Personnel Assessment and Decisions

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

Article 3

2021

# Faking and the Validity of Personality Tests: An Experimental Investigation Using Modern Forced Choice Measures

Christopher R. Huber
*Human Resources Research Organization*

Nathan R. Kuncel
*University of Minnesota-Twin Cities*

Katie B. Huber
*University of Wisconsin-River Falls*

Anthony S. Boyce
*Amazon*



## Recommended Citation

Huber, Christopher R.; Kuncel, Nathan R.; Huber, Katie B.; and Boyce, Anthony S. (2021) "Faking and the Validity of Personality Tests: An Experimental Investigation Using Modern Forced Choice Measures," *Personnel Assessment and Decisions*: Number 7 : Iss. 1 , Article 3.
DOI: https://doi.org/10.25035/pad.2021.01.003
Available at: https://scholarworks.bgsu.edu/pad/vol7/iss1/3

# Faking and the Validity of Personality Tests: An Experimental Investigation Using Modern Forced Choice Measures

**Christopher R. Huber[1], Nathan R. Kuncel[2], Katie B. Huber[3], and Anthony S. Boyce[4]**

1. Human Resources Research Organization
2. University of Minnesota-Twin Cities
3. University of Wisconsin-River Falls
4. Amazon

ABSTRACT

Despite the established validity of personality measures for personnel selection, their susceptibility to faking has been a persistent concern. However, the lack of studies that combine generalizability with experimental control makes it difficult to determine the effects of applicant faking. This study addressed this deficit in two ways. First, we compared a subtle incentive to fake with the explicit "fake-good" instructions used in most faking experiments. Second, we compared standard Likert scales to multidimensional forced choice (MFC) scales designed to resist deception, including more and less fakable versions of the same MFC inventory. MFC scales substantially reduced motivated score elevation but also appeared to elicit selective faking on work-relevant dimensions. Despite reducing the effectiveness of impression management attempts, MFC scales did not retain more validity than Likert scales when participants faked. However, results suggested that faking artificially bolstered the criterion-related validity of Likert scales while diminishing their construct validity.

KEYWORDS

faking, personality testing, validity, personnel selection

Concerns about the fakability of personality measures gained traction soon after the emergence of personality testing itself and persist to this day (Meehl & Hathaway, 1946; Rosse et al., 1998; Zickar, 2000). Substantial distortion and outright lying have been documented on a variety of predictors, including interviews, biographical information, and personality questionnaires (Anderson et al., 1984; Cascio, 1975; Pannone, 1984; Weiss & Feldman, 2006). Despite such findings, meta-analytic syntheses suggest that personality traits such as conscientiousness and emotional stability retain substantial criterion-related validity in employment settings (Barrick & Mount, 1991; Barrick et al., 2001). This evidence has led some researchers to argue that the negative effects of faking are largely exaggerated (e.g., Ones et al., 1996), whereas others remain concerned. The purpose of this study is to advance our understanding of the faking–validity relationship using a novel experimental methodology.

Directed faking studies, in which participants are explicitly instructed to "fake good" by posing as ideal job candidates, demonstrate that applicants can fake effectively if they so choose. A meta-analysis by Viswesvaran and Ones (1999) found large differences between faked and honest responses on the Big Five traits, especially in studies that used within-subjects designs. In these studies, participants elevated their scores (on average) by .47 standard deviations on agreeableness, .54 on extraversion, .76 on openness, .89 on conscientiousness, and .93 on emotional stability.

Although directed faking studies show what fakers could do in theory, comparisons between applicant and non-applicant samples are commonly used to estimate the typical degree of response distortion in operational testing. A meta-analysis by Birkeland et al. (2006) found that applicants scored somewhat higher than non-applicants on extraversion (Cohen's $d$ = .13), openness (.15), and agreeableness (.19), and much higher on emotional stability (.50)

Corresponding author:
Chris Huber
Email: huber195@umn.edu

and conscientiousness (.52). The larger effect sizes for emotional stability and conscientiousness mirror the findings from directed faking and validity generalization research, suggesting applicants selectively fake on the most universally job-relevant traits. On the other hand, the lack of experimental control in applicant/non-applicant comparisons limits their ability to isolate the effects of faking. A variety of other factors—including selection, attrition, and differential motivation to take a personality test seriously—may influence group differences in personality scores (as well as validity coefficients).

As previously mentioned, validity generalization research shows that applicant faking has not destroyed the predictive potential of personality measures. On the other hand, evidence for validity retained in spite of faking does not tell us much about the amount of potential validity lost. This loss is difficult to measure directly due to the tradeoff between experimental control and generalizability to operational testing, but there is reason to suspect that there is room for improvement. For example, recent meta-analyses have found substantially higher validity coefficients when other-reports are used instead of self-reports (Connelly & Ones, 2010; Oh et al., 2011), which may be partially attributable to differences in response distortion.

Meta-analytic research has also found higher validities for a category of faking-resistant personality measure known as quasi-ipsative multidimensional forced choice (MFC) scales (Salgado et al., 2014). Whereas single stimulus (SS) measures (e.g., Likert scales) have test takers rate one personality statement at a time, MFC items present choices between two or more statements representing different personality dimensions (see Figure 1). The statements can be paired based on estimates of their social desirability, making it difficult for test takers to discern which option will produce the most desirable personality profile.

Although the findings are promising, it is unclear whether any validity advantage of MFC scales can be attributed to their faking resistance. A few experimental studies have supported this connection by comparing MFC and SS scales while simultaneously manipulating the motivation to fake (Christiansen et al., 2005; Hirsh & Peterson, 2008; Mueller-Hanson et al., 2003). However, comparisons of MFC and SS measures cannot control for differences between the two formats other than faking resistance. In addition, all but one of these studies used fake-good instructions, which may exaggerate or otherwise distort the effects of faking—and therefore the effects of reducing faking—due to the artificial extremity of directed faking. For example, Ellingson et al. (1999) found that faked personality scores showed only modest correlations with honest scores, and a correction for socially desirable responding did not significantly improve convergence. However, as the authors noted, their conclusions about social desirability corrections could reflect the artificial nature of directed faking. Because extreme faking all but eliminated true personality variance from faked scores, the inability to recover true personality variance via a correction was almost a foregone conclusion.

The tension between experimental control and generalizability to typical applicant behavior has been a persistent issue in the faking literature, limiting our ability to draw nuanced conclusions about the effects of applicant faking. The present study was designed to address the limitations of previous research in order to provide a better understanding of the faking–validity relationship. Specifically, we employed more nuanced manipulations of motivation and ability to fake to elicit a gradient of faking behavior, allowing for a more comprehensive analysis of the effects of faking. To better approximate typical faking behavior, we manipulated faking motivation using a subtle incentive to fake. We also tested the effects of explicit fake-good instructions, allowing us to directly compare two methods to induce faking in experimental research. This produced three levels of the faking motivation variable: honest instructions, fake-good instructions, and fake-good incentive.

In addition to comparing MFC and SS scales, we manipulated the fakability of the same MFC measure to eliminate confounding differences between the two measurement formats. This was accomplished using a computer adaptive test (CAT) that allowed for varying restrictions on the social desirability matching (SDM) of statements that were paired to form a single item. Imposing stricter matching rules on the CAT algorithm has been shown to reduce fakability by increasing the perceived similarity of paired statements (Boyce & Capman, 2017).

The faking motivation and ability manipulations produced a 3x3 design that allowed us to test several methodological and theoretical hypotheses. In keeping with past research (Boyce & Capman, 2017; Drasgow et al., 2012), we hypothesized that:

> **Hypothesis 1:** MFC scales will show smaller mean differences between honest and faked responses than SS measures of the same dimensions.

> **Hypothesis 2:** Using a stricter SDM rule will reduce mean differences between honest and faked responses.

Our next set of hypotheses concerned the relationship between faking and validity. Assuming faking reduces validity, factors that mitigate faking are likely to improve validity when there is motivation to fake. Therefore, we predicted that:

> **Hypothesis 3:** MFC scales will produce higher criterion-related validity than SS measures of the same dimensions but only when respondents are instructed to fake.

**Hypothesis 4:** Using a stricter SDM rule will produce higher criterion-related validity but only when respondents are instructed to fake.

Finally, our research design allowed for a novel methodological comparison between directed and incentivized faking. Incentivized faking studies still show faking effects, but the effect sizes are more likely to resemble those found in applicant samples (e.g., Mueller-Hanson et al., 2003). Validity may be reduced but not obliterated, and mean scores may be moderately rather than severely inflated. Therefore, we proposed that:

**Hypothesis 5:** Directed faking results will replicate using an incentivized faking manipulation.

## METHOD

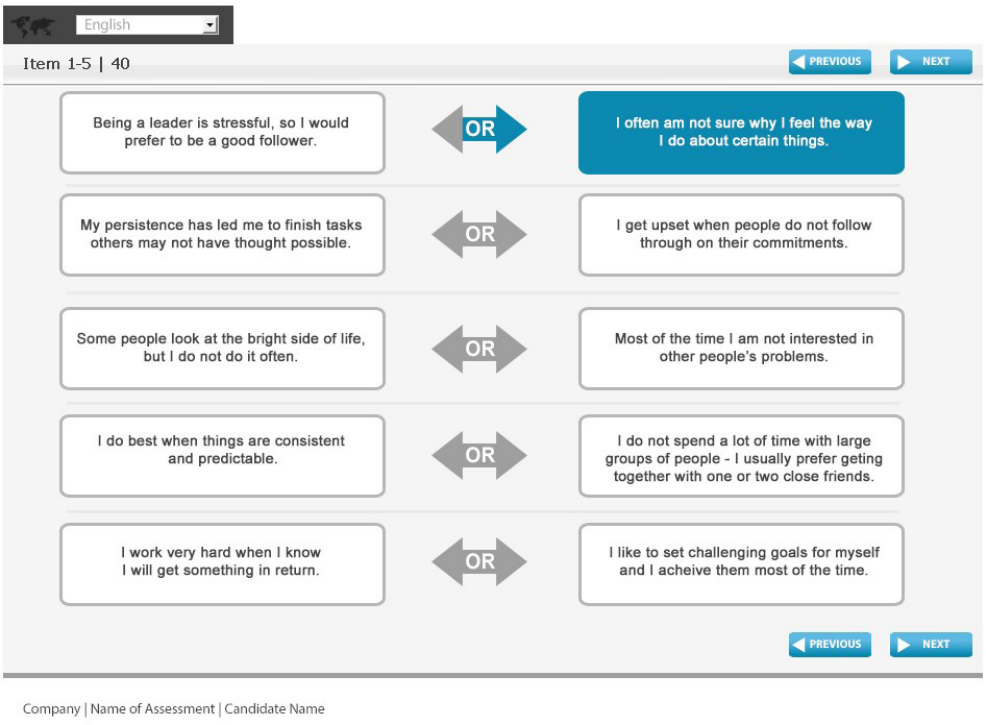
### Participants

Participants were recruited through Amazon Mechanical Turk (MTurk). Research has found that personality data from MTurk workers has comparable or superior reliability to traditional samples (Buhrmester et al., 2011). MTurk workers also appear to behave similarly to participants in traditional laboratory and field experiments (Casler et al., 2013; Horton et al., 2011).

In order to ensure the internal and external validity of results, participants were screened using a few criteria. First, we limited our participant pool to American MTurk workers over the age of 18. Second, we required participants to have at least 100 approved tasks on MTurk and an approval rate of 90% or higher. Third, participants had to be employed for at least 3 months within the past year in a position where they interacted with coworkers at least 1–2 days per week. This requirement was intended to ensure participants could complete our self-reported job performance measures (discussed below). All participants were paid $3 for their voluntary participation, and 10 were randomly selected to receive $10 bonuses.

Participants were included in the final sample if they passed two embedded attention checks, a manipulation check to ensure they had attended to their faking instructions, and a repetitive responding check. Of the 855 participants who completed the study, 652 passed these checks. The final sample was predominantly White (73%), female (57%), and currently employed (96%); see Table 1 for a breakdown of participants' occupations and educational status. Participants ranged from 19 to 70 years of age with a median age of 33. Participants were randomly assigned to one of six conditions that crossed two three-level independent variables—measurement format and faking instructions. See Table 2 for sample sizes by condition.

FIGURE 1.
Screenshot of Example Items From the MFC Inventory Used in This Study




*Note.* Copied with permission from Conway et al. (2015).

TABLE 1.
Occupational and Educational Breakdown of the Final Analysis Sample

| Occupation/highest degree | *N* |
|---|---|
| Agricultural, forestry, fishing, and related | 13 |
| Clerical and administrative support | 99 |
| Production, construction, operating maintenance, and material handling | 39 |
| Professional, paraprofessional and technical | 275 |
| Sales and related | 98 |
| Service | 96 |
| Not currently employed/full-time student | 27 |
| High school | 52 |
| Vocational/technical | 13 |
| Some college/university | 151 |
| Associate's degree | 74 |
| Bachelor's degree | 254 |
| Master's degree | 91 |
| Doctorate degree | 17 |

TABLE 2.
Summary of the Experimental Conditions

| Condition[a] | Measurement format | Faking instructions | $N_{all}$ | $N_{filtered}$ |
|---|---|---|---|---|
| 1 | SS | Honest and fake good | 141 | 99 |
| 2 | MFC–relaxed | Honest and fake good | 135 | 104 |
| 3 | MFC–strict | Honest and fake good | 126 | 87 |
| 4 | SS | Incentivized fake good | 138 | 110 |
| 5 | MFC–relaxed | Incentivized fake good | 167 | 132 |
| 6 | MFC–strict | Incentivized fake good | 148 | 120 |

*Note.* $N_{all}$ = sample size before applying the attention check filter; $N_{filtered}$ = final sample size after checks for low-effort responding; SS = single stimulus; MFC–relaxed = multidimensional forced choice with relaxed social desirability matching constraint; MFC–strict = multidimensional forced choice with strict social desirability matching constraint.
[a] Honest and faked responses from Conditions 1–3 are treated as separate conditions for data analysis purposes. Thus, the results refer to a total of nine conditions.

## Materials

***MFC Personality Inventory.*** Participants assigned to MFC conditions completed a proprietary multistage CAT developed for personnel selection and development, which measures 15 personality dimensions related to workplace outcomes (Boyce & Capman, 2017). Ten of these dimensions are based on DeYoung et al.'s (2007) Big Five Aspect Scales (BFAS), which measure two distinct aspects of each Big Five trait. DeYoung et al. (2007) validated the aspect structure through factor analysis and demonstrated convergent validity with established Big Five inventories. In addition, the individual aspects within each Big Five trait have demonstrated divergent validity with one another, including distinct relationships with other personality traits, mental abilities, neurobiological substrates, and job-related outcomes (Allen et al., 2017; DeYoung et al., 2016; DeYoung et al., 2007; DeYoung et al., 2009; Kaufman et al., 2016; Quilty et al., 2014). The remaining five dimensions of the MFC inventory capture work-relevant traits beyond the five-factor model. See Table 3 for the dimensions and their theoretical mappings.

The MFC inventory is scored using Stark's multi-unidimensional pairwise preference model, an item response theory (IRT) model for scoring binary MFC items (Stark, 2002; Stark et al., 2005). In this study, each MFC administration included 100 items. Each item consists of two personality statements selected by the CAT algorithm, resulting in approximately 13 statements per personality dimension.

In addition to IRT parameters, each statement has an associated social desirability parameter ranging from 0 to 1 (established based on a directed faking study). In the strict SDM conditions, the CAT algorithm was only allowed to

pair statements whose social desirability parameters were within .10 of one another. In the relaxed SDM conditions, the social desirability parameters of paired statements could differ by up to .20.

***SS Personality Scales.*** Participants in the SS conditions completed Likert-type measures of the 15 constructs assessed by the MFC inventory. To minimize differences with the MFC dimensions, we constructed the SS scales using items from the MFC CAT's statement pool. First, we used existing calibration data from a sample of MTurk workers ($N$ = 6,333), as well as previously estimated item location parameters, to select 12 items per dimension for pilot testing. Next, we administered the chosen items using a four-point response format, followed by the MFC inventory, to a pilot sample of 269 MTurk workers. Finally, we used the pilot data to construct reliable six-item scales that had good convergent validity with their MFC counterparts.

All 15 scales showed acceptable reliability, with coefficient alpha reliability estimates ranging from .74 to .90. In addition, the scales demonstrated convergent and discriminant validity with their MFC counterparts. Monotrait–heteromethod correlations ranged from .42 to .74 with a mean of .58, whereas the average heterotrait–heteromethod correlation was only .18. See Table S1 in the supplemental materials for reliability and convergent validity results by dimension.

***Self-Reported Job Performance.*** Participants completed Spector and Fox's 20-item organizational citizenship behavior checklist (OCB-C; Fox et al., 2012) and 10-item counterproductive work behavior checklist (CWB-C; Spector et al., 2010) as criterion measures. Fox et al. (2012) reported coefficient alphas of .89 and .94 for the OCB-C in two samples; Spector et al. (2010) reported an alpha of .79 for the CWB-C.

***Self-Reported Academic Performance.*** Participants completed three criterion items assessing academic performance and achievement. First, they reported their highest academic degree completed, which ranged from high school to doctoral degrees. Second, participants reported their GPA at that degree level on an 11-point scale ranging from A+ to E or F (Freeberg et al., 1989). Finally, they reported their high school GPA using the same scale.

A meta-analysis by Kuncel et al. (2005) found an average correlation of .84 between self-reported and school-reported GPA. However, self-reported GPAs were also higher than actual GPAs on average, and individuals with lower GPAs provided far less valid self-reports. Thus, it appears that self-reported GPA is a valid indicator of academic performance but is also susceptible to nontrivial response distortion.

***Emotion Management Task.*** To address the possibility of common method bias arising from self-report criteria, we included an objective performance task as an additional criterion measure. Specifically, we administered the 18-item Situational Test of Emotional Management–Brief (STEM-B; Allen et al., 2015), a performance-based emotional intelligence scale that requires examinees to identify the most effective response to a variety of emotional situations.

## TABLE 3.

Theoretical Mappings of Personality Dimensions From the MFC Inventory

| Five-factor model | MFC dimension | Theoretical mapping |
|---|---|---|
| Openness to Experience | Conceptual | BFAS-Intellect |
| | Flexibility | BFAS-Openness[a] |
| Conscientiousness | Structure | BFAS-Orderliness |
| | Drive | BFAS-Industriousness |
| Extraversion | Assertiveness | BFAS-Assertiveness |
| | Liveliness | BFAS-Enthusiasm |
| Agreeableness | Sensitivity | BFAS-Compassion |
| | Cooperativeness | BFAS-Politeness |
| Emotional Stability | Composure | BFAS-Volatility[b] |
| | Positivity | BFAS-Withdrawal[b] |
| N/A | Ambition | Need for Achievement |
| | Power | Need for Power |
| | Humility | HEXACO-Humility |
| | Mastery | Learning Goal Orientation |
| | Awareness | Social Effectiveness / Emotional Intelligence |

*Note.* MFC = multidimensional forced choice; BFAS = Big Five Aspect Scales (DeYoung et al., 2007). [a] The MFC Flexibility dimension is narrower in scope than BFAS–Openness, in that it focuses largely on openness to change and excludes aesthetic interests. [b] This BFAS scale reflects high neuroticism; the corresponding MFC dimension is scored to reflect low neuroticism (i.e., emotional stability).

***Risky Choice Framing.*** We administered Tversky and Kahneman's (1981) Asian disease problem as a final criterion measure. This problem requires participants to choose between two programs to combat a disease that threatens to kill 600 people. Preferences for safer or riskier options have been shown to vary depending on whether the potential outcomes are presented in positive or negative terms (i.e., lives saved vs. lives lost).

Risky choice problems can be scored to assess two distinct constructs. First, susceptibility to framing is quantified as a difference score between the negative and positive item scores. Second, general risk-taking tendency is assessed by combining the two scores.

***Faking Instructions.*** Participants received one of three instruction sets before completing a personality inventory. The honest instructions, which we borrowed from Mueller-Hanson et al. (2003), asked participants to respond as honestly as possible and emphasized their anonymity. The fake-good instructions asked participants to pretend they were applying for a job and make the best impression possible, responding as an ideal employee would. The incentivized fake-good instructions, adapted from Mueller-Hanson et al. (2003), explained that participants would automatically have a chance to receive one of ten $10 bonuses if they qualified for a fictitious "second part" of the study, which required participants with personality traits that were desired by employers. However, the instructions also warned that providing false responses could disqualify them from the study.

**Procedure**

As shown in Table 2, participants in Conditions 1–3 completed the same personality inventory (SS, MFC–relaxed, or MFC–strict) under both honest and fake-good instructions with the order of instructions counterbalanced. Thus, these conditions represented six levels of the 3x3 manipulation. Conversely, participants in Conditions 4–6 only completed a personality inventory once with incentivized fake-good instructions, and their results were compared to honest results from Conditions 1–3. The purpose of this between-person comparison was to avoid anchoring effects. Unlike directed fakers, incentivized fakers were instructed to provide honest responses. Asking them to respond honestly once and then immediately asking them to respond honestly a second time with an incentive to distort (or vice versa) would likely elicit suspicion and reluctance to deviate from their initial responses.

All participants began by reading the consent form and indicating their informed consent. They then completed screening and optional demographic questions, followed by the criterion measures. Finally, they completed one of three personality inventories under their assigned faking instructions. The purpose of administering the criterion measures before the predictors was to ensure that criterion responses were not contaminated by subsequent faking instructions.

## RESULTS

**Motivated Score Elevation**

Our first two hypotheses predicted that the degree of score elevation due to directed faking would be inversely related to the faking resistance of the measurement format. To test these hypotheses, we first transformed all personality scores to z-scores (using honest means and *SD*s) to create a common metric across measurement formats. Next, we conducted a mixed-model MANOVA to assess the combined effects of faking instructions (honest and fake-good) and measurement format (SS, MFC–relaxed, and MFC–strict) across all 15 personality traits. The main effect of instructions was significant, $F(1, 273) = 16.92$, $p < .001$, indicating participants generally increased their scores when directed to fake. Furthermore, we found a significant interaction between instructions and measurement format, $F(2, 548) = 2.73$, $p < .001$, suggesting the degree of score elevation varied by format.

To test Hypothesis 1, we conducted follow-up 2x2 MANOVAs comparing the SS format to each MFC format. These revealed significant instruction–format interactions for both the SS/MFC–relaxed comparison, $F(1, 187) = 4.05$, $p < .001$, and the SS/MFC–strict comparison, $F(1, 170) = 4.37$, $p < .001$. We also computed standardized mean differences (Glass's Δ) between honest and faked personality scores for all three measurement formats (see Tables S2, S3, and S4 in the supplemental materials for associated means and standard deviations). As shown in Table 4, directed faking produced large gains on the SS personality scales (mean Δ = .81). In support of Hypothesis 1, the degree of faking was much smaller on both MFC formats compared to the SS format, with a mean Δ of .28 for the MFC–relaxed inventory and .27 for the MFC–strict inventory.

Hypothesis 2 predicted that using a stricter SDM rule would also reduce faking gains. However, the difference between the two MFC formats was minimal, and the format–instructions interaction was nonsignificant in a follow-up 2x2 MANOVA. Thus, Hypothesis 2 was not supported.

On the other hand, comparing mean effect sizes across dimensions may not fully capture the behavior of directed fakers. At the item level, SS scales allow respondents to fake on one dimension without affecting their scores on other dimensions. By contrast, each MFC item requires examinees to choose between two personality dimensions. As a result, fakers may focus their self-presentation on the dimensions they perceive to be more work relevant (e.g., drive) at the expense of others. A stricter SDM rule could have a similar effect by reducing the salience of an alternate cue—that is, social desirability—for determining the "ideal" response.

To investigate this possibility, we calculated the standard deviation of Δ values across dimensions for each measurement format (see Table 4); a higher standard deviation indicates greater variation in faking across dimensions.

Both MFC formats—especially the MFC–strict format—had higher standard deviations than the SS format. This suggests that the MFC format, and perhaps stricter SDM, promoted a selective faking strategy.

Hypothesis 5 predicted that directed faking results would replicate using an incentivized faking manipulation. As shown in Table 4, incentivized faking produced small changes on the SS scales (mean Δ = .13) and even smaller changes on the MFC–relaxed (.08) and MFC–strict (.04) scales. A two-way MANOVA revealed a significant main effect of faking instructions, $F(1, 632) = 2.17$, $p = .006$. However, neither the main effect of measurement format nor the format–instructions interaction reached statistical significance. As such, the score elevation results did not support Hypothesis 5. More broadly, the average faking effect sizes suggested that the monetary incentive was only modestly successful at inducing faking. Without a strong incentive to fake in the first place, the relative advantage of the faking-resistant MFC format was greatly diminished.

### Criterion-Related Validity

Due to the combination of predictors, criteria, and experimental conditions, it was necessary to summarize a total of 1,080 validity coefficients to test Hypotheses 3 and 4. One option would be to simply calculate a mean validity coefficient for each experimental condition. However, this incorrectly assumes that the true correlations between all predictors and criteria are positive. In fact, a negative predictor–criterion correlation can be equally useful for selection if it represents the true direction of the relationship. Therefore, we developed a universal set of keys to indicate the appropriate signs for all 120 predictor–criterion relationships.

To do so, we first calculated an unweighted mean of validity coefficients for every predictor–criterion pair across all conditions. To minimize the effects of sampling error on keying decisions, we discarded any pair whose mean validity coefficient was less than .10 in absolute value. For each of the remaining 23 predictor–criterion pairs, we counted the sign of the grand mean validity coefficient as the true direction of the relationship and penalized conditions that produced a relationship in the opposite direction. Validity coefficients for these 23 pairs are summarized in Table S5, and validity coefficients for all 120 predictor–criterion pairs are available in Tables S6-S14.

Mean validity coefficients by condition are presented in Table 5. Under honest instructions, all three measurement formats had a mean validity of .15. Thus, as predicted in Hypotheses 3 and 4, no format was more valid than the others in the absence of faking. Contrary to our expectations, however, the SS scales had the highest overall validity under fake-good instructions (although *z*-tests contrasting the overall SS and MFC–strict/MFC–relaxed validity coefficients did not reach significance). This pattern held for every breakout category of criterion, including academic performance/achievement, job performance, and the STEM-B. Thus, the results failed to support Hypothesis 3, which predicted that MFC scales would perform better than their SS counterparts when participants were directed to fake. An alternate version of Hypothesis 3 might predict that the relative advantage of SS scales would diminish when participants faked, thereby accounting for the possibility that the SS scales could be more valid to begin with but lose some of that advantage due to faking. However, even this qualified Hypothesis 3 was not supported.

Hypothesis 4 predicted that MFC–strict scales would be more valid than MFC–relaxed scales but only under faking instructions. On average, MFC–strict validity coefficients were .05 higher than MFC–relaxed ones when participants faked, but the difference was not statistically significant. Therefore, Hypothesis 4 was not supported.

Once again, Hypothesis 5 predicted that directed faking results would replicate using an incentivized faking manipulation. Because the directed faking manipulation did not produce the expected outcomes (or other significant results to replicate), we did not formally evaluate Hypothesis 5 with respect to the validity results. Regardless, it is worth noting that the SS scales produced the highest validity coefficients among incentivized fakers, although the SS–MFC differences in the incentivized group did not reach statistical significance.

## DISCUSSION

### Motivated Score Elevation

Our directed faking results showed substantial differences between measurement formats in both the magnitude and pattern of faking. As expected, fakers were far less successful at raising their scores on the MFC scales. In addition, it appears that fakers selectively distorted on specific traits to a greater extent when responding to an MFC inventory. A closer examination of the distortion patterns suggests they favored traits with higher face validity for employee selection, including drive, cooperativeness, composure, ambition, and mastery.

Furthermore, except for openness, selective faking produced notable discrepancies between aspects of the same Big Five traits. Although the SS scales showed strong distortion on both aspects of conscientiousness, MFC fakers focused primarily on drive and had only modest score elevation on structure. Extraversion showed a similar pattern, with fakers elevating their scores by nearly half a standard deviation on MFC–Liveliness but barely at all on MFC–Assertiveness. Fakers consistently elevated their scores on both aspects of emotional stability. However, whereas faking produced almost identical (very large) increases on both SS scales, participants faked more on composure than positivity in the MFC conditions. The difference between aspects was the most pronounced for agreeableness: Participants raised their MFC–Cooperativeness scores by an average of .67 standard deviations, whereas faked MFC–Sensitivity scores were .20 standard deviations *lower* than

## TABLE 4.

Standardized Mean Differences Between Faked/Incentivized and Honest Predictor Scores

| Big Five trait | Dimension | Directed Faking vs. honest | | | Incentivized vs. honest | | |
|---|---|---|---|---|---|---|---|
| | | SS | MFC–relaxed | MFC–strict | SS | MFC–relaxed | MFC–strict |
| Openness to experience | Conceptual | 0.61*** | 0.19 | 0.12 | 0.16 | -0.05 | 0.11 |
| | Flexibility | 0.80*** | 0.18 | 0.20 | 0.15 | 0.15 | 0.17 |
| Conscientiousness | Drive | 0.98*** | 0.61*** | 0.56*** | 0.25 | 0.19 | 0.06 |
| | Structure | 0.71*** | 0.20* | 0.20 | -0.13 | 0.04 | -0.17 |
| Extraversion | Assertiveness | 0.74*** | 0.05 | -0.02 | -0.02 | -0.03 | -0.19 |
| | Liveliness | 1.02*** | 0.49*** | 0.46*** | 0.12 | 0.22 | 0.03 |
| Agreeableness | Cooperativeness | 0.97*** | 0.59*** | 0.76*** | 0.28* | 0.22 | 0.27 |
| | Sensitivity | 0.52*** | -0.19 | -0.21 | 0.34** | -0.06 | -0.07 |
| Emotional stability | Composure | 0.99*** | 0.48*** | 0.42** | 0.21 | 0.14 | 0.06 |
| | Positivity | 1.00*** | 0.23* | 0.29** | 0.23 | -0.04 | 0.05 |
| N/A | Ambition | 1.00*** | 0.60*** | 0.69*** | 0.25 | 0.12 | 0.04 |
| | Awareness | 0.64*** | -0.04 | -0.22 | -0.07 | 0.09 | -0.12 |
| | Humility | 0.37*** | -0.03 | 0.04 | 0.03 | -0.24 | 0.10 |
| | Mastery | 0.98*** | 0.62*** | 0.48*** | 0.18 | 0.31* | 0.24 |
| | Power | 0.82*** | 0.28** | 0.22 | 0.00 | 0.11 | 0.05 |
| | *Mean* | 0.81 | 0.28 | 0.27 | 0.13 | 0.08 | 0.04 |
| | *SD* | 0.21 | 0.27 | 0.30 | 0.13 | 0.09 | 0.04 |

*Note.* SS = single stimulus; MFC–relaxed = multidimensional forced choice with relaxed social desirability matching constraint; MFC–strict = multidimensional forced choice with strict social desirability matching constraint. Directed faking comparisons are within person; incentivized faking comparisons are between person. *$p < .05$, **$p < .01$, ***$p < .001$.

## TABLE 5.

Mean Validity by Condition for Empirically Keyed Predictor–Criterion Relationships

| Instructions | Predictor format | General factor variance (%)[a] | Criterion type | | | | |
|---|---|---|---|---|---|---|---|
| | | | All | All (semipartial)[b] | Education[c] | Job Performance[d] | STEM-B |
| Honest | SS | 36 | .15 | .06 | .14 | .15 | .16 |
| | MFC–relaxed | 16 | .15 | .08 | .11 | .18 | .07 |
| | MFC–strict | 16 | .15 | .10 | .16 | .15 | .15 |
| Incentive | SS | 48 | .19 | .09 | .05 | .27 | .02 |
| | MFC–relaxed | 20 | .12 | .08 | .18 | .10 | .09 |
| | MFC–strict | 12 | .11 | .08 | .10 | .12 | .06 |
| Fake | SS | 61 | .15 | .01 | .15 | .14 | .27 |
| | MFC–relaxed | 19 | .09 | .05 | .13 | .07 | .10 |
| | MFC–strict | 22 | .14 | .07 | .13 | .13 | .20 |
| Predictor–criterion pairs | | – | 23 | 23 | 6 | 15 | 2 |

*Note.* SS = single stimulus; MFC–relaxed = multidimensional forced choice with relaxed social desirability matching constraint; MFC–strict = multidimensional forced choice with strict social desirability matching constraint. STEM-B = Situational Test of Emotional Management–Brief. [a] Percentage of variance in the predictor scales attributable to a general factor. [b] Semipartial correlations controlling for general factor variance in the predictor scores. [c] Educational criteria include highest degree achieved, GPA at highest degree level, and high school GPA. [d] Job performance criteria include self-reported counterproductive work behaviors and organizational citizenship behaviors.

honest scores, on average. The subtle faking incentive generally elicited the same patterns of differential faking across aspects, albeit far less dramatically than directed faking.

Taken together, these findings suggest that faking effects may be obscured by examining the Big Five at the domain level, especially if there are tradeoffs between faking on different dimensions. Future faking research may benefit from measuring the Big Five at the aspect or facet level and

focusing on specific traits that are attractive to fakers. The latter suggestion may be especially helpful for producing clearer results with incentivized faking designs, given the modest strength of these manipulations compared to directed faking.

Our results also suggest that practitioners should consider potential tradeoffs between face validity and reducing impression management when designing selection systems. When response distortion is a concern, there may be substantial benefits to selecting on predictively valid traits that are less attractive to fakers. If an MFC inventory is used for selection, the inclusion of unscored "distractor" scales may reduce impression management on target dimensions while also increasing the assessment's face validity.

**Criterion-Related Validity**

Our validation results failed to replicate Salgado et al.'s (2014) meta-analytic findings, which suggested quasi-ipsative MFC scales should outperform their SS counterparts. As such, it is possible that quasi-ipsative MFC scales do not provide a robust validity advantage. In keeping with this possibility, Lee et al. (2018) compared three sets of personality scores obtained from an MFC measure (using one quasi-ipsative and two ipsative scoring methods) to scores from a Likert-type version of the measure. Although all four methods showed a similar pattern of correlations with criterion measures, the Likert-type measure generally produced larger validity coefficients. Although the reason for this difference was unclear, the authors speculated that it could be due to common method bias because the criterion measures were also Likert scales.

Regardless, it is interesting that even the presence of extreme response distortion did not cause a large decrement in the validity of SS scales or improve the relative advantage of faking-resistant alternatives. Furthermore, we observed a similar trend across criteria that varied in terms of potential common method variance with SS scales. On one end of this spectrum, our self-reported job performance measures shared a Likert-type response format with the SS scales, giving the SS scales a potential edge in predicting these criteria. Our measures of GPA and degree attainment requested objective information rather than self-assessments and did not use a Likert-type response scale, but they were still likely prone to some degree of socially desirable distortion (Kuncel et al., 2005). Finally, the STEM-B required participants to correctly identify the most effective responses to specific emotional situations, making it resistant to impression management (i.e., a test taker cannot "fake" knowing the correct response).

One reasonable explanation for our validity results is that faking fundamentally changed what the SS scales measured, adding a new source of variance that contributed to the prediction of various external criteria. Past factor analytic research has found evidence of a general "ideal employee" factor in applicant samples (e.g., Schmit & Ryan, 1993), which may capture predictively useful implicit theories about how to be a good employee. Although the present study was not designed to address this question, we did conduct supplemental analyses to explore the possibility. First, we computed an average correlation of only .32 between participants' honest and faked scores on the same SS scales, suggesting the faked scores no longer assessed the intended constructs. Next, we used confirmatory factor analysis to determine if faking introduced a general method factor. As shown in Table 5, faking strengthened an already substantial general factor in the SS (but not the MFC) scales.

To determine whether this general factor impacted validity, we calculated new validity coefficients with the general factor partialled out from the predictor scores (see Table 5). Removing general factor variance substantially reduced average validity coefficients for all conditions. This suggests that shared variance between personality dimensions, whether real or artifactual, did contribute to the predictive validity of the dimension scores. The SS scales showed the most precipitous decline in validity—especially in the directed faking condition, where the average validity coefficient dropped from .15 to .01. This indicates that (a) directed faking decimated the validity of the individual SS dimensions and (b) the SS scales retained their validity in the presence of faking by measuring a new construct. In other words, faking eroded the SS scales' construct validity while simultaneously preserving their criterion-related validity. This is problematic to the extent that employers are interested in selecting for specific personality traits, as opposed to simply achieving predictive validity. On the other hand, it is unclear to what extent this phenomenon occurs given typical levels of distortion in preemployment testing.

**Future Directions**

A key feature of this study was that it manipulated both motivation and ability to fake in multiple ways. However, the observed patterns of faking suggested that the faking incentive and SDM manipulations were fairly weak, making it difficult to fully parse their effects. This limited our ability to make nuanced inferences about the effects of typical applicant faking or the merits of directed faking manipulations. Future research could remedy this issue with stronger incentives to fake and larger discrepancies between strict and relaxed SDM rules.

To the extent that quasi-ipsative MFC scales are generally better predictors of performance, it remains unclear why this is the case. The magnitude and causes of their predictive advantage remain important questions for the future of personality testing. Further experimental research using finely tuned faking manipulations, coupled with an increased focus on underlying constructs, should provide valuable insights and could substantially improve the accuracy of high-stakes personality assessment.

## REFERENCES

Allen, V., Rahman, N., Weissman, A., MacCann, C., Lewis, C., & Roberts, R. D. (2015). The Situational Test of Emotional Management–Brief (STEM-B): Development and validation using item response theory and latent class analysis. Personality and Individual Differences, 81, 195-200. https://dx.doi.org/10.1016/j.paid.2015.01.053

Allen, T. A., Rueter, A. R., Abram, S. V., Brown, J. S., & DeYoung, C. G. (2017). Personality and neural correlates of mentalizing ability. European Journal of Personality, 31, 599-613. https://doi.org/10.1002/per.2133

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. Journal of Applied Psychology, 69(4), 574-580. https://dx.doi.org/10.1037/0021-9010.69.4.574

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

Barrick, M. R., Mount, M., & Judge, T. (2001). The FFM personality dimensions and job performance: Meta-analysis of meta-analyses. International Journal of Selection and Assessment, 9, 9-30.

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://dx.doi.org/10.1111/j.1468-2389.2006.00354.x

Boyce, A. S., & Capman, J. F. (2017). ADEPT-15® technical documentation (2nd Ed.). Aon Hewitt Consulting.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6, 3-5. https://dx.doi.org/10.1177/1745691610393980

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. Journal of Applied Psychology, 60, 767-769. https://dx.doi.org/10.1037/0021-9010.60.6.767

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Computers in Human Behavior, 29, 2156-2160. https://dx.doi.org/10.1016/j.chb.2013.05.009

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. Human Performance, 18, 267-307. https://dx.doi.org/10.1207/s15327043hup1803_4

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136(6), 1092-1122. https://dx.doi.org/10.1037/a0021212

Conway, J., Boyce, A., Caputo, P. & Huber, C. (2015, April). Development of a computer adaptive forced-choice personality test. Paper presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

DeYoung, C. G., Carey, B. E., Krueger, R. F., & Ross, S. R. (2016). Ten aspects of the Big Five in the Personality Inventory for DSM–5. Personality Disorders: Theory, Research, and Treatment, 7(2), 113-123. https://doi.org/10.1037/per0000170

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology, 93(5), 880-896. https://dx.doi.org/10.1037/0022-3514.93.5.880

DeYoung, C. G., Shamosh, N. A., Green, A. E., Braver, T. S., & Gray, J. R. (2009). Intellect as distinct from openness: Differences revealed by fMRI of working memory. Journal of Personality and Social Psychology, 97(5), 883-892. https://doi.org/10.1037/a0016615

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support army personnel selection and classification decisions. U.S. Army Research Institute for the Behavioral and Social Sciences.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. Journal of Applied Psychology, 84(2), 155-166. https://doi.org/10.1037/0021-9010.84.2.155

Freeberg, N. E., Rock, D. A., & Pollack, J. (1989). Analysis of the Revised Student Descriptive Questionnaire: Phase II. Predictive validity of academic self-report. College Board Report No. 89-8. College Entrance Examination Board.

Fox, S., Spector, P. E., Goh, A., Bruursema, K., & Kessler, S. R. (2012). The deviant citizen: Measuring potential positive relations between counterproductive work behaviour and organizational citizenship behaviour. Journal of Occupational and Organizational Psychology, 85(1), 199-220. https://dx.doi.org/10.1111/j.2044-8325.2011.02032.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://dx.doi.org/10.1016/j.jrp.2008.04.006

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. Experimental Economics, 14, 399-425. https://dx.doi.org/10.1007/s10683-011-9273-9

Kaufman, S. B., Quilty, L. C., Grazioplene, R. G., Hirsh, J. B., Gray, J. R., Peterson, J. B., & DeYoung, C. G. (2016). Openness to experience and intellect differentially predict creative achievement in the arts and sciences. Journal of Personality, 84(2), 247-258. https://doi.org/ 10.1111/jopy.12156

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. Review of Educational Research, 75(1), 63-82. https://doi.org/10.3102/00346543075001063

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

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. Journal of Applied Psychology, 30(5), 525-564. https://dx.doi.org/10.1037/h0053634

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

of Applied Psychology, 88(2), 348-355. https://dx.doi.org/10.1037/0021-9010.88.2.348

Oh, I., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. Journal of Applied Psychology, 96(4), 762-773. https://dx.doi.org/10.1037/a0021832

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81(6), 660-679. https://dx.doi.org/10.1037/0021-9010.81.6.660

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. Personnel Psychology, 37(3), 507-514. https://dx.doi.org/10.1111/j.1744-6570.1984.tb00526.x

Quilty, L. C., DeYoung, C. G., Oakman, J. M., & Bagby, R. M. (2014). Extraversion and behavioral activation: Integrating the components of approach. Journal of Personality Assessment, 96(1), 87-94. https://doi.org/10.1080/00223891.2013.834440

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83(4), 634-644. https://dx.doi.org/10.1037/0021-9010.83.4.634

Salgado, J. F., Anderson, N., & Tauriz, G. (2014). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. Journal of Occupational and Organizational Psychology, 4, 797-834. https://dx.doi.org/10.1111/joop.12098

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. Journal of Applied Psychology, 78(6), 966-974. https://dx.doi.org/10.1037/0021-9010.78.6.966

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? Journal of Applied Psychology, 95(4), 781-790. https://dx.doi.org/10.1037/a0019477

Stark, S. (2002). A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The Generalized Graded Unfolding Model for multi-unidimensional paired comparison responses [Doctoral dissertation]. University of Illinois at Urbana-Champaign.

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

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458. https://dx.doi.org/10.1126/science.7455683

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

Weiss, B., & Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews. Journal of Applied Social Psychology, 36(4), 1070-1086. https://dx.doi.org/10.1111/j.0021-9029.2006.00055.x

Zickar, M. J. (2000). Modeling faking on personality tests. In D. R. Ilgen & C. L. Hulin (Eds.), Computational modeling of behavior in organizations: The third scientific discipline (pp. 95-113). American Psychological Association. https://dx.doi.org/10.1037/10375-005

RECEIVED 10/01/20 ACCEPTED 03/04/21