DOI: 10.1002/hrdq.21521

EMPIRICAL RESEARCH QUANTITATIVE

WILEY

Examining gender differences in the use of multidimensional forced-choice measures of personality in terms of test-taker reactions and test fairness

Steven Zhou | **Philseok Lee** | **Shea Fyffe**

Department of Psychology, George Mason University, Fairfax, Virginia, USA

Correspondence
Steven Zhou, 4400 University Drive, David King Hall Room 3041, Fairfax, VA 22030, USA.
Email: szhou9@gmu.edu

Abstract

Human resource (HR) practices have been focused on using assessments that are robust to faking and response biases associated with Likert-type scales. As an alternative, multidimensional forced-choice (MFC) measures have recently shown advances in reducing faking and response biases while retaining similar levels of validity to Likert-type measures. Although research evidence supports the effectiveness of MFC measures, fairness issues resulting from gender biases in the use of MFC measures have not yet been investigated in the literature. Given the importance of gender equity in HR development, it is vital that new assessments improve upon known gender biases in the historical use of Likert-type measures and do not lead to gender discrimination in HR practices. In this vein, our investigation focuses specifically on potential gender biases in the use of MFC measures for HR development. Specifically, our study examines differential test-taker reactions and differential prediction of self-assessed leadership ability between genders when using the MFC personality measure. In an experimental study with college students, we found no evidence of gender differences in test-taker reactions to MFC

"This article was reviewed, and accepted by the editorship of Dr. Thomas Reio".

measures. In a second cross-sectional study with full-time employees, we found evidence of intercept differences, such that females were frequently underpredicted when using MFC personality measures to predict self-assessed leadership ability. Moreover, the pattern of differential prediction using MFC measures was similar to that of Likert-type measures. Implications for MFC personality measures in applied practice are discussed.

KEYWORDS
differential validity, gender bias, multidimensional forced-choice format, personality tests, personnel selection, test fairness, test-taker reaction

1 | INTRODUCTION

In human resource (HR) development, personality tests have been considered one of the most important assessment tools for purposes such as selection, vocational discernment, and self-assessment (Hough et al., 2015). They predict important outcomes such as leadership (Bono & Judge, 2004), teamwork and team performance (Peeters et al., 2006), overall job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000), and organizational citizenship behaviors (Borman et al., 2001). Outside of selection, personality tests are often used for HR development in areas such as vocational counseling (McCrae & Costa Jr, 1991) and personal development (e.g., employee self-reflection; Moyle & Hackston, 2018). Furthermore, personality tests are relatively easy to administer and inexpensive (Anderson et al., 2010).

Despite its merits and popularity, Likert-type personality measures (e.g., 1 = Strongly Disagree to 5 = Strongly Agree), which are commonly used in HR development, have been criticized due to problems of faking (Wetzel, Frick, & Brown, 2020). This is particularly concerning in selection settings, where applicants are motivated to fake their responses to be hired. The resulting responses could alter the rank order of job applicants and distort the factor structure, reliability, and validity evidence of personality tests, consequently harming the utility of selection systems (e.g., Birkeland et al., 2006; Komar et al., 2008; Zickar et al., 2004).

In response to faking and other response bias concerns surrounding Likert-type tests, organizations have increasingly turned to the use of multidimensional forced-choice (MFC) measures (e.g., *StrengthsFinder* [Rath, 2007]; the Tailored Adaptive Personality Assessment System [Stark et al., 2014]; the Adaptive Employee Personality Test [Adept-15; Boyce et al., 2014]). MFC measures present multiple statements within an item block and ask respondents to rank each statement from "most like me" to "least like me" or choose "most like me" or "least like me." The MFC design incorporates items measuring different psychological traits within each block of statements, matched based on a similar level of social desirability and/or item extremity. As a result, it is impossible for respondents to equally endorse all items with high social desirability, thus reducing the potential for faking (Wetzel, Frick, & Greiff, 2020). Many researchers have shown that MFC measures successfully reduce faked score inflation (e.g., Cao & Drasgow, 2019; Fisher et al., 2019; Hirsh & Peterson, 2008; Lee et al., 2019; 2021; Martinez Gómez & Salgado, 2021; Wetzel, Frick, & Brown, 2020), while maintaining criterion-related validity similar to that of Likert-type measures (e.g., Bartram, 2007; Lee et al., 2018; O'Neill et al., 2017; Salgado & Tauriz, 2014; Zhang et al., 2020). For example, meta-analytic findings suggested that the average mean difference in MFC measures between high-stakes settings (i.e., either applicant-incumbent designs or simulated selection scenarios) and low-stakes settings (i.e., participants informed that the test was for research purposes only) was only 0.06 (Cao & Drasgow, 2019), which

was considerably smaller than those of Likert-type measures in the previous meta-analytic research (e.g., 0.48–0.65 from Viswesvaran & Ones, 1999; 0.11–0.45 from Birkeland et al., 2006).

Despite these substantial advances in the use of MFC measures in HR selection, there have been very few investigations into the potential gender differences in test-taker reactions to and validity evidence for MFC measures. As Hough and Oswald (2008) note, "Most organizations around the world are interested in fair employment practices" (p. 284). Given growing demands for equity and fairness in testing situations, especially in personnel selection (Konradt et al., 2017; Stobart & Eggen, 2012; Truxillo et al., 2015), it is of vital importance to examine whether MFC measures are equitable and fair to be used for assessment in selection. For example, if MFC measures engendered more negative test-taker reactions among underrepresented groups (e.g., women), or if they systematically underpredicted performance among underrepresented groups, then widespread use of MFC measures for HR selection could potentially exacerbate gender differences and inequity in hiring. Nevertheless, to date, this critical issue has not been appropriately investigated in the literature. To address the issue, this study focuses on gender differences using MFC personality measures, compared with traditional Likert-type personality measures, in terms of test-taker reactions (Study 1) and predictive bias (Study 2). Specifically, we employ the induced selection scenario design with self-report measures (e.g., Byle & Holtgraves, 2008; Cao & Drasgow, 2019) to simulate employee selection; though limitations of this design are discussed later, this design approximates the experiences that actual applicants may have when taking personality tests.

2 | TEST-TAKER REACTIONS TO PERSONALITY MEASURES ACROSS GENDERS

Test-taker reactions to assessments used in HR development are of vital importance to both the individual and the organization. At the individual level, test-taker reactions are related to the intention to accept a job offer and extend to one's overall intention to support an organization (e.g., "That was such a bad interview experience, I would never spend money at that business again."), and even future work performance upon hiring (Konradt et al., 2017; J. M. McCarthy et al., 2013; Rynes & Barber, 1990; Smither et al., 1993). At the organizational level, test-taker reactions are important to maintaining the organization's public renown, reducing the risk of litigation, and encouraging future candidates to apply (Harris et al., 2020; Hausknecht et al., 2004; J. M. McCarthy et al., 2017). Thus, maintaining positive test-taker reactions to tests used in personnel selection is of critical importance for organizations. Test-taker reactions are also related to the perceived efficacy and value of personality tests used in settings such as vocational counseling and employee self-development (Lundgren et al., 2019).

Test-taker reactions that potentially differ between genders could create problems for organizations, especially in terms of gender equity. For example, differential reactions could negatively affect recruitment. If one gender group shows more strong negative reactions to personnel assessments or procedures than the other group, they would be less likely to apply to the job (Hausknecht et al., 2004) or more likely to dissuade others from applying to the job (Smither et al., 1993), thus reducing the pool of qualified applicants and subsequently lead to differential hiring rates (Newman et al., 2014). Moreover, the concerns over an organization's public reputation and litigation risk could likewise lead to reduced hiring rates among protected gender groups. This is particularly relevant for gender-unbalanced industries or domains, where there are already substantial differences in the size of the pool of qualified applicants (e.g., engineering or social services; see https://www.bls.gov/cps/cpsaat11.htm). Thus, it is crucial for organizations to ensure that they are attracting equal numbers of qualified applicants between gender groups. Thus, Study 1 focuses on identifying potential gender differences in test-taker reactions to MFC and Likert-type personality tests.

Prior studies have focused primarily on how test-taker reactions to MFC measures differ from reactions to Likert-type measures, given the fact that response processes to the two types of measures differ. Specifically, Likert-type measures elicit absolute decision-making (i.e., simply choose the degree of agreement within a single statement), whereas MFC measures elicit comparative decision-making (i.e., compare the relative preferences among

multiple statements within an item block and determine their ranks). Harland (2003) found that MFC personality measures elicited more negative test-taker reactions than Likert-type measures in a leadership development context, especially in terms of perceived accuracy, usefulness, and respectfulness. These findings also have been replicated in a personnel selection context (e.g., Converse et al., 2008). More recently, Dalal et al. (2019) proposed a self-concept theory explanation of these findings, suggesting that MFC measures threaten the test taker's self-concept due to restriction of choice, forced endorsement of negative items, and lack of feedback. They found that test-taker reactions indeed improved when these factors were addressed, using graded ranking MFC measures (as opposed to pairwise preference), removing socially undesirable items, and providing post-assessment feedback.

Our study is the first to examine gender differences in test-taker reactions for MFC measures, but there is empirical evidence to suggest that gender differences might exist in the use of MFC measures. First, research on decision-making suggests that gender differences in the decision-making process could impact test-taker reactions. For example, De Acedo Lizárraga et al. (2007) found that "[females] are more concerned with uncertainty, doubts, and the dynamism that are involved in the decision" (p. 387). Other studies have also found that females feel much less comfortable with guessing in test-taking (Adam, 1999; Baldiga, 2014). Applied to MFC measures, this would suggest that females might be more likely to react negatively due to the forced-choice nature of comparative decision-making, thus leading to increased ratings on perceived cognitive demand.

Additionally, after considering the influences of social roles in the workplace, men and women might react differently to forced-choice questions (see Social Role Theory; Eagly et al., 2000). For example, if the item block consists of statements favoring social roles or expectations toward males (e.g., ranking "I am assertive" over "I easily make friends"), females may feel that the MFC items are less predictive of their future performance, thus leading to more negative test-taker reactions in terms of perceived validity. Finally, Lishner et al. (2008) explored different types of forced-choice measures of sexual and emotional infidelity, finding gender differences in how upsetting participants found each of the measures. Though a completely different content area, this suggests that gender differences might exist in test-taker reactions to forced-choice measures. Altogether, this lends support to our argument that there could be important gender differences in test-taker reactions to MFC measures (compared to Likert-type measures). These gender differences could lead to negative downstream effects in HR management, such as lower offer acceptance rates, decreased diversity of the applicant pool, and growing mistrust in personality tests for developmental purposes. Thus, we propose our first set of hypotheses and research questions:

H1. Gender differences exist in test-taker reactions to Likert-type and MFC personality measures.

RQ1. Are there different patterns of gender differences between test-taker reactions to Likert-type compared with MFC personality measures?

3 | DIFFERENTIAL PREDICTION BY PERSONALITY ACROSS GENDER GROUPS

Our second study focused on gender differences in terms of predictive bias. Personality tests can be used as a predictor for several different outcomes that organizations might be interested in. For formal selection purposes, personality tests would be used to predict desirable job-related outcomes; but in other HR development settings, personality tests could also be used as an employee self-assessment tool (e.g., in a leadership training "course" offered to encourage employee development and professional growth; Moyle & Hackston, 2018). One important way to look at the fairness and equity of personality tests is to examine for differential prediction. Differential prediction is a form of test bias, which is when "some aspect of the test [causes] it to work systematically differently across subgroups" (Berry, 2015, p. 442). Specifically, differential prediction occurs when "for a given subgroup, consistent nonzero errors of prediction are made for members of the subgroup" (Society for Industrial and

Organizational Psychology, 2003, p. 32). Statistically, differential prediction can take the form of slope differences (i.e., differences in the validity coefficient between groups) and/or intercept differences (i.e., systematic over- or under-prediction for a given subgroup). In other words, if personality tests show differential prediction between genders, this means they will systematically overpredict and underpredict a given outcome variable for the two groups.

This has important practical implications when the test is used for HR development. In a selection setting, if personality does not predict job-related outcomes as well for females compared with males, then there would be more errors involved in using personality as a selection tool for females, which could result in unfair hiring decisions (Berry et al., 2013). Moreover, if personality tests systematically underpredict female scores on job-related outcomes, then hiring decisions made based on a single cut score on the predictor from a common regression line (which is required by law; Saad & Sackett, 2002) could lead to fewer females being selected than would deserve to be selected (Berry et al., 2013). In other HR development settings, if personality tests show differential prediction between genders on a self-assessed outcome (e.g., self-perceptions of leadership potential), then one gender might be systematically underrepresented in terms of their perceived leadership abilities. Taking this a step further, if personality tests are used to identify high-potential employees (e.g., "Hi-Po's", see Bialek & Hagen, 2022), it could also lead to implications for women's self-selection for leadership roles. Thus, the investigation of differential prediction is crucial to understanding whether the use of personality tests for HR development is fair and equitable based on gender.

Very few studies have tested the differential prediction of personality tests, and those that have been published have shown somewhat inconsistent results and exclusively focused on Likert-type personality tests. In Saad and Sackett's (2002) ground-breaking study on the topic, they used three personality composites to predict five job performance dimensions among military participants, and they found little to no slope differences but substantial intercept differences. Interestingly, the intercept differences resulted in overprediction of females, and most were found when personality was used to predict leadership performance outcomes. Berry et al. (2013) built on this study, expanding it to all Big Five personality factors and to nonmilitary populations (specifically, two samples of middle managers at a US energy company, and Chinese MBA students). They generally replicated earlier findings that Likert-type personality measures showed little evidence of differential prediction, with only 3.3% of cases showing differential prediction across both samples. On the other hand, Duehr (2006) used the Big Five personality test to predict nine dimensions of the transformational-transactional leadership behavior scale. Drawing from gender roles theory, Duehr hypothesized and found that 44% of cases showed intercept differences that underpredicted female scores on desirable leadership behaviors, but overpredicted female scores on undesirable leadership behaviors. In short, the research on differential prediction among personality tests is inconclusive and severely lacking; Berry et al. (2013) recommended extensive future research with larger sample sizes, employees from different companies or industries, and using different types of personality measures. Since then, there has been little to no effort in answering these important questions to our knowledge.

In our study, we focus on leadership outcomes as opposed to generic performance for several reasons. First, there is growing consensus that personality is an important predictor of leadership behaviors above and beyond the influence of the situation (Bono & Judge, 2004). As a result, personality tests have become more popular in personnel selection for leadership positions (Salgado & De Fruyt, 2017). Moreover, as Duehr (2006) extensively discusses, gender stereotypes are particularly strong with regard to leadership, which could lead to more evidence showing differential prediction (contrary to the lack of evidence found by Berry et al., 2013). Duehr (2006) argues that personality traits that are stereotypically masculine or feminine "may function differently for men and women in the prediction of transformational leadership… [and] influence the degree to which identical levels of a personality trait are predictive of leadership behavior for men relative to women" (p. 61). Moreover, prior studies have highlighted known mean differences between genders on both the predictor side (personality; Weisberg et al., 2011) and the outcome side (leadership; Stelter, 2002). Thus, if women, in fact, demonstrate more people-oriented leadership behavior (i.e., the mean difference in the outcome variable), which is driven primarily by the agreeableness personality trait (De Vries, 2012), and women tend to score higher on agreeableness (i.e., the mean difference in the predictor variable), then it follows that women would be expected to have higher performance on a measure of people-oriented leadership compared with men. Furthermore, following the recommendation of an anonymous reviewer, we also investigate how the combination of Big Five personality traits might differently predict leadership

characteristics between men and women. That is, some personality traits might be more predictive of specific leadership behaviors (relative to other personality traits) for men when compared with women. Finally, recently growing calls for "breaking the glass ceiling" center around the lack of females in leadership positions (Johns, 2013; Kalaitzi et al., 2017; Sims et al., 2021). Public policy and public opinion likewise are highly concerned with gender equity in leadership; for example, California mandated in 2018 that all publicly traded companies have at least one female board director (Jamali, 2020). Thus, it is important to ensure that the selection methods used to appoint organizational leaders do not show gender biases.

Specifically, we examined differential prediction in MFC and Likert-type measures predicting four popular leadership outcomes: *task-oriented leadership* (i.e., "initiating structure"), *people-oriented leadership* (i.e., "consideration," Stogdill, 1963), *charismatic leadership* (House & Howell, 1992), and *ethical leadership* (M. E. Brown & Treviño, 2006). These four leadership outcomes, while by no means an exhaustive list of the variety of theories of leadership, are among the most popular and have been used as key outcome variables in prior studies linking personality to leadership (e.g., De Vries, 2012). Moreover, Duehr's (2006) study focused solely on transformational leadership with only three personality predictors as opposed to five. Given recent concerns over the internal validity of transformational leadership as a construct (Van Knippenberg & Sitkin, 2013), we argue that it is wise to extend the existing findings to the Big Five and a more comprehensive leader behavior outcome model.

We are among the first to consider the important aspect of the type of personality measure. If MFC measures produce greater levels of differential prediction across genders than the Likert-type measures have shown, this would severely threaten their usability and fairness in a selection context. To our knowledge, only one study investigated the differential prediction of MFC personality measures so far. Nye et al. (2020) recently examined the differential validity of an MFC personality measure between occupational classes in a military context. Their results indicated that the correlations from personality to attrition did in fact differ in size between military occupational specialties. However, the subgroups used in this study were *occupational specialties* within the military (e.g., infantry, medics, motor transport operators), as opposed to demographic subgroups such as gender and race. Given the aforementioned legal concerns over differential prediction between protected subgroups such as gender and race, our study adds unique information about the fairness of MFC personality tests in a legal context by focusing on gender differences. If overlooking possible gender biases of MFC personality measures relative to Likert-type personality measures, efforts to use MFC measures in HR development could be subject to issues of fairness, equity, and legality. Because the literature does not clearly state a direction of the effect (e.g., how much bias there should be), we leave these as exploratory research questions:

RQ2. Is there evidence of differential prediction between genders when using an MFC personality measure to predict leadership-related outcomes?

RQ3. How does differential prediction differ across Likert and MFC personality measures?

4 | STUDY 1: DIFFERENTIAL TEST-TAKER REACTIONS BETWEEN GENDER GROUPS

4.1 | Methods

4.1.1 | Participants and procedures

Participants were undergraduate psychology students at a large land-grant mid-Atlantic public research university. Students received course credit in exchange for participation in this study. A total of 306 full-time students were recruited through the university's Psychology Research Participation System. Participants were told to imagine they had been recruited by a large private organization and asked to take a series of personality assessments. Participants

were randomly assigned to take either the Likert-type personality measure or the MFC personality measure, followed by the test-taker reaction questions. After 2 weeks, participants were invited via email to return to complete the opposite personality measure format that they did not complete the first time. For example, if they completed Likert-type items in the first session, they completed MFC items in the second session. Therefore, the data collection of the Likert-type and MFC personality measures was counter-balanced. Fifty-one participants failed to attend a second lab session; these participants and two others who did not report gender were excluded from analyses. Thus, 253 participant records were retained for use in the study. The sample was primarily female (64%).

4.1.2 | Measures

Participants completed the Big Five personality items in a 20-triplet MFC format (i.e., 60 personality items appearing in sets of three) used by A. Brown and Maydeu-Olivares (2011). For the Likert-type format, the same 60 items that were in the MFC format were converted to Likert-type statements. For test-taker reactions, participants completed Likert-type items consisting of (a) six items measuring perceived cognitive load (e.g., "I feel this test is easy to complete"; Converse et al., 2008), (b) two items measuring perceived respectfulness (e.g., "I feel this test is respectful of my feelings"; Harland, 2003), (c) five items measuring perceived validity (e.g., "I feel this test is relevant to the job."; Bauer et al., 2001; J. McCarthy et al., 2009), and (d) three items measuring perceived accuracy (e.g., "I feel this test allows me to accurately depict my personality."; Speer et al., 2016). Internal consistency (omega coefficient) for each measure was 0.71, 0.76, 0.75, and 0.69.

4.2 | Results

We compared test-taker reactions between genders on four different aspects: perceived validity, perceived accuracy, perceived cognitive load, and perceived respectfulness (i.e., the main effect of gender). We conducted mixed factor ANOVAs using the *rstatix* package (Kassambara, 2021) with response format (MFC vs. Likert) as a within-subjects factor and gender (male vs. female) as a between-subjects factor, after controlling for the effects of order (group A which completed Likert first then MFC vs. group B which completed MFC first then Likert). *p*-values were then adjusted for multiple comparisons (4 DVs) using the Benjamini–Hochberg correction (Huang, 2020).[1] Table 3 reports the ANOVAs with adjusted *p*-values, and Figure 1 visualizes the results. Overall, the main effect of gender was not significant for the four dependent variables. Thus, Hypothesis 1 (i.e., gender differences exist in test-taker reactions to Likert-type and MFC personality measures) was not supported. We noted however that the response format was significant for all four test-taker reactions, such that applicants reacted more negatively to the MFC measure. This finding is in line with prior studies on MFC test-taker reactions (e.g., Converse et al., 2008). Moreover, in response to RQ1 (i.e., different patterns of gender differences between test-taker reactions to Likert-type compared to MFC personality measures), the interaction effect was not significant for all four dependent variables. Thus, we concluded that there was no evidence of gender differences in test-taker reactions.

5 | STUDY 2: DIFFERENTIAL PREDICTION ACROSS GENDER GROUPS

5.1 | Methods

5.1.1 | Participants and procedure

For our second study on differential prediction, data were collected from an online sample of full-time working adults via Amazon Mturk, an online crowdsourcing website for survey participants. Study 2 included adults over the age of

TABLE 1 Means, SDs, correlations, and effect sizes for Study 1 variables.

		Likert-type		MFC		MFC vs. Likert	Correlations with…			
Variable	Gender	Mean	SD	Mean	SD	Cohen's *d*	Validity	Accuracy	Load	Respect
Perceived Validity	Male	3.16	0.67	2.94	0.72	−0.38	1.00	-	-	-
	Female	3.25	0.60	2.99	0.65					
	Combined	3.22	0.62	2.97	0.67					
Perceived Accuracy	Male	2.90	0.74	2.64	0.88	−0.27	0.68	1.00	-	-
	Female	2.93	0.79	2.73	0.76					
	Combined	2.92	0.77	2.71	0.81					
Perceived Cognitive Load	Male	2.12	0.46	2.34	0.64	0.38	−0.23	−0.19	1.00	-
	Female	2.21	0.47	2.40	0.54					
	Combined	2.18	0.47	2.38	0.58					
Perceived Respectfulness	Male	3.57	0.73	3.30	0.74	−0.35	0.54	0.53	−0.20	1.00
	Female	3.60	0.70	3.35	0.77					
	Combined	3.59	0.71	3.33	0.76					

Note: Validity = Perceived Validity; accuracy = Perceived Accuracy; load = Perceived Cognitive Load; respect = Perceived Respectfulness.
Abbreviation: MFC, multidimensional forced-choice.

18 who work at least 32 h/week. In addition, we used quotas so that we could collect an equal number of male and female responses, which addresses concerns over unequal proportions distorting earlier findings (Saad & Sackett, 2002). Participants completed a 45-minute online survey including both Likert-type and MFC questions for Study 2. One thousand eighty-three responses were initially collected in March 2020 and were paid 3.00 USD each. After filtering out participants who failed at least one of the three attention check questions (e.g., Please answer '2 – Somewhat Disagree' for this question; see Kung et al., 2018), we were left with a sample size of 876. The average age of the sample was 38.86 (SD = 10.18), and the sample was 68.84% white/Caucasian, 16.89% Asian or Asian-American, 6.96% Black or African American, and the rest other or mixed. Participants worked an average of 43.08 h/week (SD = 6.01) (Figures 2 and 3).

5.1.2 | Measures

After completing a set of demographic questions, participants responded to the same two personality measures (Likert-type and MFC) used in Study 1. The order of presentation of each type of personality measure was randomly distributed throughout the survey to minimize order effects. Next, we calculated empirical reliability estimates.[2] The reliability estimates for the Likert-type personality scales were comparable to prior studies of gender differences in personality (e.g., Weisberg et al., 2011): 0.90 for agreeableness, 0.89 for conscientiousness, 0.95 for extraversion, 0.94 for neuroticism, and 0.88 for openness. The reliability estimates of MFC measures were computed as well. Like the Likert-type scales, reliability estimates of the MFC measure were adequate—0.76 for agreeableness, 0.78 for conscientiousness, 0.83 for extraversion, 0.83 for neuroticism, and 0.75 for openness. These reliabilities are comparable with previous research findings of MFC personality measures (e.g., agreeableness = 0.70, conscientiousness = 0.75, extraversion = 0.83, neuroticism = 0.80, and openness = 0.72 from Lee et al., 2018).

Participants then completed a set of measures of outcome variables including four types of leadership behavior. These measures were presented at the end of the study, maximizing psychological separation from the independent

TABLE 2 Means, SDs, and correlations for Study 2 variables.

	Mean	SD	Likert_O	Likert_C	Likert_E	Likert_A	Likert_N	MFC_O	MFC_C	MFC_E	MFC_A	MFC_N	Task	Peop	Char	Ethic
Likert_O	−0.001	0.945	1.000													
Likert_C	−0.001	0.951	0.272	1.000												
Likert_E	0.000	0.974	0.362	0.231	1.000											
Likert_A	−0.001	0.954	0.339	0.370	0.340	1.000										
Likert_N	0.002	0.970	−0.402	−0.305	−0.400	−0.231	1.000									
MFC_O	−0.001	0.882	0.620	0.125	0.255	0.121	−0.357	1.000								
MFC_C	−0.029	0.889	0.154	0.645	0.101	0.178	−0.196	0.426	1.000							
MFC_E	0.009	0.912	0.236	0.021	0.783	0.234	−0.291	0.252	−0.027	1.000						
MFC_A	−0.030	0.883	0.193	0.222	0.299	0.657	−0.201	0.371	0.410	0.310	1.000					
MFC_N	0.017	0.910	−0.298	−0.212	−0.339	−0.174	0.770	−0.512	−0.362	−0.321	−0.353	1.000				
task	0.000	0.697	0.385	0.489	0.352	0.371	−0.249	0.197	0.275	0.207	0.216	−0.186	1.000			
peop	0.000	0.464	0.451	0.423	0.261	0.582	−0.362	0.240	0.223	0.151	0.365	−0.283	0.589	1.000		
char	0.000	0.662	0.437	0.448	0.459	0.486	−0.384	0.257	0.255	0.321	0.342	−0.315	0.773	0.759	1.000	
ethic	0.000	0.467	0.443	0.448	0.249	0.548	−0.323	0.224	0.244	0.130	0.328	−0.245	0.694	0.944	0.756	1.000

Note: Likert_A = factor score for Likert-type measure of agreeableness, Likert_C = factor score for Likert-type measure of conscientiousness, Likert_E = factor score for Likert-type measure of extraversion, Likert_N = factor score for Likert-type measure of neuroticism, Likert_O = factor score for Likert-type measure of openness, MFC_O = factor score for MFC measure of openness, MFC_C = factor score for MFC measure of conscientiousness, MFC_E = factor score for MFC measure of extraversion, MFC_A = factor score for MFC measure of agreeableness, MFC_N = factor score for MFC measure of neuroticism, task = factor score for task-oriented leadership, peop = factor score for people-oriented leadership, char = factor score for charismatic leadership, and ethic = factor score for ethical leadership.
Abbreviation: MFC, multidimensional forced-choice.

TABLE 3 Univariate ANOVA results from Study 1.

Dependent variable	Gender	Response format	Interaction
Perceived Validity	$F(1, 249) = 1.034$, $p = 0.511$, $\eta^2 = 0.003$	$F(1, 249) = 31.125$, $p < 0.001$, $\eta^2 = 0.032$	$F(1, 249) = 0.273$, $p = 0.733$, $\eta^2 < 0.001$
Perceived Accuracy	$F(1, 249) = 0.659$, $p = 0.622$, $\eta^2 = 0.002$	$F(1, 249) = 21.183$, $p < 0.001$, $\eta^2 = 0.019$	$F(1, 249) = 0.648$, $p = 0.622$, $\eta^2 = 0.001$
Perceived Cognitive Load	$F(1, 249) = 1.683$, $p = 0.366$, $\eta^2 = 0.005$	$F(1, 249) = 32.790$, $p < 0.001$, $\eta^2 = 0.034$	$F(1, 249) = 0.158$, $p = 0.807$, $\eta^2 < 0.001$
Perceived Respectfulness	$F(1, 249) = 0.361$, $p = 0.733$, $\eta^2 = 0.001$	$F(1, 249) = 28.953$, $p < 0.001$, $\eta^2 = 0.028$	$F(1, 249) = 0.002$, $p = 0.960$, $\eta^2 < 0.001$

Note: *p*-values are adjusted using the Benjamini–Hochberg correction. Results are after controlling for the effect of order. η^2 = eta-squared.

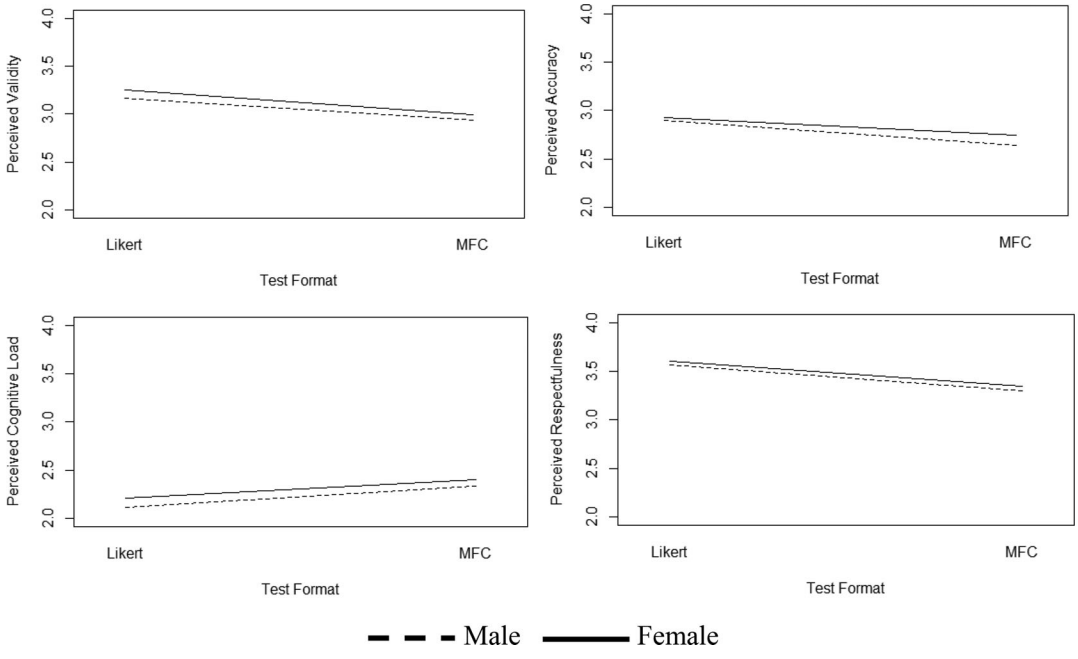

FIGURE 1 Interaction charts of gender and test format on test-taker reactions.

variables to reduce the likelihood of common method bias (Podsakoff et al., 2003). *Task-oriented leadership* was measured using five items from Schriesheim and Stodgill's (1975) revised Leader Behavior Description Questionnaire. We selected five items from the original 10-item measure based on the highest factor loadings (e.g., "I schedule the work to be done" and "I maintain definite standards of performance"). Internal consistency (omega coefficient) was 0.83. *People-oriented leadership* was similarly measured using 5 out of the 10 items from Schriesheim and Stogdill (1975) based on the highest factor loadings (e.g., "I treat all group members as my equals" and "I am friendly and approachable"). Internal consistency (omega coefficient) was 0.71. *Charismatic leadership* was measured using De Hoogh et al.'s (2005) four-item measure of inspirational motivation. Items included "I talk optimistically about the future" and "I express confidence that goals will be achieved." Internal consistency (omega coefficient) was 0.87. Finally, *ethical leadership* was measured using five items selected based on the highest factor loadings from M. E. Brown et al.'s (2005) 10-item measure of ethical leadership (e.g., "I have the best interests of employees in mind" and "I make fair and balanced decisions"). Internal consistency (omega coefficient) was 0.83.

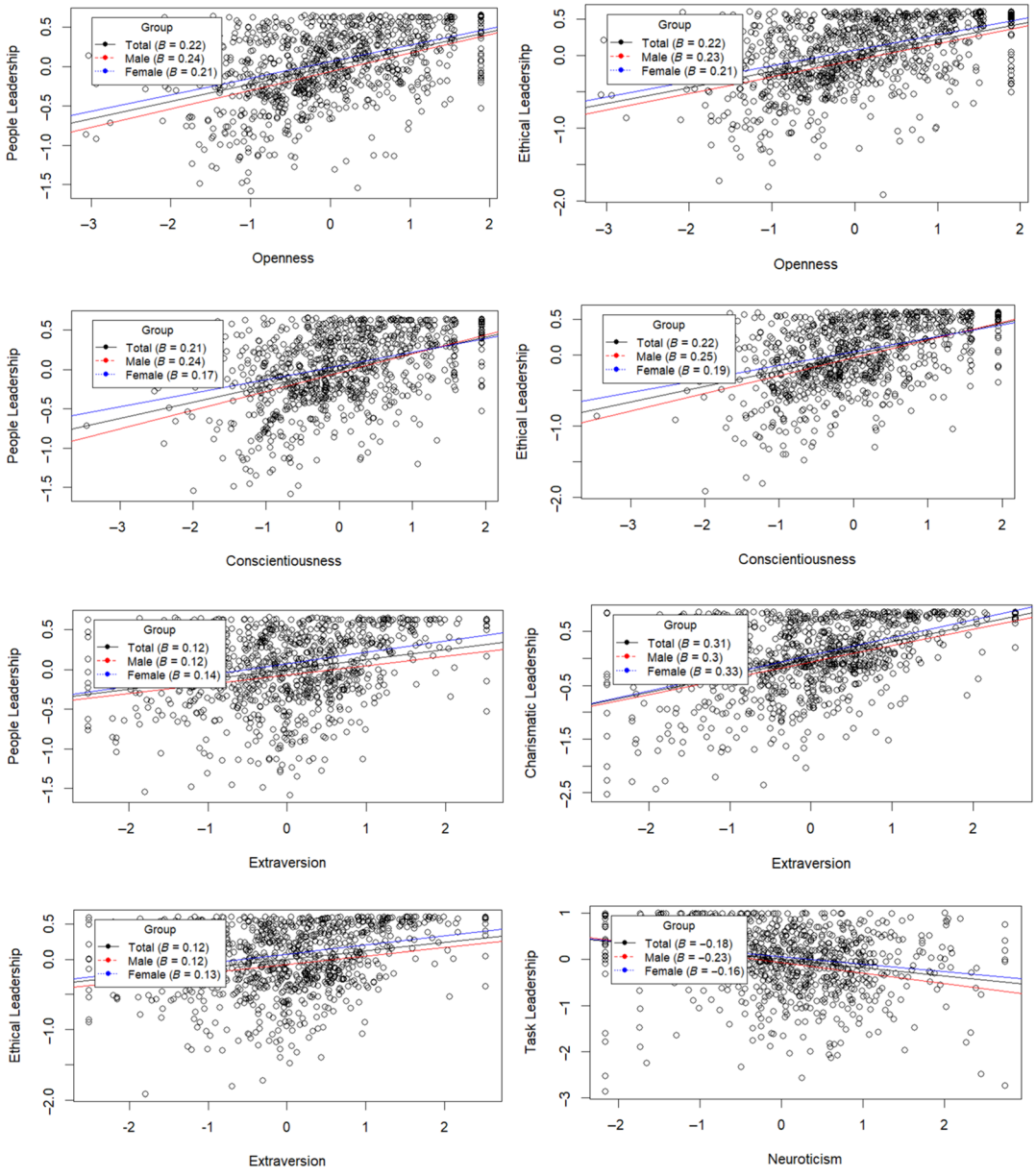

FIGURE 2 Graphs of significant differential prediction using Likert-type measures.

5.2 | Results

5.2.1 | Preliminary analyses

Because both predictors and outcomes were collected via self-reported measures, we first assessed the potential of common method bias by conducting Harman's single-factor test (Podsakoff et al., 2003). We found that the common factor extracted only accounted for 22% of the variance. According to conventional cutoffs (e.g., 50% or 70%, see Fuller et al., 2016 for a simulation study), this suggests that common method variance is not substantially present in

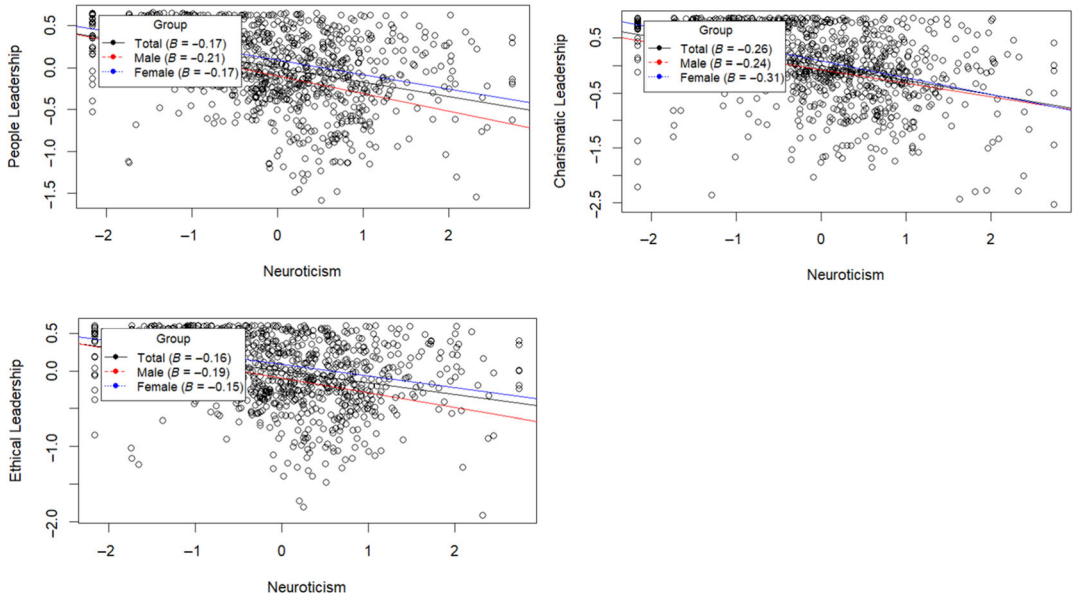

FIGURE 2 (Continued)

our data. Due to recommendations that Harman's test can be biased downward due to the number of variables, we scored the personality data (five predictors) and leadership data (four outcomes) separately to partially control for potential common method bias (Podsakoff et al., 2003; see next paragraph for details).

Participants' latent traits on the MFC triplet data were scored using a Thurstonian item response theory (TIRT) model with ULSMV estimator (A. Brown & Maydeu-Olivares, 2012) in Mplus 8 (Muthén & Muthén, 1998–2017). We refer readers to A. Brown and Maydeu-Olivares (2012) for details of the TIRT model scoring process. Model fit was as follows: CFI = 0.872, TLI = 0.863, RMSEA = 0.033 (90% CI = 0.032, 0.025). We note that recent MFC study using the TIRT model also reported similar fit results with our findings (e.g., RMSEA = 0.03, CFI = 0.85, and SRMR = 0.098 from Guenole et al., 2018; RMSEA = 0.04, CFI = 0.90, TLI = 0.88 from Morillo et al., 2016; RMSEA = 0.03, CFI = 0.89, TLI = 0.89 from Lee et al., 2018).

Next, to stay consistent with the IRT-based estimation used for the MFC measure, we estimated scores on the Likert-type personality measure using the multidimensional version of the graded response model (Samejima, 1997) in R using the *mirt* package with the default expected a-posteriori factor score estimation (Chalmers, 2012). Fit statistics at the item-level were good, with an average RMSEA of 0.015 across 60 items and a maximum RMSEA of 0.046. Finally, scores on the leadership outcome measures were estimated using traditional CFA using the MLR estimator to account for potential non-normality. Model fit was as follows: CFI = 0.891, TLI = 0.873, RMSEA = 0.078 (90% CI = 0.072, 0.084), SRMR = 0.067.

5.2.2 | Primary analyses

To assess for differential prediction, we followed the step-down hierarchical regression procedure outlined by Lautenschlager and Mendoza (1986). In this method, four regression models are fitted: Model 1 (predictor only), Model 2 (predictor, demographic variable, and product of the two), Model 3 (predictor and product only), and Model 4 (predictor and demographic variable only). In Step 1, Model 1 is compared with Model 2; a significant improvement in prediction suggests that bias exists because of the demographic variable (nonsignificant results mean there is no

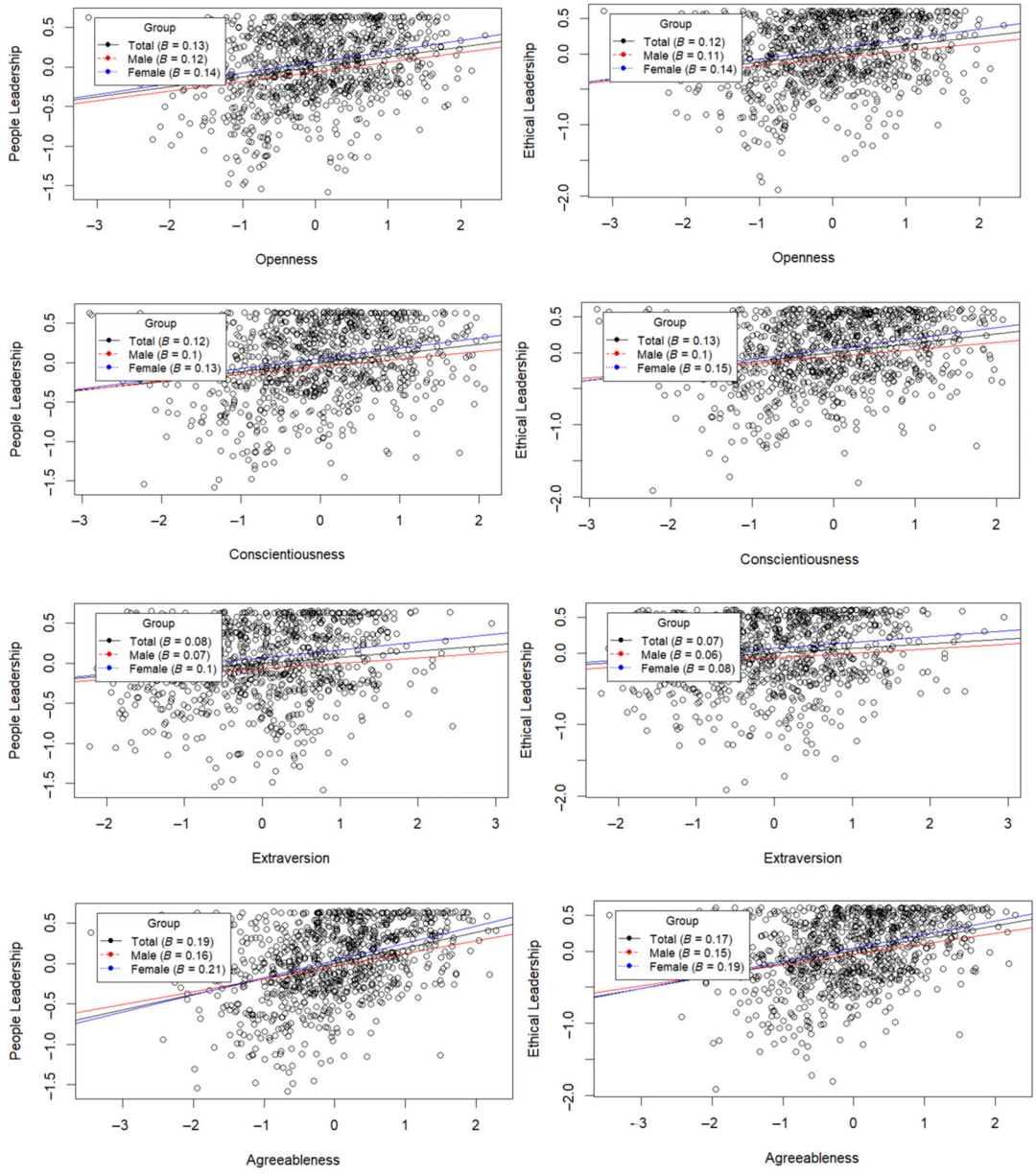

FIGURE 3 Graphs of significant differential prediction using multidimensional forced-choice measures.

bias, so no further testing is conducted). For Step 2, Model 2 is compared with Model 4. If significant, it suggests that there are slope differences, and Step 3 compares Model 3 with Model 2 to test whether there are also intercept differences. If Step 2 is not significant, it suggests that there are no slope differences, and Step 3 compares Model 4 with Model 1 to test whether there are intercept differences instead. We refer interested readers to Lautenschlager and Mendoza (1986) for more details on the differential prediction testing procedure.

We first tested for evidence of differential prediction between genders when using an MFC personality measure to predict self-reported leadership outcomes. Table 4 presents the results of differential prediction for MFC measures. Across the 20 analyses (each of five personality dimensions with each of four outcome variables), slope

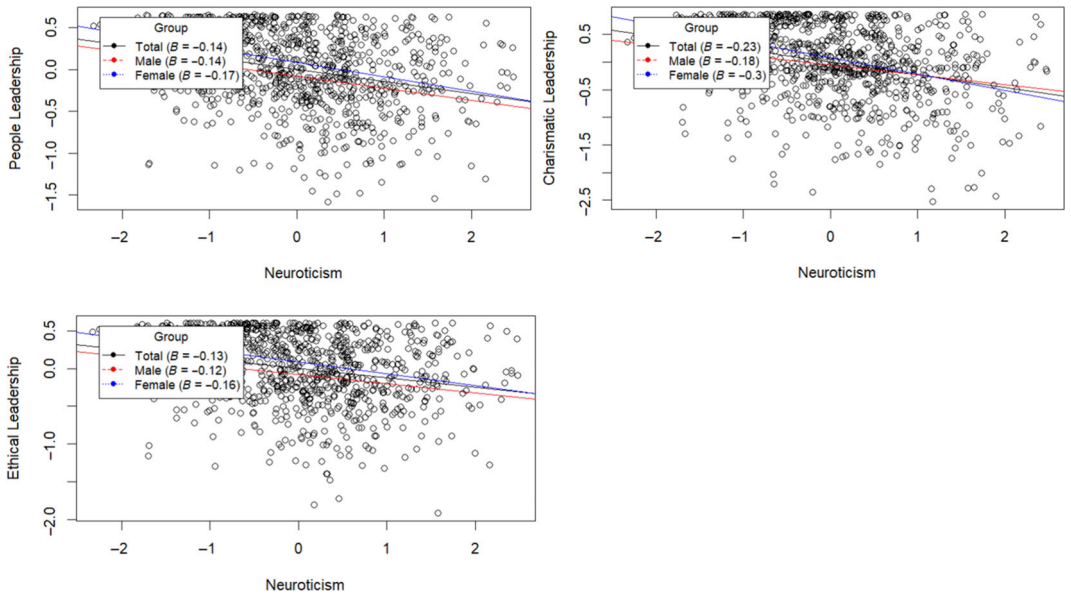

FIGURE 3 (Continued)

differences were identified in just one case: Neuroticism predicting charismatic leadership ($\Delta R^2 = 0.010$, $p < 0.01$). Following the interpretation method demonstrated by Berry et al. (2013), this means that MFC personality scores were differentially related to self-reported leadership outcomes between genders in 5% of cases, which is around what might be expected to occur as a chance phenomenon given the large number of pairwise comparisons being tested. Thus, there do not seem to be meaningful slope differences. However, there were numerous instances of intercept differences; 10 of the 20 cases (i.e., 50%) using MFC measures showed intercept differences, which were far greater than the low levels of differential prediction reported previously (Berry et al., 2013). In general, intercept differences were primarily found in the prediction of people-oriented leadership and ethical leadership. Moreover, all instances of intercept differences resulted in an underprediction of female scores, much like what Duehr (2006) found. Thus, we concluded that the MFC personality measures showed evidence of differential prediction, primarily in intercept differences, when used to predict job-related outcomes; specifically, female scores on leadership outcomes tended to be underpredicted (Tables 5 and 6).

We also compared this pattern with the differential prediction results of Likert-type personality measures (see Table 7). Overall, there was a consistent match. Likert-type measures showed two cases of slope differences (Conscientiousness predicting people-oriented leadership [$\Delta R^2 = 0.008$, $p < 0.01$] and ethical leadership [$\Delta R^2 = 0.008$, $p < 0.01$]), like the one case found using MFC measures; thus, there do not appear to be meaningful slope differences. Additionally, Likert-type measures showed nine cases of intercept differences, like the 10 cases found using MFC measures. Once again, all instances of intercept differences resulted in an underprediction of female scores. Moreover, of almost all the combinations where Likert-type measures showed differential prediction, MFC measures showed differential prediction in the same combinations. For both measures, the differential prediction was most common when predicting people-oriented and ethical leadership. Lastly, in most of the cases, the amount of differential prediction was very small, with the change in R^2 less than 0.1% in about half of the cases. In general, the change in R^2 in cases of differential prediction was smaller when using MFC measures compared with Likert-type measures. Overall, we concluded that there was little to no difference in differential prediction between Likert-type and MFC personality measures.

TABLE 4 Tests of differential prediction of MFC personality across gender groups.

Predictor	Outcome	Step 1	Step 2	Step 3	Conclusion
Openness (MFC)	Task leadership	$\Delta R^2 = 0.004$, $\Delta F = 1.788$			No bias
	People leadership	$\Delta R^2 = 0.018$, $\Delta F = 8.680^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.194$	$\Delta R^2 = 0.018$, $\Delta F = 17.183^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.005$, $\Delta F = 2.397$			No bias
	Ethical leadership	$\Delta R^2 = 0.020$, $\Delta F = 9.545^{**}$	$\Delta R^2 = 0.001$, $\Delta F = 0.886$	$\Delta R^2 = 0.019$, $\Delta F = 18.207^{**}$	Intercept differences only[b]
Conscientiousness (MFC)	Task leadership	$\Delta R^2 = 0.002$, $\Delta F = 0.885$			No bias
	People leadership	$\Delta R^2 = 0.014$, $\Delta F = 6.460^{**}$	$\Delta R^2 = 0.001$, $\Delta F = 0.849$	$\Delta R^2 = 0.013$, $\Delta F = 12.074^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.004$, $\Delta F = 1.948$			No bias
	Ethical leadership	$\Delta R^2 = 0.016$, $\Delta F = 7.482^{**}$	$\Delta R^2 = 0.002$, $\Delta F = 1.921$	$\Delta R^2 = 0.014$, $\Delta F = 13.029^{**}$	Intercept differences only[b]
Extraversion (MFC)	Task leadership	$\Delta R^2 = 0.004$, $\Delta F = 1.664$			No bias
	People leadership	$\Delta R^2 = 0.021$, $\Delta F = 9.424^{**}$	$\Delta R^2 = 0.001$, $\Delta F = 0.892$	$\Delta R^2 = 0.020$, $\Delta F = 17.958^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.006$, $\Delta F = 2.986$			No bias
	Ethical leadership	$\Delta R^2 = 0.021$, $\Delta F = 9.468^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.283$	$\Delta R^2 = 0.021$, $\Delta F = 18.669^{**}$	Intercept differences only[b]
Agreeableness (MFC)	Task leadership	$\Delta R^2 < 0.001$, $\Delta F = 0.062$			No bias
	People leadership	$\Delta R^2 = 0.008$, $\Delta F = 4.241^{*}$	$\Delta R^2 = 0.003$, $\Delta F = 2.734$	$\Delta R^2 = 0.005$, $\Delta F = 5.736^{*}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.002$, $\Delta F = 1.051$			No bias
	Ethical leadership	$\Delta R^2 = 0.009$, $\Delta F = 4.313^{*}$	$\Delta R^2 = 0.001$, $\Delta F = 1.392$	$\Delta R^2 = 0.007$, $\Delta F = 7.223^{**}$	Intercept differences only[b]
Neuroticism (MFC)	Task leadership	$\Delta R^2 = 0.006$, $\Delta F = 2.874$			No bias
	People leadership	$\Delta R^2 = 0.031$, $\Delta F = 15.083^{**}$	$\Delta R^2 = 0.001$, $\Delta F = 0.678$	$\Delta R^2 = 0.030$, $\Delta F = 29.499^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.016$, $\Delta F = 7.904^{**}$	$\Delta R^2 = 0.006$, $\Delta F = 6.369^{*}$	$\Delta R^2 = 0.010$, $\Delta F = 9.602^{**}$	Slope & intercept differences
	Ethical leadership	$\Delta R^2 = 0.031$, $\Delta F = 14.771^{**}$	$\Delta R^2 = 0.001$, $\Delta F = 1.020$	$\Delta R^2 = 0.030$, $\Delta F = 28.520^{**}$	Intercept differences only[b]

Note: Differential prediction was conducted following the steps from Lautenschlager and Mendoza (1986).
Abbreviation: MFC, multidimensional forced-choice.
[a]Females are overpredicted.
[b]Females are underpredicted.
$^{*}p < 0.05$; $^{**}p < 0.01$.

TABLE 5 Cohen's *d* between genders on the MFC and Likert-type measures in Study 2.

		Male		Female		
Variable	Gender	Mean	SD	Mean	SD	Cohen's *d*
Openness	Likert-type	0.024	0.951	−0.026	0.940	0.053
	MFC	0.005	0.922	−0.054	0.839	0.066
Conscientiousness	Likert-type	−0.086	0.900	0.084	0.993	−0.179
	MFC	−0.082	0.858	0.023	0.917	−0.118
Extraversion	Likert-type	0.088	0.926	−0.089	1.010	0.183
	MFC	0.087	0.905	−0.069	0.913	0.172
Agreeableness	Likert-type	−0.208	0.903	0.206	0.960	−0.445
	MFC	−0.156	0.865	0.097	0.883	−0.290
Neuroticism	Likert-type	−0.177	0.911	0.180	0.996	−0.374
	MFC	−0.125	0.874	0.160	0.924	−0.317

Note: Mean differences were calculated on the factor scores on each of the Big Five (Likert-type and MFC) from the Graded Response Model and TIRT respectively.
Abbreviations: MFC, multidimensional forced-choice; TIRT, Thurstonian item response theory.

6 | DISCUSSION

Our research fits neatly into the landscape of research in the use of personality tests for HR development, by examining gender biases in the use of MFC measures of personality to predict self-assessed leadership. The main findings are as follows. First, we found no support for our H1 that there would be gender differences in test-taker reactions (e.g., perceived validity, perceived respectfulness). We also found no significant interactions between gender and test format in explaining test-taker reactions (RQ1). Second, we found little evidence for differential prediction in terms of slope differences between genders on MFC measures predicting self-assessed leadership (only 5% of cases), and some evidence of differential prediction between genders such that intercept differences led to females being generally underpredicted (50% of cases; RQ2). Third, we noted that the pattern of results for MFC measures was consistent with the pattern of results in differential prediction using Likert-type measures. Specifically, the MFC measure produced slightly fewer cases of slope differences and slightly more cases of intercept differences; however, the pattern of results in terms of where slope and/or intercept differences were found largely matched between Likert-type and MFC. This suggests that MFC measures display similar levels of differential prediction compared with Likert-type measures (RQ3).

Our research findings have important implications for the potential use of MFC personality measures in HR development. First, in terms of test-taker reactions, our findings were somewhat contrary to expectations based on theory, as there was no evidence of gender differences in test-taker reactions to MFC measures. It is possible that our null findings in this area are a result of the efforts made in the design of A. Brown and Maydeu-Olivares' (2011) five-factor personality test to balance dimension specifications and social desirability. Thus, our results provide some evidence that men and women would not respond differently to the use of a well-designed and balanced MFC measure. Our sample focused on undergraduate college students, many of whom likely will or have already completed personality tests in HR development contexts such as vocational counseling (McCrae & Costa Jr, 1991). Therefore, our results suggest that the use of personality tests in such a context (e.g., helping students assess their career directions) is unlikely to be differentially affected between genders in terms of test-taker reactions.

Additionally, while the sample was not identical to a real-life employee selection setting, it is a common study design used when one cannot obtain data from actual applicants and employees (e.g., 59 out of 74 studies in the Cao & Drasgow, 2019, meta-analysis). Thus, while not ideal, Study 1 could approximate applicant reactions to

TABLE 6 Means, SDs, and correlations for Study 2 variables: (a) male-only and (b) female only.

	Likert_O	Likert_C	Likert_E	Likert_A	Likert_N	MFC_O	MFC_C	MFC_E	MFC_A	MFC_N	task	peop	char	ethic
(Male-only group)														
Likert_O	1.000													
Likert_C	0.398	1.000												
Likert_E	0.372	0.267	1.000											
Likert_A	0.385	0.398	0.409	1.000										
Likert_N	−0.463	−0.381	−0.368	−0.265	1.000									
MFC_O	0.597	0.177	0.222	0.128	−0.349	1.000								
MFC_C	0.241	0.579	0.074	0.167	−0.214	0.527	1.000							
MFC_E	0.209	0.047	0.764	0.293	−0.227	0.245	0.016	1.000						
MFC_A	0.182	0.208	0.300	0.620	−0.185	0.388	0.415	0.351	1.000					
MFC_N	−0.297	−0.196	−0.277	−0.177	0.675	−0.562	−0.409	−0.319	−0.382	1.000				
task	0.401	0.515	0.365	0.403	−0.296	0.170	0.228	0.200	0.205	−0.154	1.000			
peop	0.481	0.461	0.232	0.553	−0.413	0.240	0.177	0.130	0.292	−0.270	0.579	1.000		
char	0.449	0.462	0.451	0.491	−0.358	0.243	0.199	0.323	0.308	−0.250	0.802	0.734	1.000	
ethic	0.462	0.477	0.237	0.539	−0.378	0.208	0.181	0.123	0.270	−0.227	0.688	0.947	0.752	1.000
(Female-only group)														
Likert_O	1.000													
Likert_C	0.163	1.000												
Likert_E	0.352	0.221	1.000											
Likert_A	0.323	0.330	0.337	1.000										
Likert_N	−0.351	−0.284	−0.411	−0.299	1.000									
MFC_O	0.646	0.082	0.286	0.136	−0.368	1.000								
MFC_C	0.075	0.698	0.135	0.172	−0.208	0.330	1.000							
MFC_E	0.261	0.013	0.798	0.229	−0.330	0.257	−0.057	1.000						
MFC_A	0.215	0.216	0.331	0.674	−0.279	0.373	0.398	0.304	1.000					
MFC_N	−0.299	−0.259	−0.376	−0.249	0.841	−0.465	−0.348	−0.306	−0.386	1.000				

(Continues)

TABLE 6 (Continued)

	Likert_O	Likert_C	Likert_E	Likert_A	Likert_N	MFC_O	MFC_C	MFC_E	MFC_A	MFC_N	task	peop	char	ethic
task	0.371	0.464	0.351	0.341	−0.231	0.231	0.316	0.223	0.220	−0.234	1.000			
peop	0.434	0.376	0.317	0.593	−0.380	0.253	0.258	0.199	0.414	−0.348	0.599	1.000		
char	0.431	0.433	0.478	0.486	−0.435	0.278	0.299	0.331	0.368	−0.392	0.748	0.787	1.000	
ethic	0.439	0.411	0.290	0.536	−0.337	0.255	0.294	0.163	0.360	−0.315	0.702	0.938	0.766	1.000

Note: Likert_A = factor score for Likert-type measure of agreeableness, Likert_C = factor score for Likert-type measure of conscientiousness, Likert_E = factor score for Likert-type measure of extraversion, Likert_N = factor score for Likert-type measure of neuroticism, Likert_O = factor score for Likert-type measure of openness, MFC_O = factor score for MFC measure of openness, MFC_C = factor score for MFC measure of conscientiousness, MFC_E = factor score for MFC measure of extraversion, MFC_A = factor score for MFC measure of agreeableness, MFC_N = factor score for MFC measure of neuroticism, task = factor score for task-oriented leadership, peop = factor score for people-oriented leadership, char = factor score for charismatic leadership, and ethic = factor score for ethical leadership.

TABLE 7 Tests of differential prediction of Likert-type personality across gender groups.

Predictor	Outcome	Step 1	Step 2	Step 3	Conclusion
Openness (Likert)	Task leadership	$\Delta R^2 = 0.003$, $\Delta F = 1.435$			No bias
	People leadership	$\Delta R^2 = 0.020$, $\Delta F = 11.201^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.787$	$\Delta R^2 = 0.019$, $\Delta F = 21.619^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.004$, $\Delta F = 2.073$			No bias
	Ethical leadership	$\Delta R^2 = 0.021$, $\Delta F = 11.592^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.242$	$\Delta R^2 = 0.021$, $\Delta F = 22.962^{**}$	Intercept differences only[b]
Conscientiousness (Likert)	Task leadership	$\Delta R^2 = 0.002$, $\Delta F = 1.360$			No bias
	People leadership	$\Delta R^2 = 0.013$, $\Delta F = 6.825^{**}$	$\Delta R^2 = 0.005$, $\Delta F = 5.005^{*}$	$\Delta R^2 = 0.008$, $\Delta F = 8.517^{**}$	Intercept and slope differences
	Charismatic leadership	$\Delta R^2 < 0.001$, $\Delta F = 0.088$			No bias
	Ethical leadership	$\Delta R^2 = 0.012$, $\Delta F = 6.681^{**}$	$\Delta R^2 = 0.004$, $\Delta F = 3.936^{*}$	$\Delta R^2 = 0.008$, $\Delta F = 9.308^{**}$	Intercept and slope differences
Extra-version (Likert)	Task leadership	$\Delta R^2 = 0.006$, $\Delta F = 2.979$			No bias
	People leadership	$\Delta R^2 = 0.024$, $\Delta F = 11.316^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.673$	$\Delta R^2 = 0.023$, $\Delta F = 21.967^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.008$, $\Delta F = 4.682^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.423$	$\Delta R^2 = 0.008$, $\Delta F = 8.948^{**}$	Intercept differences only[b]
	Ethical leadership	$\Delta R^2 = 0.024$, $\Delta F = 11.538^{**}$	$\Delta R^2 < 0.001$, $\Delta F = 0.115$	$\Delta R^2 = 0.024$, $\Delta F = 22.983^{**}$	Intercept differences only[b]
Agreeableness (Likert)	Task leadership	$\Delta R^2 = 0.003$, $\Delta F = 1.697$			No bias
	People leadership	$\Delta R^2 < 0.001$, $\Delta F = 0.021$			No bias
	Charismatic leadership	$\Delta R^2 = 0.004$, $\Delta F = 2.129$			No bias
	Ethical leadership	$\Delta R^2 < 0.001$, $\Delta F = 0.480$			No bias
Neuroticism (Likert)	Task leadership	$\Delta R^2 = 0.010$, $\Delta F = 4.593^{*}$	$\Delta R^2 = 0.002$, $\Delta F = 1.778$	$\Delta R^2 = 0.008$, $\Delta F = 7.402^{**}$	Intercept differences only[b]
	People leadership	$\Delta R^2 = 0.040$, $\Delta F = 21.201^{**}$	$\Delta R^2 = 0.002$, $\Delta F = 1.620$	$\Delta R^2 = 0.039$, $\Delta F = 40.752^{**}$	Intercept differences only[b]
	Charismatic leadership	$\Delta R^2 = 0.016$, $\Delta F = 8.500^{**}$	$\Delta R^2 = 0.002$, $\Delta F = 1.951$	$\Delta R^2 = 0.014$, $\Delta F = 15.031^{**}$	Intercept differences only[b]
	Ethical leadership	$\Delta R^2 = 0.040$, $\Delta F = 20.121^{**}$	$\Delta R^2 = 0.002$, $\Delta F = 1.689$	$\Delta R^2 = 0.038$, $\Delta F = 38.522^{**}$	Intercept differences only[b]

Note: Differential prediction was conducted following the steps from Lautenschlager and Mendoza (1986).

[a]Females are overpredicted.

[b]Females are underpredicted.

$^{*}p < 0.05$; $^{**}p < 0.01$.

personality measures in a selection setting; future research could use actual job applicant reactions to validate and support our findings. If future research supports our finding of little to no differential test-taker reactions between males and females among actual job applicants, it suggests that there would be fewer fairness and legal problems in

terms of gender bias when using MFC assessments. Meaning, while prior research has suggested that negative test-taker reactions can directly impact the diversity of the applicant pool, the perceptions of organizational justice, and the likelihood of accepting a job offer (Konradt et al., 2017; J. M. McCarthy et al., 2013; Rynes & Barber, 1990; Smither et al., 1993), the lack of significant findings suggests that using MFC measures would not lead to fewer females in the applicant pool or fewer females accepting a job offer.

Second, we found that slopes of the relationships between personality dimensions and self-assessed leadership, compared between males and females, were consistent (i.e., no meaningful evidence of differential prediction in terms of slope differences). Slope differences were only evident in one pairwise comparison using MFC measures: the slope of neuroticism predicting charismatic leadership was stronger among females than males. This is potentially explainable by previous research finding the largest mean differences between genders on neuroticism, such that females tended to score higher (Vianello et al., 2013). It is thus possible that the stronger endorsement of such traits also leads to stronger *expression* of such traits in leadership, thus inflating the slope and size of the relationship. For example, prior research has suggested that individuals vary in how much they express their personality while at work, depending on the work environment (Barrick et al., 2004; Colbert et al., 2004). Moreover, overall, the amount of differential prediction based on the change in R^2 was generally lower among MFC measures compared with Likert type. This again supports the viability of using MFC measures for HR development, especially in predicting self-assessed leadership.

Thus, our study has important implications in providing evidence that MFC personality measures can be used to not only obtain better fake-resistant estimates of applicant personality but also to identify individuals with perceived high leadership abilities (e.g., in identifying high-potential employees; Bialek & Hagen, 2022). This directly addresses a major gap in the literature of a lack of research on differential prediction as it pertains to personality and leadership, with even more of a gap in research on how MFC measures relate to differential prediction. With future studies to validate these findings in other contexts, the implications of gender fairness in terms of differential prediction would be vital for both research and practice.

Finally, our study contributes unique findings identifying differential prediction in terms of intercept differences on MFC personality measures, such that self-reported leadership outcomes are often underpredicted for females. This finding is of vital importance for the feasibility of using MFC measures for HR development, as systematic underprediction using a flawed predictor measure could result in decision-making that might lead to unfair outcomes. This is especially important as there is an urgent need for organizations to better identify, support, and elevate female leaders. The use of biased selection criteria would systematically underpredict (and thus, lead to under selection) of female applicants. Our evidence for intercept differences suggests that females' personality traits are related to undervaluing their self-assessed leadership ability. Thus, it is possible that an HR development program that uses a common regression line for MFC personality measures predicting leadership outcomes would result in lower leadership scores predicted for females than what their true score would be, leading to systematic under-selection of females for leadership-related roles. However, we note that the pattern of underprediction using MFC measures was very similar to the pattern of underprediction using Likert-type measures. Thus, our findings again suggest that MFC measures are no less appropriate than Likert-type measures in terms of differential prediction. Given the prevalence of underprediction, extensive future research and development are needed to identify the source of the intercept differences and remove as much bias as possible in the measures. Doing so would have important implications in better addressing the gender differences currently seen in organizational leadership selection.

An anonymous reviewer also recommended that we analyze the data using a weighted composite score of all five personality traits. We performed this for each of the four outcome variables separately. For example, we regressed task leadership onto the five MFC personality traits, extracted the standardized coefficients, and then used them to create a weighted composite of the five personality traits. We then tested for differential prediction. For all four outcome variables, the weighted composite did not show any differential prediction bias. This was interesting in that it suggested that the use of weighted composite personality scores could minimize the gender biases found when using each personality trait separately. In real selection procedures where composite scores are common, this is good news for researchers and practitioners who want to use MFC measures.

Taken together, this study contributes to the literature on test fairness in MFC personality tests. Lee et al. (2021) recently investigated issues of measurement bias in MFC personality tests using a differential item functioning (DIF) method and found only 1 out of 20 MFC item blocks showed DIF between male and female groups. However, no research has yet examined the fairness and predictive bias of MFC tests between male and female groups via a differential prediction approach. Our study addresses this research gap by providing additional insights into the test fairness issues of MFC personality tests. These findings can help researchers and practitioners better understand the validity and fairness of MFC personality tests.

6.1 | Limitations and future research

As with any study, our research has limitations that should be addressed in future research. First, the data in the test-taker reactions study were collected from undergraduate students, and the employee sample via the Mturk was collected in research settings. While the former setting approximated the use of personality tests in vocational counseling, the latter relies on self-report scores for outcome variables such as leadership. However, we noted that most prior studies used similar simulated selection scenario designs, and those studies that did tend to report stronger faking effects, which suggests that the design does approximate a selection setting (Cao & Drasgow, 2019). Additionally, we argued that self-assessed leadership is still a valuable outcome variable to study in other HR development settings such as identifying employees for professional growth and training programs. That being said, additional research on actual applicants and real selection settings would be necessary to be able to validate and generalize our findings to a selection system. Similarly, data collected in cross-sectional self-report designs like the one used in Study 2 runs the risk of common method bias (Podsakoff et al., 2003) and bias in terms of self-reported leadership compared with the others-report leadership measures that would be used in actual performance measurement (Warech et al., 1998). While we took several steps throughout our methods and analysis to mitigate these concerns, a better study design would be to collect data at two time points, with the leadership outcome data collected from others-report sources to produce more accurate 360° evaluations of leadership.

In addition, adjustments to the MFC format could potentially address other test-taker concerns related to the use of MFC measures. For example, our results showed that test-taker reactions to MFC measures were overall more negative than reactions to Likert-type measures. According to Sass et al. (2020), negative reactions could potentially be reduced through modifications to the MFC format that reduce the number of choices that need to be made. For future research, they suggest that research on test-taker reactions should focus on *motivation* as opposed to anxiety or cognitive load. Other suggestions have also been made—such as reducing negatively worded items—to improve overall test-taker reactions to MFC measures (Dalal et al., 2019); however, this would need to be balanced with the need for negatively worded items to have reliable scoring (Lee et al., 2022). Moreover, as MFC measures become more popular and widespread, it is possible that applicants will become more used to the format and thus grow to have fewer negative reactions.

In terms of differential prediction, we emphasize that our findings suggest that MFC measures underpredict leadership outcomes for females just as frequently as Likert-type measures do. However, the evidence does suggest that there is some differential validity and extensive intercept differences, contrary to some prior findings (Berry et al., 2013). Thus, in order to ensure that personality measures (regardless of test format) would not be a limiting factor (i.e., "glass ceiling") preventing females from reaching higher levels of leadership positions, further research is necessary to identify the root cause of differential prediction. Meade and Fetzer (2009) describe how differential prediction could be caused by sources unrelated to the test itself, such as stronger mean differences on the criterion side than on the predictor side, reliability, or omitted variables. For example, Keiser et al. (2016) found that the underprediction of female performance in college based on admissions tests was explained by course-taking patterns. Similarly, Schmitt et al. (2017) found that gender differences in personality could be explained by cultural values such as egalitarianism.

Future research identifying the *source* of differential prediction would be crucial to understanding the appropriateness of MFC personality measures in selection contexts. Additionally, future research should focus on the *size* of these differences. As Berry (2015) described, differences may be "sizable in percentage terms, in absolute terms these are relatively small differences," and for applied purposes, the magnitude may not be large enough to actually lead to biased hiring results (p. 459). Finally, future research could expand our findings to other job-related outcomes such as satisfaction and attrition. Given that the previous evidence for a *lack* of differential prediction was based on performance outcomes, it is possible that the true cause is on the criterion side of what is being predicted.

7 | CONCLUSION

In conclusion, our study builds on increasing interest in the use of MFC personality measures in applied settings by uniquely examining questions of fairness and bias based on gender. Given the growing interest in MFC measures for selection and applied practice, it is of utmost importance to ensure that any tests used maintain fairness and equity among protected groups, such as gender. Our research contributes to the literature on MFC personnel assessments by examining gender biases in test-taker reactions and differential prediction. We hope that our research provides a springboard and poses important questions for future research to build on and develop in pursuit of identifying a better way to measure personality for the purposes of personnel selection and applied practice.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at https://osf.io/5rdx8.

ORCID

Steven Zhou https://orcid.org/0000-0003-0710-7065
Philseok Lee https://orcid.org/0000-0002-6965-0808
Shea Fyffe https://orcid.org/0000-0003-0312-7915

ENDNOTES

[1] Because our research question is at the univariate level and due to lower levels of correlations between dependent variables (see Tables 1 and 2), recent scholars have recommended the use of univariate ANOVAs with the Benjamini-Hochberg correction instead of a MANOVA (Benjamini & Hochberg, 1995; Huang, 2020).

[2] Please refer to Dueber et al. (2019) for further technical detail regarding empirical reliability.

REFERENCES

Adam, J. J. (1999). Gender differences in choice reaction time: Evidence for differential strategies. *Ergonomics*, *42*(2), 327–335. https://doi.org/10.1080/001401399185685

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291–304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448. https://doi.org/10.1287/mnsc.2013.1776

Barrick, M. R., Mitchell, T. R., & Stewart, G. L. (2004). Situational and motivation influences on trait-behavior relationships. In M. R. Barrick & A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 60–82). John Wiley & Sons.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

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

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Berry, C. M., Kim, A., Wang, Y., Thompson, R., & Mobley, W. H. (2013). Five-factor model personality measures and sex-based differential prediction of performance. *Applied Psychology*, *62*(1), 13–43. https://doi.org/10.1111/j.1464-0597.2012.00493.x

Bialek, T. K., & Hagen, M. S. (2022). Cohort-based leadership development for high-potential employees: A model for programmatic design. *Human Resource Development Quarterly*, *33*(4), 361–382. https://doi.org/10.1002/hrdq.21459

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

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology*, *89*(5), 901–910. https://doi.org/10.1037/0021-9010.89.5.901

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, *9*(1–2), 52–69. https://doi.org/10.1111/1468-2389.00163

Boyce, A. S., Conway, J. S., & Caputo, P. M. (2014). *Development and validation of Aon Hewitt's personality model and adaptive employee personality test (ADEPT-15)*. Aon Hewitt.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, M. E., & Treviño, L. K. (2006). Ethical leadership: A review and future directions. *The Leadership Quarterly*, *17*(6), 595–616. https://doi.org/10.1016/j.leaqua.2006.10.004

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes*, *97*(2), 117–134. https://doi.org/10.1016/j.obhdp.2005.03.002

Byle, K. A., & Holtgraves, T. M. (2008). Integrity testing, personality, and design: Interpreting the personnel reaction blank. *Journal of Business and Psychology*, *22*(4), 287–295. https://doi.org/10.1007/s10869-008-9059-z

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

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Colbert, A. E., Mount, M. K., Harter, J. K., Witt, L. A., & Barrick, M. R. (2004). Interactive effects of personality and perceptions of the work situation on workplace deviance. *Journal of Applied Psychology*, *89*(4), 599–609. https://doi.org/10.1037/0021-9010.89.4.599

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. (2019). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*. Advance online publication., *36*, 55–70. https://doi.org/10.1007/s10869-019-09655-6

De Acedo Lizárraga, M. L. S., De Acedo Baquedano, M. T. S., & Cardelle-Elawar, M. (2007). Factors that affect decision making: Gender and age differences. *International Journal of Psychology and Psychological Therapy*, *7*(3), 381–391. https://www.redalyc.org/articulo.oa?id=56070306

De Hoogh, A. H., Den Hartog, D. N., Koopman, P. L., Thierry, H., Van den Berg, P. T., Van der Weide, J. G., & Wilderom, C. P. (2005). Leader motives, charismatic leadership, and subordinates' work attitude in the profit and voluntary sector. *The Leadership Quarterly*, *16*(1), 17–38. https://doi.org/10.1016/j.leaqua.2004.10.001

De Vries, R. E. (2012). Personality predictors of leadership styles and the self–other agreement problem. *The Leadership Quarterly*, *23*(5), 809–821. https://doi.org/10.1016/j.leaqua.2012.03.002

Dueber, D. M., Love, A. M., 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

Duehr, E. E. (2006). *Personality, gender, and transformational leadership: Investigating differential prediction for male and female leaders* (Publication No. 3235398). [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations Publishing.

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Taylor & Francis Group.

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

Fuller, C. M., Simmering, M. J., Atinc, G., Atinc, Y., & Babin, B. J. (2016). Common methods variance detection in business research. *Journal of Business Research*, *69*(8), 3192–3198. https://doi.org/10.1016/j.jbusres.2015.12.008

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment*, *25*(4), 513–526. https://doi.org/10.1177/1073191116641181

Harland, L. K. (2003). Using personality tests in leadership development: Test format effects and the mitigating impact of explanations and feedback. *Human Resource Development Quarterly*, *14*(3), 285–301. https://doi.org/10.1002/hrdq.1067

Harris, A. M., McMillan, J. T., & Carter, N. T. (2020). Test-taker reactions to ideal point measures of personality. *Journal of Business and Psychology*. Advance online publication., *36*, 513–532. https://doi.org/10.1007/s10869-020-09682-8

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

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*(5), 1323–1333. https://doi.org/10.1016/j.jrp.2008.04.006

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology*, *1*(3), 272–290. https://doi.org/10.1111/j.1754-9434.2008.00048.x

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*(1), 183–209. https://doi.org/10.1146/annurev-orgpsych-032414-111441

House, R. J., & Howell, J. M. (1992). Personality and charismatic leadership. *The Leadership Quarterly*, *3*(2), 81–108. https://doi.org/10.1016/1048-9843(92)90028-E

Huang, F. L. (2020). MANOVA: A procedure whose time has passed? *Gifted Child Quarterly*, *64*(1), 56–60. https://doi.org/10.1177/0016986219887200

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology*, *85*(6), 869–879. https://doi.org/10.1037/0021-9010.85.6.869

Jamali, L. (2020). *A push to get more women on corporate boards gains momentum*. NPR. https://www.npr.org/2020/03/05/811192459/a-push-to-get-more-women-on-corporate-boards-gains-momentum

Johns, M. L. (2013). Breaking the glass ceiling: Structural, cultural, and organizational barriers preventing women from achieving senior and executive positions. *Perspectives in Health Information Management*, *10*(Winter), 1–11. https://www.ncbi.nlm.nih.gov/pmc/articles/Pmc3544145

Kalaitzi, S., Czabanowska, K., Fowler-Davis, S., & Brand, H. (2017). Women leadership barriers in healthcare, academia and business. *Equality, Diversity and Inclusion: An International Journal*, *36*(5), 457–474. https://doi.org/10.1108/EDI-03-2017-0058

Kassambara, A. (2021). *rstatix: Pipe-friendly framework for basic statistical tests*. R Package Version 0.7.0. Retrieved from https://CRAN.R-project.org/package=rstatix

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology*, *101*(4), 569–581. https://doi.org/10.1037/apl0000069

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*(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

Konradt, U., Garbers, Y., Böge, M., Erdogan, B., & Bauer, T. N. (2017). Antecedents and consequences of fairness perceptions in personnel selection: A 3-year longitudinal study. *Group & Organization Management*, *42*(1), 113–146. https://doi.org/10.1177/1059601115617665

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology*, *67*(2), 264–283. https://doi.org/10.1111/apps.12108

Lautenschlager, G. J., & Mendoza, J. L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement*, *10*(2), 133–139. https://doi.org/10.1177/014662168601000202

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. https://doi.org/10.1016/j.paid.2019.01.022

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

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

Lishner, D. A., Nguyen, S., Stocks, E. L., & Zillmer, E. J. (2008). Are sexual and emotional infidelity equally upsetting to men and women? Making sense of forced-choice responses. *Evolutionary Psychology*, *6*(4), 667–675. https://doi.org/10.1177/147470490800600412

Lundgren, H., Poell, R. F., & Kroon, B. (2019). "This is not a test": How do human resource development professionals use personality tests as tools of their professional practice? *Human Resource Development Quarterly*, *30*(2), 175–196. https://doi.org/10.1002/hrdq.21338

Martinez Gómez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, *12*(732241), 1–19. https://doi.org/10.3389/fpsyg.2021.732241

McCarthy, J., Hrabluik, C., & Jelley, R. B. (2009). Progression through the ranks: Assessing employee reactions to high-stakes employment testing. *Personnel Psychology*, *62*(4), 793–832. https://doi.org/10.1111/j.1744-6570.2009.01158.x

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "so what?," "What's new?," and "where to next?". *Journal of Management*, *43*(6), 1693–1725. https://doi.org/10.1177/0149206316681846

McCarthy, J. M., Van Iddekinge, C. H., Lievens, F., Kung, M. C., Sinar, E. F., & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect criterion-related validity? A multistudy investigation of relations among reactions, selection test scores, and job performance. *Journal of Applied Psychology*, *98*(5), 701–719. https://doi.org/10.1037/a0034089

McCrae, R. R., & Costa, P. T., Jr. (1991). The NEO personality inventory: Using the five-factor model in counseling. *Journal of Counseling & Development*, *69*(4), 367–372. https://doi.org/10.1002/j.1556-6676.1991.tb01524.x

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, *12*(4), 738–761. https://doi.org/10.1177/1094428109331487

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Moyle, P., & Hackston, J. (2018). Personality assessment for employee development: Ivory tower or real world? *Journal of Personality Assessment*, *100*(5), 507–517. https://doi.org/10.1080/00223891.2018.1481078

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

Newman, D. A., Jones, K. S., Fraley, R. C., Lyon, J. S., & Mullaney, K. M. (2014). Why minority recruiting doesn't often work, and what can be done about it: Applicant qualifications and the 4-group model of targeted recruiting. In K. Y. T. Yu & D. M. Cable (Eds.), *Oxford library of psychology. The Oxford handbook of recruitment* (pp. 492–556). Oxford University Press.

Nye, C. D., White, L. A., Drasgow, F., Prasad, J., Chernyshenko, O. S., & Stark, S. (2020). Examining personality for the selection and classification of soldiers: Validity and differential validity across jobs. *Military Psychology*, *32*(1), 60–70. https://doi.org/10.1080/08995605.2019.1652482

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. https://doi.org/10.1016/j.paid.2016.03.075

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*(5), 377–396. https://doi.org/10.1002/per.588

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Rath, T. (2007). *StrengthsFinder 2.0*. Simon and Schuster.

Rynes, S. L., & Barber, A. E. (1990). Applicant attraction strategies: An organizational perspective. *Academy of Management Review*, *15*(2), 286–310. https://doi.org/10.5465/amr.1990.4308158

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology*, *87*(4), 667–674. https://doi.org/10.1037/0021-9010.87.4.667

Salgado, J. F., & De Fruyt, F. (2017). Personality in personnel selection. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 174–198). Blackwell Publishing.

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

Samejima, F. (1997). Graded response model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.

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

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology*, *52*(S1), 45–56. https://doi.org/10.1002/ijop.12265

Schriesheim, C. A., & Stogdill, R. M. (1975). Differences in factor structure across three versions of the Ohio State Leadership Scales. *Personnel Psychology*, *28*(2), 189–206. https://doi.org/10.1111/j.1744-6570.1975.tb01380.x

Sims, C., Carter, A., & Moore De Peralta, A. (2021). Do servant, transformational, transactional, and passive avoidant leadership styles influence mentoring competencies for faculty? A study of a gender equity leadership development program. *Human Resource Development Quarterly*, 32*(1)*, 55–75. https://doi.org/10.1002/hrdq.21408

Smither, J. W., Reilly, R. R., Millsap, R. E., Kenneth Pearlman AT&T, & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*(1), 49–76. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Society for Industrial and Organizational Psychology.

Speer, A. B., King, B. S., & Grossenbacher, M. (2016). Applicant reactions as a function of test length. *Journal of Personnel Psychology*, *15*(1), 15–24. https://doi.org/10.1027/1866-5888/a000145

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*(3), 153–164. https://doi.org/10.1037/mil0000044

Stelter, N. Z. (2002). Gender differences in leadership: Current social issues and future organizational implications. *Journal of Leadership Studies*, *8*(4), 88–99. https://doi.org/10.1177/107179190200800408

Stobart, G., & Eggen, T. (2012). High-stakes testing: Value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. https://doi.org/10.1080/0969594X.2012.639191

Stogdill, R. M. (1963). *Manual for the leader behavior description questionnaire, Form XII*. Bureau of Business Research, Ohio State University.

Truxillo, D. M., Bauer, T. N., & McCarthy, J. M. (2015). Applicant fairness reactions to the selection process. In R. S. Cropanzano & M. L. Ambrose (Eds.), *Oxford library of psychology. The Oxford handbook of justice in the workplace* (pp. 621–640). Oxford University Press.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic—Transformational leadership research: Back to the drawing board? *Academy of Management Annals*, *7*(1), 1–60. https://doi.org/10.5465/19416520.2013.759433

Vianello, M., Schnabel, K., Sriram, N., & Nosek, B. (2013). Gender differences in implicit and explicit personality traits. *Personality and Individual Differences*, *55*(8), 994–999. https://doi.org/10.1016/j.paid.2013.08.008

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Warech, M. A., Smither, J. W., Reilly, R. R., Millsap, R. E., & Reilly, S. P. (1998). Self-monitoring and 360-degree ratings. *The Leadership Quarterly*, *9*(4), 449–473. https://doi.org/10.1016/S1048-9843(98)90011-X

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*(178), 1–11. https://doi.org/10.3389/fpsyg.2011.00178

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., *33*, 156–170. https://doi.org/10.1037/pas0000971

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. https://doi.org/10.1027/1015-5759/a000609

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*(3), 569–590. https://doi.org/10.1177/1094428119836486

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*(2), 168–190. https://doi.org/10.1177/1094428104263674

AUTHOR BIOGRAPHIES

Steven Zhou is a PhD candidate in industrial and organizational psychology at George Mason University. His primary research areas are in leadership, psychometrics, careers and calling, and the academic-practitioner gap. His prior work experience includes HR, data analytics, and nonprofit management.

Dr. Philseok Lee is an Assistant Professor of Psychology at George Mason University. His primary research areas are in psychometrics, big data and machine learning, faking issues in personnel selection, and personality. He earned his PhD in industrial and organizational psychology from the University of South Florida.

Shea Fyffe is a PhD candidate in industrial and organizational psychology at George Mason University. His primary research areas are in psychometrics, natural language processing, personality, and statistical programming. He previously worked as a psychometric data analyst before starting the PhD program.

How to cite this article: Zhou, S., Lee, P., & Fyffe, S. (2024). Examining gender differences in the use of multidimensional forced-choice measures of personality in terms of test-taker reactions and test fairness. *Human Resource Development Quarterly*, 1–27. https://doi.org/10.1002/hrdq.21521