Personality and Individual Differences
2024, Vol. 222, No. 112581

What are you comparing it to? Investigating order effects in presentation of multidimensional forced choice personality items

Steven Zhou, Virginia Cheng, and Philseok Lee
George Mason University

Despite growing interest in multidimensional forced choice (MFC) measures, there has been relatively little research on the *design* of MFC measures and how it affects their psychometric properties. This study focuses on the prevalence of order effects *within* MFC blocks, that is, the degree to which the first item in a MFC block unduly influences participant responses or rankings on subsequent items. We focus on conscientiousness and neuroticism, the two personality dimensions with evidence of the largest social desirability and faking effects, to examine if putting the conscientiousness or neuroticism item *first* within a block elicits a participant's comparative judgment cognitive processes that lead to systematically biased lower or higher rankings on subsequent items in the block. Through an experimental mixed within- and between- persons study comparing a random-order MFC with adjusted MFC measures that place the conscientiousness or neuroticism item first in each block, we found little to no evidence of order effects. Rankings and factor scores on the personality traits remained consistent despite changing the order of items within MFC blocks. However, placing the conscientiousness item first did lead to decreases in criterion-related validity. Implications for future researchers and practitioners using MFC measures are discussed.

Keywords: multidimensional forced choice, personality, order effects, social desirability, bias, validity

Over the past two decades, multidimensional forced choice (MFC) measures have gained increasing attention as a viable alternative to traditional Likert-type measures in the field of industrial and organizational (I-O) psychology (Speer et al., 2023). Unlike Likert-type measures, MFC measures present respondents with multiple statements within an item block and require respondents to select or rank from “most like me” to “least like me” statements. To date, research has demonstrated several key findings: MFC measures can (i) provide similar or even better criterion-related validity compared to Likert-type measures (e.g., Lee et al., 2018; Salgado & Tauriz, 2014; Wetzel & Frick, 2020), (ii) mitigate faking responses (e.g., Cao & Drasgow, 2019; Lee & Joo, 2021; Speer et al., 2023), (iii) can be effectively estimated by various item response theory (IRT) models (e.g., Brown & Maydeu-Olivares, 2012; Stark et al., 2005; Joo et al., 2023), (vi) are robust to internal and external measurement biases (e.g., Morillo et al., 2019; Zhou et al., in press), and (v) may elicit more negative test-taker reactions due to increased cognitive demand and fatigue (e.g., Dalal et al., 2021; Sass et al., 2020).

Building on the growing prominence of MFC measures, recent research has increasingly emphasized the *test design* of MFC measures, suggesting that their effectiveness and fake-resistance depend on their test design (e.g., Kreichmann et al., 2023; Lee et al., 2022; Pavlov et al., 2021; 2024). Thus, given the unique characteristics of the MFC format, researchers must consider a variety of issues during the test development stage. For example, MFC measures must determine the optimal number of statements to include within blocks, strategically align social desirability, decide the number of dimensions to include, and devise effective mixtures of positively and negatively keyed statements. Recent empirical studies have shown that MFC measures can be more fake-resistant when item social desirability levels within blocks are appropriately matched (e.g., Kreitchmann et al., 2023; Pavlov et al., 2021). In addition, more accurate estimation of MFC scores can be achieved by including negatively keyed statements within MFC blocks (e.g., Frick et al., 2023; Lee et al., 2022). Despite recent efforts to explore the test design of MFC measures, a crucial aspect of the

Preprint template (#prettypreprint) created by Brenton M. Wiernik and downloaded from www.osf.io/hsv6a

Correspondence concerning this paper should be addressed to Steven Zhou, George Mason University, szhou9@gmu.edu

test design remains unaddressed, namely the impact of *order effects* within MFC blocks.

Order effects occur when a previous question item or response item cognitively affects the subsequent question item or response item, which may introduce confounding variables (Rasinski et al., 2012). Previous research has exclusively focused on Likert-type measures when exploring and analyzing order effects in test design or, relatedly, grouping of items within Likert-type personality measures (McFarland et al., 2002). Specifically, there are effects in: response presentation placement (Maddocks, 2005; Krosnick & Alwin, 1987; Abakoumkin, 2011), complexity of item content (Sanjeev & Balyan, 2014), positive- or negative-first ordered responses (Chan, 1991), spacing of items (Tourangeau et al., 2004), and vertically-placed responses (Tourangeau et al., 2004). Order effects can also cause measurement errors in Likert-type measures, leading to inaccurate scoring, measurement biases, and biased decision-making (Krosnick & Alwin, 1987; Maddocks, 2005; McFarland, 1981; Rasinski et al., 2012).

These effects drew upon multiple theories, such as the satisficing principle and cognitive theory. Simons' (1957) satisficing principle states that a respondent will seek and select the first satisfying option rather than examining all response options for the best solutions in order to minimize cognitive energy or psychological costs. Building off this principle, the cognitive theory offers that respondents may be influenced by two sorts of cognitive effects: primacy and recency (Abakoumkin, 2011; Krosnick & Alwin, 1987). Primacy effects occur when items placed at the beginning of a list have an increased selection chance given individual differences in cognitive load and acceptability of option. On the other hand, recency effects occur when items presented last are more likely to be selected based on individual differences in cognitive load and acceptability of option.

Although previous research provides significant insights into the order effects of psychological measurement, there remains a significant dearth of research that examines the order effects within an MFC block. Order effects might be particularly problematic with MFC measures, considering participants' tendency to read statements from top to bottom even when all statements are concurrently presented within an item block. Upon reading the statements, participants are expected to engage in a comparative judgment process (Thurstone, 1927). In the case of a triplet-MFC measure, participants are assumed to compare the first and second statements, the first and third statements, and the second and third statements.

Specifically, an item presented first with a strong priming effect could influence participant ratings on the remaining two items in the block. Here, social desirability may be considered a strong priming effect in that responding in a manner to be viewed as favorable for one item may influence the respondent to continue this response behavior for subsequent items. For example, similar effects connecting social desirability and order effects have been shown in religiosity and drinking (Rodriguez et al., 2014). Furthermore, given that MFC measures are more cognitively demanding (Bowen et al., 2002; Dalal et al., 2021), respondents may be tempted to answer based on minimizing cognitive energy and thus rely unknowingly on the aforementioned priming effects. Moreover, if order effects occur differently across different groups, they could lead to various negative psychometric issues such as differential item functioning (Lee et al., 2021). Prior studies have found some evidence for how the interaction of items within a block is important. For example, Morillo et al. (2019) found evidence for complex interaction in blocks whereby having openness as the first item and emotional stability as the second item influenced item parameters.

Our study aims to assess the impact of order effects within item blocks using various MFC designs. We propose that an item with high positive social desirability that is displayed first in the block would result in subsequent items to be rated systematically lower. For example, the aforementioned religiosity and drinking study (Rodriguez et al., 2014) found that asking about religiosity (which scored high on social desirability) *first* led to lower scores on drinking measures. Applied to our study on within-block order effects, we expect that a personality trait that is highly socially desirable, if presented first in a block, would bias participants towards rating it systematically higher within the block. Likewise, an item with high negative social desirability that is displayed first in the block would result in subsequent items to be rated systematically higher. Here, we focus on *conscientiousness* and *neuroticism*, which have been shown in prior meta-analytic studies to have the largest effects in social desirability and faking (Speer et al., 2023; Martínez and Salgado., 2021).

H1. Participants who complete a MFC personality measure where the first statement in a block is always the Conscientiousness statement will have inflated scores on Conscientiousness (compared to a control random-order MFC measure).

H2. Participants who complete a MFC personality measure where the first statement in a block is always the Neuroticism statement will have suppressed scores on Neuroticism (compared to a control random-order MFC measure).

Moreover, these order effects may consequently have downstream impact on psychometric properties such as convergent validity with traditional Likert-type scales and criterion-related validity with outcome variables. For example, McFarland (1981) found that altering the question order led to some differences in correlations on subsequent questions. Similarly, Schell and Oswald (2013) reported that item order in a personality measure affects scale-level correlations.

RQ1. Assuming the presence of order effects, does presenting the conscientiousness or neuroticism items first in the MFC measure result in different convergent and criterion-related validity estimates?

Method

Participants

Undergraduate students from a university were recruited through the university Psychology Research Participation System. We collected a quality-controlled sample of 428 respondents (60% female) with a mean age of 22. Participants described themselves as White (29.2%), Asian (29.2%), Hispanic (17.5%), and Black (10.7%) while the remaining 13.4% identified as other.

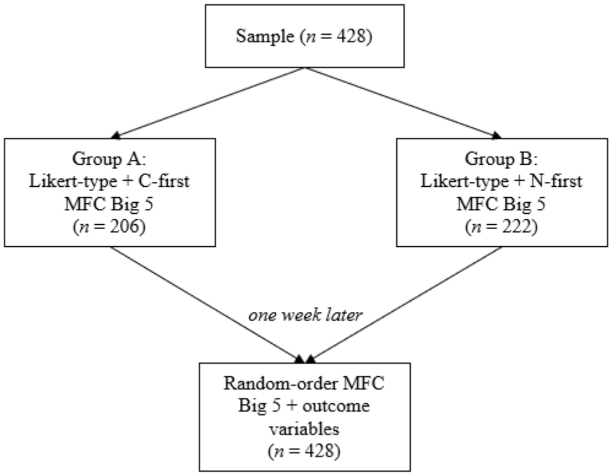
Measures

We created three different versions of a 20-triplet MFC measure of the Big Five (i.e., 12 items per personality dimension) by adapting the Brown (2010) measure of Big Five, which was developed by matching 60 IPIP items (Goldberg, 1992) based on similar item characteristics in single-stimulus IRT estimation (see Brown, 2010 for details). One version consistently showed Conscientiousness items first within each block. A second version consistently showed Neuroticism items first within each block. A third version randomly ordered the MFC items. The two modified MFC measures are included in the Appendix. All three versions presented the statement items in a vertical placement within each block. According to Tourangeau et al. (2004), top items within a vertically-placed list are often seen as most desirable given the logic of “deeply-rooted metaphors” (e.g., heaven is up and hell is down); therefore there are potential implications that the Conscientiousness or Neuroticism items will be chosen more often. Each block was presented one at a time to eliminate potential distractions. We also measured personality using the same 60 item measure as the MFC, but in Likert-type format. Finally, to examine criterion-related validity, we measured four outcome variables: satisfaction (adapted from the Diener et al., 1985 satisfaction with life scale), perseverance and adaptability (from the Oswald et al., 2004 student biodata measures), and self-reported GPA.

Procedure

Participants were randomly split into two groups, Group A and Group B, to complete two versions of the 20-minute online Qualtrics surveys. At time 1, Group A was given the Qualtrics survey that contained the Likert-type personality scale and the Conscientiousness-first MFC version (i.e., C-first MFC); while, Group B was first given the Qualtrics survey that contained the Likert-type personality scale and the Neuroticism-first MFC version (i.e., N-first MFC). One week later (Time 2), both groups (i.e., Group A and B) completed the second survey that contained the random-MFC and outcome measures (see Figure 1).

Figure 1. Flow chart of study design.

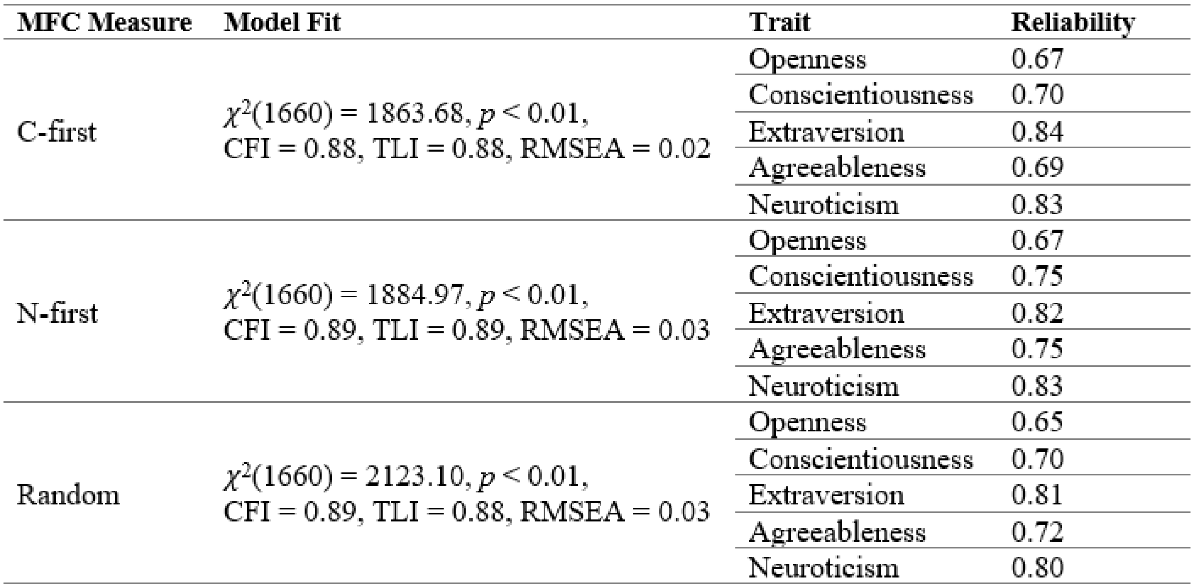

Sample (n = 428)
Group A: Likert-type + C-first MFC Big 5 (n = 206)
Group B: Likert-type + N-first MFC Big 5 (n = 222)
one week later
Random-order MFC Big 5 + outcome variables (n = 428)

Analytic Strategy

MFC scores were estimated using the Thurstonian IRT model (Brown & Maydeu-Olivares, 2012) in all three MFC datasets with the *Mplus* program (Muthén & Muthén, 1998-2017). For the estimation, the mean- and variance-adjusted unweighted least squares estimator was used. Model fit and estimated reliability for each of the three MFC measures are found in Table 1 below. Empirical reliability was computed based on Dueber et al. (2019) using the *thurstonianIRT* package in R (Bürkner et al., 2023).

Hypothesis 1 and 2 were tested using two approaches. First, to directly assess for order effects, we computed a simple count of how many times each participant ranked the C-statement (H1) or N-statement (H2) first in the block (range of 0 to 12, because there

Table 1. Model fit and reliability for three MFC measures.



MFC Measure	Model Fit	Trait	Reliability
C-first	$\chi^2(1660) = 1863.68, p < 0.01$, CFI = 0.88, TLI = 0.88, RMSEA = 0.02	Openness	0.67
		Conscientiousness	0.70
		Extraversion	0.84
		Agreeableness	0.69
		Neuroticism	0.83
N-first	$\chi^2(1660) = 1884.97, p < 0.01$, CFI = 0.89, TLI = 0.89, RMSEA = 0.03	Openness	0.67
		Conscientiousness	0.75
		Extraversion	0.82
		Agreeableness	0.75
		Neuroticism	0.83
Random	$\chi^2(1660) = 2123.10, p < 0.01$, CFI = 0.89, TLI = 0.88, RMSEA = 0.03	Openness	0.65
		Conscientiousness	0.70
		Extraversion	0.81
		Agreeableness	0.72
		Neuroticism	0.80

are 12 statements for each). We then conducted a dependent sample t-test comparing the C-first MFC and the random-MFC for Group A participants (H1) and the same test comparing the N-first MFC and random-MFC for Group B participants (H2). Next, to identify if order effects led to higher scores on conscientiousness or neuroticism, we conducted dependent sample t-tests on the factor scores for conscientiousness (H1) and neuroticism (H2).

For RQ1, we first tested for differences in convergent validity between each of the MFC measures and the Likert-type measures. Specifically, we compared the correlations between (i) Likert-type conscientiousness and C-first MFC conscientiousness vs. Likert-type conscientiousness and random-MFC conscientiousness; and (ii) Likert-type neuroticism and N-first MFC neuroticism vs. Likert-type neuroticism and random-MFC neuroticism. If order effects exist, then this analysis will reveal a change in the convergent validity of the MFC measure with the Likert-type measure. Finally, to assess for differences in criterion-related validity, we ran two sets of four analyses. For the first set, we ran four multigroup regressions of conscientiousness scores predicting each of the four outcome variables (GPA, satisfaction, perseverance, and adaptability), moderated by the *type* of conscientiousness score (i.e., C-first MFC, random-MFC, or Likert-type). The second set repeated this using neuroticism as the predictor, moderated by type of score (i.e., N-first MFC, random-MFC, or Likert-type).

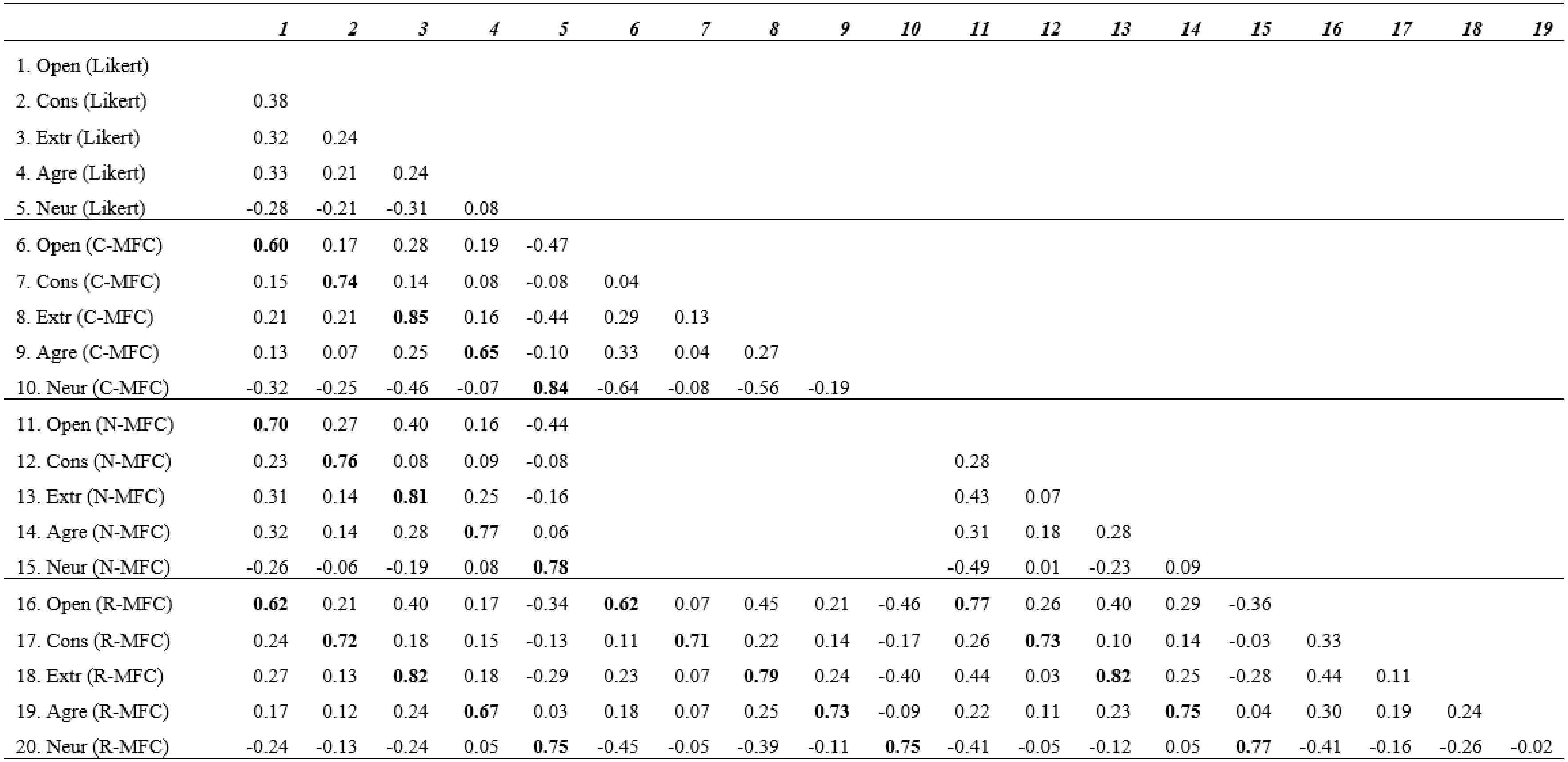
Results

Table 2 depicts the zero-order correlations between each of the 15 MFC Big Five variables (i.e., five traits X three MFC types) and the five Likert-type Big Five variables.

H1. Hypothesis 1 was not supported. For the direct test of order effects (i.e., how many times the conscientiousness item was ranked first), the difference between the C-first MFC and random-MFC was not significant: $t(205) = 1.42$, $p = 0.08$. For the test on difference in factor scores, the result was also non-significant: $t(205) = 0.39$, $p = 0.35$. We followed up with a test of equivalence via two one-sided tests (TOST; Lakens et al., 2018) to assess for evidence *supporting* the null finding (i.e., no difference in scores) based on a smallest effect size of interest of 0.11 (Gignac & Szodorai, 2016). The purpose of this test is to assess if the true effect size is *less* than the smallest effect size of interest (Lakens et al., 2018). The equivalence test was significant, $t(205) = -1.94$, $p = 0.03$ (Hedges' $g = 0.03$, 90% C.I. = [-0.06, 0.10]). This suggests that the true effect size is between -0.11 and 0.11, which is small enough to be arguably negligible (Gignac & Szodorai, 2016). Thus, the combination of the two dependent sample *t*-tests and the equivalence test suggests that there is no meaningful difference in conscientiousness scores when measured via the C-first MFC versus the random-MFC; in other words, order does not matter for measuring conscientiousness.

H2. Again, the test of the difference in number of times neuroticism was ranked first was not significant: $t(221)$

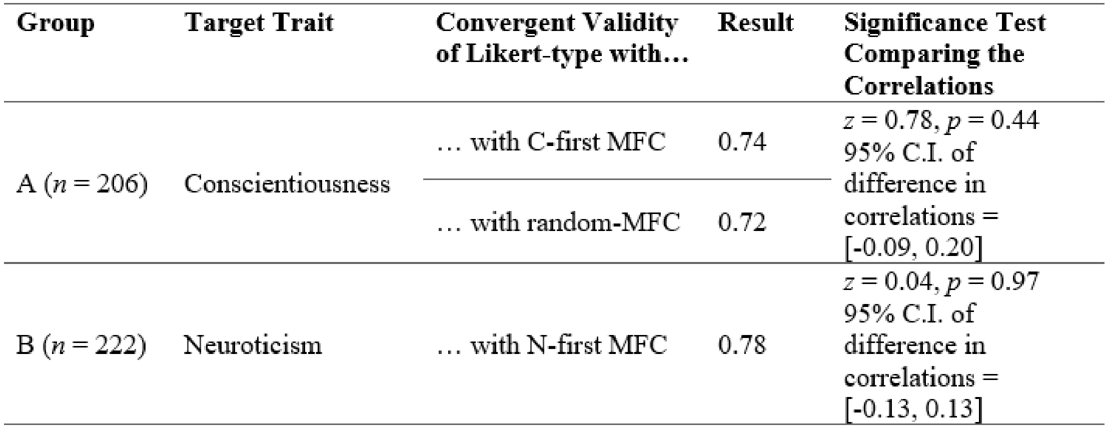
Table 2. Correlation of study variables.



	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1. Open (Likert)																			
2. Cons (Likert)	0.38																		
3. Extr (Likert)	0.32	0.24																	
4. Agre (Likert)	0.33	0.21	0.24																
5. Neur (Likert)	-0.28	-0.21	-0.31	0.08															
6. Open (C-MFC)	**0.60**	0.17	0.28	0.19	-0.47														
7. Cons (C-MFC)	0.15	**0.74**	0.14	0.08	-0.08	0.04													
8. Extr (C-MFC)	0.21	0.21	**0.85**	0.16	-0.44	0.29	0.13												
9. Agre (C-MFC)	0.13	0.07	0.25	**0.65**	-0.10	0.33	0.04	0.27											
10. Neur (C-MFC)	-0.32	-0.25	-0.46	-0.07	**0.84**	-0.64	-0.08	-0.56	-0.19										
11. Open (N-MFC)	**0.70**	0.27	0.40	0.16	-0.44														
12. Cons (N-MFC)	0.23	**0.76**	0.08	0.09	-0.08						0.28								
13. Extr (N-MFC)	0.31	0.14	**0.81**	0.25	-0.16						0.43	0.07							
14. Agre (N-MFC)	0.32	0.14	0.28	**0.77**	0.06						0.31	0.18	0.28						
15. Neur (N-MFC)	-0.26	-0.06	-0.19	0.08	**0.78**						-0.49	0.01	-0.23	0.09					
16. Open (R-MFC)	**0.62**	0.21	0.40	0.17	-0.34	**0.62**	0.07	0.45	0.21	-0.46	**0.77**	0.26	0.40	0.29	-0.36				
17. Cons (R-MFC)	0.24	**0.72**	0.18	0.15	-0.13	0.11	**0.71**	0.22	0.14	-0.17	0.26	**0.73**	0.10	0.14	-0.03	0.33			
18. Extr (R-MFC)	0.27	0.13	**0.82**	0.18	-0.29	0.23	0.07	**0.79**	0.24	-0.40	0.44	0.03	**0.82**	0.25	-0.28	0.44	0.11		
19. Agre (R-MFC)	0.17	0.12	0.24	**0.67**	0.03	0.18	0.07	0.25	**0.73**	-0.09	0.22	0.11	0.23	**0.75**	0.04	0.30	0.19	0.24	
20. Neur (R-MFC)	-0.24	-0.13	-0.24	0.05	**0.75**	-0.45	-0.05	-0.39	-0.11	**0.75**	-0.41	-0.05	-0.12	0.05	**0.77**	-0.41	-0.16	-0.26	-0.02

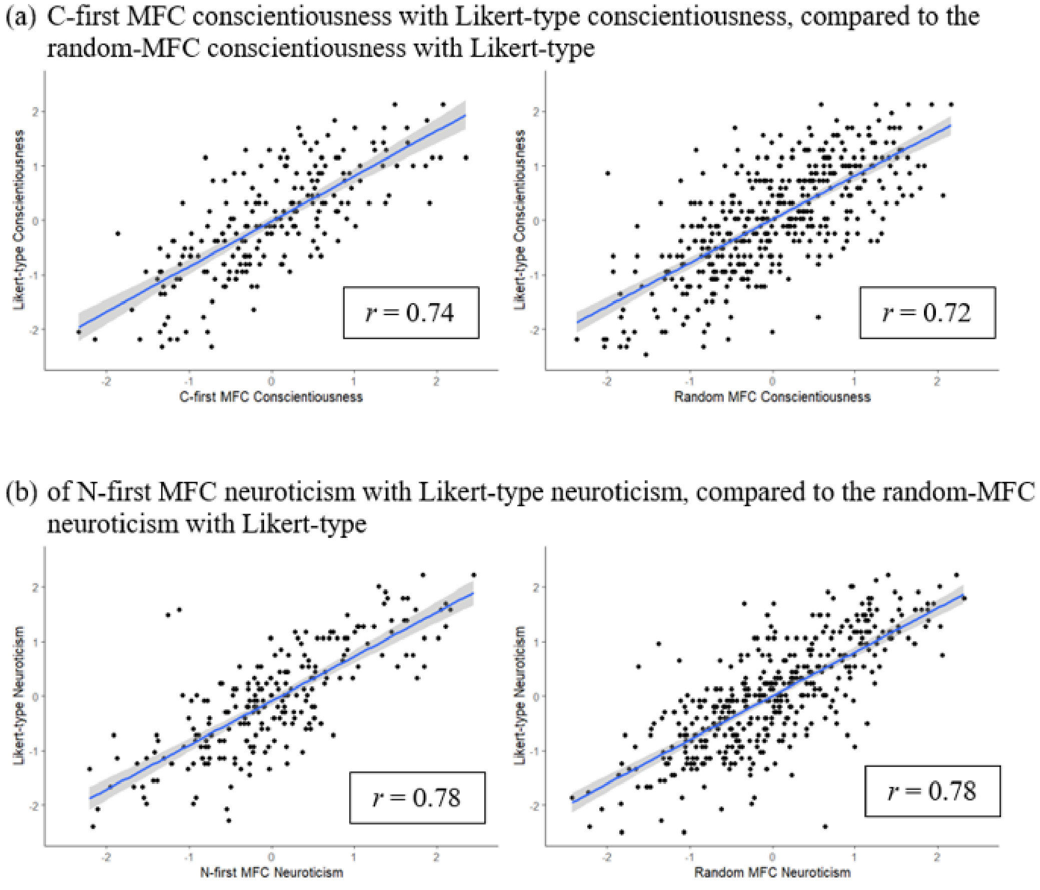
Notes: Open = openness, Cons = conscientiousness, Extr = extraversion, Agre = agreeableness, Neur = neuroticism, C-MFC = C-first MFC, N-MFC = N-first MFC, R-MFC = random-order MFC. Correlations between variables 6-10 with 11-15 are blank because of the between-subjects component of the study design (i.e., participants completed *either* the C-first MFC or the N-first MFC, but not both). Convergent validities between the different measures and their corresponding personality traits are bolded.

Table 3. Results of comparison of convergent validities.



Group	Target Trait	Convergent Validity of Likert-type with…	Result	Significance Test Comparing the Correlations
A (n = 206)	Conscientiousness	… with C-first MFC	0.74	z = 0.78, p = 0.44 95% C.I. of difference in correlations = [-0.09, 0.20]
		… with random-MFC	0.72	
B (n = 222)	Neuroticism	… with N-first MFC	0.78	z = 0.04, p = 0.97 95% C.I. of difference in correlations = [-0.13, 0.13]

Figure 2. Visualization of the correlations between key variables.



(a) C-first MFC conscientiousness with Likert-type conscientiousness, compared to the random-MFC conscientiousness with Likert-type

(b) of N-first MFC neuroticism with Likert-type neuroticism, compared to the random-MFC neuroticism with Likert-type

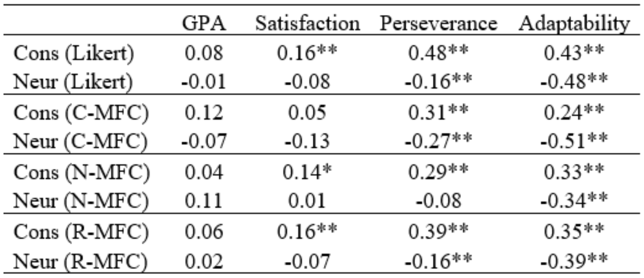
= -0.34, p = 0.63. Likewise, test of the difference in factor scores was not significant: t(221) = 1.33, p = 0.09. However, this time, the equivalence test was also not significant, t(221) = -1.25, p = 0.11 (Hedges' g = 0.09, 90% C.I. = [-0.01, 0.13]). Thus, the results are inconclusive; while there was no statistically significant difference between neuroticism scores, the data also does not support a conclusion that there is *no meaningful difference* and thus no order effect.

RQ1. Table 3 and Figure 2 depict the results of the tests of differences in convergent validities; neither were significant. In other words, placing the conscientiousness item first did not impact the convergent validity of

MFC conscientiousness with Likert-type conscientiousness; likewise with neuroticism. As an auxiliary finding, we noted that the convergent validities of the random-MFC were stronger with the N-first MFC than with the C-first MFC (see Table 2 diagonals).

Interestingly, there were some significant differences in the criterion-related validity estimates predicting satisfaction, perseverance, and adaptability (but not GPA); zero-order correlations for these variables are in Table 4. Specifically, simple slopes analyses revealed that conscientiousness was the strongest predictor of satisfaction when using the random-MFC (B = 0.88, SE = 0.34, p = 0.01), a non-significant predictor when using the C-first MFC (B = 0.25, SE = 0.34, p = 0.46), and a weak predictor when using the Likert-type measure (B = 0.09, SE = 0.04, p = 0.03). Similarly, conscientiousness was the strongest predictor of perseverance when using the random-MFC (B = 2.12, SE = 0.36, p < 0.01) and the weakest predictor when using the Likert-type measure (B = 0.34, SE = 0.04, p < 0.01). Finally, conscientiousness was the strongest predictor of adaptability when using the random-MFC (B = 1.99, SE = 0.36, p < 0.01) and the weakest predictor when using the Likert-type measure (B = 0.27, SE = 0.04, p < 0.01).

Table 4. Correlation of personality predictors with outcome variables, by measure format.



	GPA	Satisfaction	Perseverance	Adaptability
Cons (Likert)	0.08	0.16**	0.48**	0.43**
Neur (Likert)	-0.01	-0.08	-0.16**	-0.48**
Cons (C-MFC)	0.12	0.05	0.31**	0.24**
Neur (C-MFC)	-0.07	-0.13	-0.27**	-0.51**
Cons (N-MFC)	0.04	0.14*	0.29**	0.33**
Neur (N-MFC)	0.11	0.01	-0.08	-0.34**
Cons (R-MFC)	0.06	0.16**	0.39**	0.35**
Neur (R-MFC)	0.02	-0.07	-0.16**	-0.39**

* p < 0.05, ** p < 0.01

We repeated this set of four regressions with neuroticism scores predicting each outcome variable, again moderated by the *type* of neuroticism score. This time, there were no significant differences in the criterion-related validity estimates predicting the outcomes. Simple slopes analyses revealed only that neuroticism was the strongest predictor of adaptability when using the random-MFC (B = -2.03, SE = 0.35, p < 0.01) and the weakest predictor when using the Likert-type measure (B = -0.22, SE = 0.03, p < 0.01). Put together, the evidence suggests that there are order effects in terms of criterion-related validity of conscientiousness as a predictor, but no order effects in terms of criterion-related validity of neuroticism as a predictor.

Discussion

Overall, the data presents a mixed picture. The evidence suggests that placing the conscientiousness item first in the MFC does *not* create order effects in terms of score inflation, but the results placing the neuroticism item first was inconclusive. There was no evidence of order effects leading to downstream impact on convergent validities, but there was substantial evidence of impact on criterion-related validity when using conscientiousness (but not neuroticism) as a predictor.

These findings have important theoretical and practical implications. Despite the assumption of comparative judgment used in MFC tests, our study suggests that the initial item in a block has little influence on respondent scores to the remaining items (Thurstone, 1927). In other words, an item that is presented first (e.g., conscientiousness-first or neuroticism-first blocks) did not influence participant rankings for that item, did not influence participants' overall factor scores, and did not diminish the MFC test's convergent validity with Likert-type tests. One plausible explanation may be that MFC presents similarly desirable options within blocks; therefore, respondents read all options without the temptation to answer based on minimization of cognitive energy given the satisficing principle and cognitive theory (Simons, 1957; Krosnick & Alwin, 1987). Another possibility would be that MFC assessments inherently requires respondents to engage in comparative judgement; therefore respondents, may not make decisions based on the order of statements, but rather compare all possible combinations of pairwise comparisons and then made their decisions. This suggests that, to some extent, the order of statements within an MFC block may have a reduced effect on respondents' decision, but rather participants prioritize the content of a statement over the presentation order in their decision-making.

Overall, our findings can provide reassurance for test developers and practitioners who plan to use the MFC assessments, as it implies that the order effect might be less influential than previous believed. Furthermore, this highlights the good news that there is less of a need to impose additional strategies within personality assessments (e.g., Krosnick and Alwin (1987) strategy to increase concentration) or alternative costly approaches (e.g., Krosnick and Alwin (1987) suggestion to randomize presentation order for each respondent) as the effects of response order may not be a cause for concern. However, our study did provide evidence of order effects on the criterion-related validity of the MFC test when conscientiousness is used to predict various student outcomes. Interestingly, criterion-related validity was generally strongest when using the random-MFC and weakest when using Likert-type measures. This is consistent with previous literature of non-significant criterion-related validity difference for GPA between MFC and Likert for neuroticism (Christiansen et al., 2005; Huber et al., 2021). This is also similar to findings from studies on item groupings within personality tests (e.g., McFarland et al., 2002), which suggested that randomly changing the order of items within a personality test (as opposed to grouping all conscientiousness items together, for example) led to improvements in faking resistance and model fit. Additionally, our results suggest that order effects detract from criterion-related validity, because the C-first MFC measures showed significantly weaker criterion-related validity compared to the random-MFC measures. Based on recent findings, a possible reason may be that respondents that perceive conscientiousness as worthy may “anchor” their judgment as a basis for evaluating subsequent items, therefore distorting how conscientiousness predicts student scores (Lee, 2023). Although most of our results suggest no evidence of order effects in terms of factor scores and convergent validity, this clearly demonstrates that placing the conscientiousness item first within a block result in a reduction of criterion-related validity. Future researchers and practitioners should thus be wary of the potential impact of an item’s position within a block that might induce participants to respond similarly to those items when compared to items on similar outcome variables. Finally, our results only found these effects for conscientiousness and not neuroticism; therefore, indicating that the impact of order effects on criterion-related validity varies across different personality dimensions.

Limitations and Future Directions

There are several limitations to this study that point to future research directions. First and foremost, our study was conducted on a sample of undergraduate college students with a specific focus on student-related outcomes, thus clearly delineating the generalizability of our study to college students only. Future studies can expand on our study by using actual employee samples to enhance the generalizability of these findings beyond students. Moreover, although our study employed an experimental design with a time lag to reduce the likelihood of common method bias found in cross-sectional surveys, future studies can employ more sophisticated experimental designs that incorporate others-report data to obtain more accurate estimates of study variables.

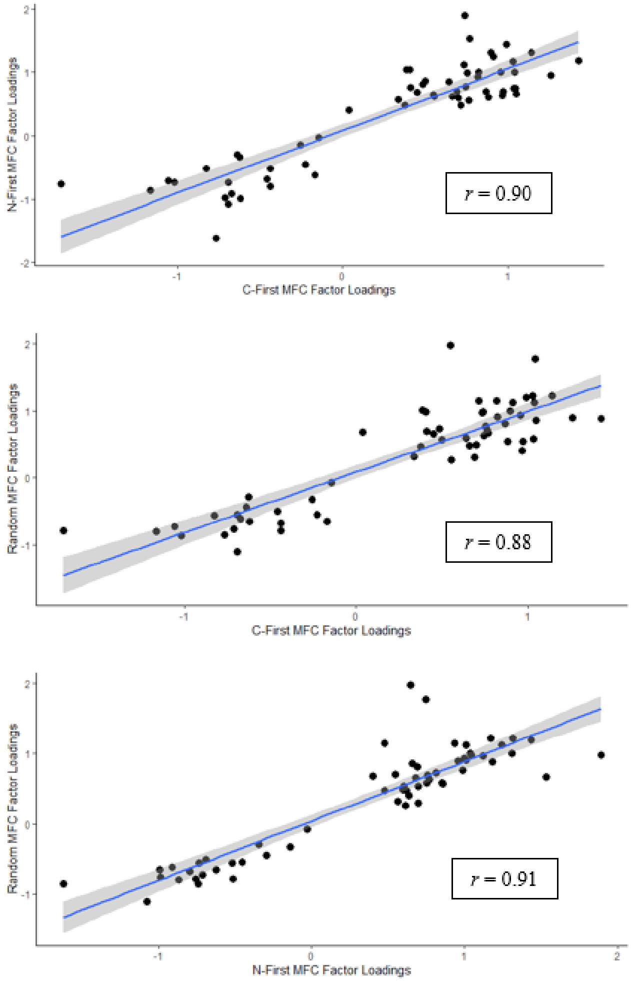
Second, we limited our investigation to order effects using conscientiousness and neuroticism, the two factors that show the strongest social desirability and faking effects (Speer et al., 2023; Martínez and Salgado, 2021). Future studies can expand by investigating order effects with other items, or order effects based on other sources beyond social desirability. For example, there may be order effects in agreeableness, whether due to social desirability or prosocial item wording, especially if there are agreeableness-relevant situational contexts (e.g., survey of potential romantic partners). Furthermore, future research could assess degrees of order effects through varying levels of faking induction (e.g., high or low levels of instructed faking) as order effects may emerge given a higher press to fake.

Finally, future research can investigate order effects with more granularity at the level of each individual item parameter and in the context of other test design elements. Following the advice of an anonymous reviewer, we investigated item invariance across three different versions via Morillo et al.'s (2019) approach. To do so, we compared the factor loadings of each of the 60 item parameters between the C-first, N-first, and random MFC measures (see Figure 3). The visualizations demonstrate that the factor loadings are all very similar. For example, when comparing the C-first and N-first MFC measures, the factor loadings for each of the 60 items were strongly correlated (*r* = 0.90). Future studies could conduct differential item functioning testing on the item parameters, especially given recent advances in methodology for measurement invariance among MFC measures (see Lee et al., 2021).

Conclusion

Our study is the first to investigate *order effects* within MFC blocks. It contributes to the growing literature surrounding MFC measures, with particular emphasis on how the *test design* might impact psychometric properties and lead to downstream effects on criterion-related validity and the usefulness of the test in applied settings. Our study uses a student sample to demonstrate that while there are little to no order effects on factor scores and convergent validity, there are order effects when conscientiousness is used to predict various student outcomes. We hope that this study motivates future researchers to further investigate order effects, and that it provides useful guidance for applied practitioners seeking to implement MFC measures in their workplace or other settings.

Figure 3. Correlation of factor loadings between C-First, N-First, and Random MFC.



References

Abakoumkin, G. (2011). Forming choice preferences the easy way: order and familiarity effects in elections. *Journal of Applied Social Psychology*, *41*(11), 2689-2707. https://doi.org/10.1111/j.1559-1816.2011.00845.x

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*(3), 240-259. https://doi.org/10.1108/eb028952

Brown, A. (2010). *How item response theory can solve problems of ipsative data* [Doctoral dissertation, Universitat de Barcelona].

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*, 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Bürkner, P.-C., Hughes, A., & Trustees of Columbia University. (2023). *thurstonianIRT* package (version 0.12.4). https://github.com/paul-buerkner/thurstonianIRT

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

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, *51*(3), 531–540. https://doi.org/10.1177/0013164491513002

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. https://doi.org/10.1207/s15327043hup1803_4

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*(2), 155-169. https://doi.org/10.1111/j.1468-2389.2008.00420.x

Dalal, D. K., Zhu, X. S., Rangel, B., Boyce, A. S., & Lobene, E. (2021). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*, *36*(1), 55–70. https://doi.org/10.1007/s10869-019-09655-6

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71-75. https://doi.org/10.1207/s15327752jpa4901_13

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement*, *79*(1), 108-128. https://doi.org/10.1177/0013164417752782

Frick, S., Brown, A., & Wetzel, E. (2023). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, *58*(1), 1-29. https://doi.org/10.1080/00273171.2021.1938960

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment*, *4*(1), 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Huber, C. R., Kuncel, N. R., Huber, K. B., & Boyce, A. S. (2021). Personality tests: An experimental investigation using modern forced choice measures. *Personnel Assessment and Decisions*, *7*(1), 3. https://doi.org/10.25035/pad.2021.01.003

Joo, S., Lee, P., & Stark, S. (2023). Modeling multidimensional forced choice measures with the Zinnes and Griggs Pairwise Preference Item Response Theory model. *Multivariate Behavioral Research*, *58*(2), 241-261. https://doi.org/10.1080/00273171.2021.1960142

Kreitchmann, R. S., Sorrel, M. A., & Abad, F. J. (2023). On bank assembly and block selection in multidimensional forced-choice adaptive assessments. *Educational and Psychological Measurement*, *83*(2), 294–321. https://doi.org/10.1177/00131644221087986

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201-219. https://doi.org/10.1086/269029

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Lee, H. (2023). Reduction of faking with the use of a forced-choice personality test: Cross-cultural comparisons between South Korea and the United States. *International Journal of Selection and Assessment*, *31*(1), 147-162. https://doi.org/10.1111/ijsa.12408

Lee, P., & Joo, S. H. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions*, *7*(1), 4. https://doi.org/10.25035/pad.2021.01.004

Lee, P., Joo, S. H., & Jia, Z. (2022). Opening the black box of the response process to personality faking: An application of item response tree models. *Journal of Business and Psychology*, *37*(6), 1199-1214. https://doi.org/10.1007/s10869-022-09791-6

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*, 229-235. https://doi.org/10.1016/j.paid.2017.11.031

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*(4), 739-771. https://doi.org/10.1177/1094428120959822

Lee, P., Joo, S. H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences*, *191*, 111555. https://doi.org/10.1016/j.paid.2022.111555

Maddocks, A. M. (2005). *A meta-analysis of item wording and response order effects on attitude surveys*. Loyola University Chicago.

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, *12*, 732241. https://doi.org/10.3389/fpsyg.2021.732241

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*(2), 208–215. https://doi.org/10.1086/268651

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Journal of Work and Organizational Psychology*, *35*(2), 75-83. https://doi.org/10.5093/jwop2019a11

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.) Muthén & Muthén.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187-207. https://doi.org/10.1037/0021-9010.89.2.187

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

Pavlov, G. (2024). An investigation of effects of instruction set on item desirability matching. *Personality and Individual Differences*, 216, 1123423. https://doi.org/10.1016/j.paid.2023.112423

Rasinski, K. A., Lee, L., & Krishnamurty, P. (2012). Question order effects. In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology* (pp. 229–248). American Psychological Association. https://doi.org/10.1037/13619-014

Rodriguez, L. M., Neighbors, C., & Foster, D. W. (2014). Priming effects of self-reported drinking and religiosity. *Psychology of Addictive Behaviors*, *28*(1), 1-9. https://doi.org/10.1037/a0031828

Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*(1), 3-30. https://doi.org/10.1080/1359432X.2012.716198

Sanjeev, M. A., & Balyan, P. (2014). Response order effects in online surveys: An empirical investigation. *International Journal of Online Marketing*, *4*(2), 28-44. https://doi.org/10.4018/ijom.2014040103

Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (2020). Taking the test taker’s perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *27*(3), 572-584. https://doi.org/10.1177/1073191118762049

Schell, K. L., & Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences*, *55*(3), 317-321. https://doi.org/10.1016/j.paid.2013.03.008

Simons, H.A. (1957). *Models of Man; social and rational*. New York: Wiley.

Speer, A. B., Wegmeyer, L. J., Tenbrink, A. P., Delacruz, A. Y., Christiansen, N. D., & Salim, R. M. (2023). Comparing forced-choice and single-stimulus personality scores on a level playing field: A meta-analysis of psychometric properties and susceptibility to faking. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001099

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *101*(2), 266–270. https://doi.org/10.1037/0033-295X.101.2.266

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368-393. https://doi.org/10.1093/poq/nfh035

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional Forced choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. https://doi.org/10.1037/pas0000781

Zhou, S., Lee, P., & Fyffe, S. (in press). Examining gender differences in the use of multidimensional forced-choice (MFC) measures of personality in terms of test-taker reactions and test fairness. *Human Resource Development Quarterly*.

Appendix: MFC Measures

C-First MFC Measure

Block	Statement	Item	Direction
1	1	n1	-
	2	e1	+
	3	o1	+
2	4	c1	+
	5	a1	+
	6	n2	-
3	7	o2	-
	8	e2	+
	9	a2	+
4	10	c2	+
	11	o3	+
	12	n3	+
5	13	a3	+
	14	n4	+
	15	e3	+
6	16	c3	-
	17	o4	+
	18	e4	-
7	19	e5	-
	20	n5	+
	21	a4	+
8	22	c4	+
	23	o5	+
	24	e6	-
9	25	o6	+
	26	n6	+
	27	a5	-
10	28	c5	-
	29	n7	+
	30	e7	+
11	31	c6	+
	32	e8	+
	33	a6	+
12	34	n8	-
	35	a7	+
	36	o7	-
13	37	c7	-
	38	e9	-
	39	n9	+

N-First MFC Measure

Block	Statement	Item	Direction
1	1	n1	-
	2	e1	+
	3	o1	+
2	4	n2	-
	5	a1	+
	6	c1	+
3	7	o2	-
	8	e2	+
	9	a2	+
4	10	n3	+
	11	c2	+
	12	o3	+
5	13	n4	+
	14	a3	+
	15	e3	+
6	16	o4	+
	17	e4	-
	18	c3	-
7	19	n5	+
	20	e5	-
	21	a4	+
8	22	c4	+
	23	o5	+
	24	e6	-
9	25	n6	+
	26	o6	+
	27	a5	-
10	28	n7	+
	29	c5	-
	30	e7	+
11	31	e8	+
	32	a6	+
	33	c6	+
12	34	n8	-
	35	a7	+
	36	o7	-
13	37	n9	+
	38	e9	-
	39	c7	-

C-First MFC Measure (continued)

14	40	c8	+
	41	a8	+
	42	o8	+
15	43	e10	+
	44	o9	+
	45	n10	+
16	46	c9	+
	47	n11	-
	48	a9	-
17	49	c10	-
	50	a10	+
	51	o10	+
18	52	a11	-
	53	e11	+
	54	o11	-
19	55	c11	+
	56	o12	-
	57	n12	+
20	58	c12	+
	59	a12	-
	60	e12	+

N-First MFC Measure (continued)

14	40	a8	+
	41	c8	+
	42	o8	+
15	43	n10	+
	44	e10	+
	45	o9	+
16	46	n11	-
	47	c9	+
	48	a9	-
17	49	c10	-
	50	a10	+
	51	o10	+
18	52	a11	-
	53	e11	+
	54	o11	-
19	55	n12	+
	56	o12	-
	57	c11	+
20	58	c12	+
	59	a12	-
	60	e12	+

Random MFC Measure

Block	Statement	Item	Direction
1	1	n1	-
	2	e1	+
	3	o1	+
2	4	a1	+
	5	c1	+
	6	n2	-
3	7	o2	-
	8	e2	+
	9	a2	+
4	10	c2	+
	11	o3	+
	12	n3	+
5	13	a3	+
	14	n4	+
	15	e3	+
6	16	o4	+
	17	e4	-
	18	c3	-
7	19	e5	-
	20	n5	+
	21	a4	+
8	22	c4	+
	23	o5	+
	24	e6	-
9	25	o6	+
	26	n6	+
	27	a5	-
10	28	c5	-
	29	n7	+
	30	e7	+
11	31	e8	+
	32	a6	+
	33	c6	+
12	34	n8	-
	35	a7	+
	36	o7	-

Random MFC Measure (continued)

13	37	e9	-
	38	n9	+
	39	c7	-
14	40	a8	+
	41	c8	+
	42	o8	+
15	43	e10	+
	44	o9	+
	45	n10	+
16	46	c9	+
	47	n11	-
	48	a9	-
17	49	c10	-
	50	a10	+
	51	o10	+
18	52	a11	-
	53	e11	+
	54	o11	-
19	55	o12	-
	56	c11	+
	57	n12	+
20	58	c12	+
	59	a12	-
	60	e12	+