisfy the goals of validity and scoring accuracy (e.g., how many mixed blocks are needed? how to create effective mixed blocks?).

Limitations

This research has several limitations. First, this study used student samples in the experimental settings rather than job applicant samples from real organizations. Future research could examine whether the results of this study can be generalized in real personnel selection settings. Second, this study used a somewhat unclear instruction for the faking test condition. Respondents were asked to imagine their "dream job." However, as an anonymous reviewer pointed out, this method could be problematic in faking research because faking can be differently emerged depending on job types. Future research could provide respondents with more specific job instructions or could use real job applicants engaged in a real selection process. Third, this study used a MFC measure developed only for the research purpose (not developed for the personnel selection purpose). Future research could verify this study's results by using a more elaborately developed personnel selection purpose. Last, Lee et al. (2020) showed the TIRT DIF method was effective for detecting DIF blocks with the large DIF size under n = 500 condition and the type I errors were well-controlled. However, the DIF tests were substantially underpowered in the small DIF size condition. This study's sample size of n = 417 may be too small to detect DIF blocks with small DIF sizes. Although an evaluation of DIF and DTF effect sizes was more considered rather than statistical significance DIF test results in our study, future research should conduct a measurement invariance using a larger sample to achieve good power even in small DIF cases.

Conclusions

In sum, MFC measures have been widely applied in noncognitive assessments in industrial and organizational psychology and education (Burrus et al., 2012). Overall, we supported measurement invariance of MFC measures (compared to Likert-type measures) at the item and test level between honest and faking conditions via advanced IRT methodology. However, we do not argue the MFC format itself is essentially more fake resistant than Likert-type measures. As noted by Griffith and Robie (2013), "forced-choice measures of personality may both reduce faking and attain adequate levels of predictive validity if *properly developed*" (p. 272). We hope that practitioners and researchers ensure the quality of MFC items by testing test measurement invariance and properly developing more fake-resistant MFC noncognitive assessment for various industrial and organizational settings.

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.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? International Journal of Selection and Assessment, 9, 9-30.

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

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, 317-335.

Bott, J. P., O'Connell, M. S., Ramakrishnan, M., & Doverspike, D. (2007). Practical limitations in making decisions regarding the distribution of applicant personality test scores based on incumbent data. Journal of Business and Psychology, 22, 123-134.

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.

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

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.

Burrus, J., Naemi, B., & Kyllonen, P. C. (2012). Intentional and unintentional faking in education. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), New perspective on faking in personality assessment, (pp. 282–306). Oxford University Press.

Camilli, G., Shepard, L. A., & Shepard, L. (1994). Methods for identifying biased test items (Vol. 4). Sage.

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.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. Human Performance, 18(3), 267-307.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. Applied Psychological Measurement, 40, 486-499.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. Journal of Applied Psychology, 85, 678-707.

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

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., pp. 577-635). Consulting Psychologists Press.

Drasgow, F., Nye, C. D., Stark, S., & Chernyshenko, O. S. (2018). Differential item and test functioning. In P. Irwing, T. Booth, & D. Hughes (Eds.), The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (pp. 885-899). Wiley-Blackwell.

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.

Griffin, B., Hesketh, B., & Grayson, D. (2004). Applicants faking good: Evidence of item bias in the NEO PI-R. Personality and Individual Differences, 36, 1545-1558.

Griffith, R. L., & Robie, C. (2013). Personality testing and the "f-word": Revisiting seven questions about faking. In N. Christiansen, & R. Tett (Eds.), Handbook of personality at work (pp. 253–280). Taylor & Francis.

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.

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.

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. Journal of Research in Personality, 42, 1323-1333.

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

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. Industrial and Organizational Psychology: Perspectives on Science and Practice, 1, 272-290.

Hough, L. M., Oswald, F. L., & Ock, J. (2015). Beyond the Big Five: New directions for personality research and practice in organizations. Annual Review of Organizational Psychology and Organizational Behavior, 2, 183-209.

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.

Joo, S. H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. Behavior Research Methods, 52, 761–772.

Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. Structural Equation Modeling: A Multidisciplinary Journal, 23, 870-887.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. Journal of Applied Psychology, 93, 140-154.

Lee, P., Mahoney, K. T., & Lee, S. (2017). An application of the exploratory structural equation modeling framework to the study of personality faking. Personality and Individual Differences, 119, 220-226.

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.

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. (2020). Detecting DIF in multidimensional forced choice measures using the Thurstonian item response theory model. Organizational Research Methods. Advance online publication. https://doi.org/10.1177/1094428120959822

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

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. Applied Psychological Measurement, 33, 251-265.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? Personality and Individual Differences, 32, 247-256.

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.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. Journal of Applied Psychology, 97, 1016-1031.

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsada, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. Journal of Work and Organizational Psychology, 35, 75-83.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. Journal of Applied Psychology, 88, 348-355.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. Psychology Science, 48, 288-312.

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.

Nye, C. D. (2011). The development and validation of effect size measures for IRT and CFA studies of measurement equivalence. Doctoral dissertation, Department of Psychology, University of Illinois at Urbana-Champaign.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. Journal of Applied Psychology, 96(5), 966-980.

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

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and Likert scores. Organizational Research Methods, 22, 710-739.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. Educational and Psychological Measurement, 65, 70-89.

Peeters, M. A., Van Tuijl, H. F., Rutte, C. G., & Reymen, I. M. (2006). Personality and team performance: A meta-analysis. European Journal of Personality, 20, 377-396.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. Human Performance, 14, 187-207.

Salgado, J. F. (2016). A theoretical model of psychometric effects of faking on assessment procedures: Empirical findings and implications for personality at work. International Journal of Selection and Assessment, 24, 209-228.

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(3), 213-216.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. Journal of Applied Psychology, 86, 943-953.

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). Oxford University Press.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? Journal of Applied Psychology, 89, 497-508.

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.

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

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, 175-199.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. Educational and Psychological Measurement, 59, 197-210.

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. doi:10.1177/0146621617703183

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.

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

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., Frick, S., & Brown, A. (2020). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. Psychological Assessment. Advance online publication. https://doi.org/10.1037/pas0000971

White, L. A., & Young, M. C. (1998, August). Development and validation of the assessment of individual motivation (AIM). Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Young, A. L. (2018). Faking Resistance of a Forced-Choice Measure of the Dark Triad. Doctoral dissertation, Department of Psychology, North Carolina State University.

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. Organizational Research Methods, 23, 569-590.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. Journal of Applied Psychology, 84, 551-563.

Zickar, M. J. (2000). Modeling faking on personality tests. In D. R. Ilgen & C. L. Hulin (Eds.), Computational modeling of behavior in organizations: The third scientific discipline (pp. 95–113). American Psychological Association.

RECEIVED 09/29/20 ACCEPTED 02/02/21

Appendix
Analytical Strategy for TIRT DIF and Categorical MACS DIF Method

Analytical strategy for TIRT DIF involved two steps. First, a constrained baseline DIF test was performed to identify discriminating non-DIF blocks that could serve as an anchor subset for subsequent free baseline DIF tests. For the constrained baseline TIRT DIF method, item parameters for the studied MFC block were freely estimated across the honest and the faking conditions, whereas parameters for all other blocks were constrained to be equal. Then the item parameters of the studied block were tested with six *df* Wald tests (i.e., three loadings and three intercepts per block) using the MODEL TEST command in the Mplus program. If the Wald test was statistically significant, the studied block was identified as DIF. In this research, the constrained baseline model identified 13 non-DIF blocks. Based on the block-level discrimination, blocks 1, 8, and 12 were chosen as an anchor subset for the subsequent free baseline DIF tests on the remaining 15 blocks. For the next step, the free baseline DIF analysis was conducted with three anchor blocks. For the free baseline TIRT DIF analysis, item parameters for all MFC blocks were freely estimated across test conditions, except for anchor blocks. Item parameters of the studied blocks were then tested for DIF one at a time on their parameters with six *df* Wald tests. If the DIF test was statistically significant, the studied MFC block was classified as DIF. To control Type I error of multiple DIF tests, we used a Bonferroni corrected critical *p*-value (*p* = 0.00333 [0.05/15]). See Supplemental Materials for an example Mplus syntax of the free baseline TIRT DIF method.

For the DIF test of Likert-type measure, we conducted a categorical mean and covariance structure (MACS) DIF analysis at each scale-level. To this end, the s*equential free baseline approach* was also applied. First, the constrained baseline model was specified, where it constrained each item's loading and threshold to be equal across conditions. This model was compared with each of the models in which respective loading and threshold are freely estimated for each item. By comparing the respective changes in chi-square using two *df* (i.e., loading and threshold), the DIF item was tested (i.e., likelihood ratio test [LRT]). The constrained baseline LRT was conducted for all items and highly discriminating non-DIF items were used as anchor items for subsequent free-baseline DIF tests. Consequently, items 7, 8, 37, 41, and 43 were identified as anchor items for neuroticism, openness, agreeableness, extraversion, and conscientiousness, respectively. Using these items as anchor items, the free baseline model was specified, where item parameters were freely estimated and only the anchor item are constrained across the honest and faking conditions. Then, a series of constrained models that tested one item for DIF were formed by constraining loading and threshold parameters simultaneously to be equal across conditions. Finally, each DIF item was tested one at a time using the Bonferroni corrected critical *p*-values for each dimension (*p* = .00455 (0.05/11) for openness; *p* = .00556 (0.05/9) for conscientiousness; *p* = .005 (0.05/10) for extraversion; *p* = .005 (0.05/10) for agreeableness, *p* = .00556 (0.05/9) for neuroticism). For the categorical MACS DIF analysis, we used the DIF TEST function implemented in the Mplus program.