American Psychological Association

Journal of Applied Psychology

0021-9010/19/$12.00

2019, Vol. 104, No. 11, 1347–1368
http://dx.doi.org/10.1037/apl0000414

Does Forcing Reduce Faking? A Meta-Analytic Review of Forced-Choice Personality Measures in High-Stakes Situations

Mengyang Cao and Fritz Drasgow
University of Illinois at Urbana-Champaign

Forced-choice (FC) is a popular format for developing personality measures, where individuals must choose 1 or multiple statements from several options. Although FC measures have been proposed to reduce score inflation in high-stakes assessments, inconsistent results have been found in empirical studies regarding their effectiveness. In this study, we conducted a meta-analysis of studies comparing FC personality measure scores between low-stakes and (both simulated and actual) high-stakes situations. Results suggest that the overall score inflation effect size for FC personality measures is 0.06. In selection scenarios, score inflation for FC scales is much lower than the meta-analytic effect size for single-statement personality measures across most personality facets. The score inflation effect size was also found to vary across FC scale characteristics and study design factors. Specifically, FC scales were consistently found to be more faking-resistant when constructed with statements balanced in social desirability and with responses scored via a normative approach. FC scales constructed with the PICK format were also found to be faking-resistant, while more applicant-incumbent studies are needed to examine the fakability of MOLE FC scales. Evidence at the overall level supports the use of multidimensional scales and extremity balance of statements, but results are not consistent across personality facets, or when large samples are excluded. Personality facets of high relevance to the target job were found to exhibit larger inflation than facets of low relevance to the target job. Practical guidance on constructing and using FC personality measures for personnel selection purposes is provided.

Keywords: forced-choice, personality, faking, meta-analysis

With the increasing popularity of using personality inventories for personnel selection, there is also a growing concern about the possibility of response distortion (e.g., Dilchert, Ones, Viswesvaran, & Deller, 2006; Ellingson, Smith, & Sackett, 2001; Griffith, Chmielowski, & Yoshita, 2007; Hough & Oswald, 2008; Zickar & Robie, 1999). Response distortion, or *faking*, is commonly referred to as the tendency to respond in a way that creates a favorable impression when personality measures are implemented in high-stake contexts (Paulhus, 2002). Researchers and practitioners have raised concerns over the potential negative consequences of faking, such as inflated scores, decreased validity, and distorted rank orders, though empirical research remains inconclusive on whether or not the consequences are nontrivial (e.g., Berry & Sackett, 2009; Griffith et al., 2007; Komar, Brown, Komar, & Robie, 2008; Marcus, 2006; Mueller-Hanson, Heggestad, & Thornton, 2003; Schmit & Ryan, 1993).

To mitigate the faking problem, scholars have proposed using the forced-choice (FC) format as an alternative to single-statement Likert-type scales when constructing personality measures. The FC format forces test-takers to compare statements within an item block, and choose the statement that most/least describes themselves. Although the FC format is believed to reduce motivational response inflation by forcing a choice among equally desirable statements, inconsistent empirical results have been found regarding the effectiveness of FC measures in reducing faking (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006; Jackson, Wroblewski, & Ashton, 2000). Such inconsistency in empirical findings is likely due to several factors, such as the different characteristics of FC measures used in the studies, and the different study designs adopted in each study. For example, the Heggestad et al. (2006) study used an FC scale with a tetrad format, whereas the Christiansen, Burns, and Montgomery (2005) study used the pairwise preference format to construct FC scales. Moreover, although using the same type of tetrad FC scales, the Heggestad et al. (2006) study compared the score inflation across two samples, whereas the Jackson, Wroblewski, and Ashton (2000) study tracked the same group of participants in both low-stakes and high-stakes situations. To date, no comprehensive investigation has been con-

This article was published Online First May 9, 2019.

Mengyang Cao, Department of Psychology, University of Illinois at Urbana-Champaign; Fritz Drasgow, School of Labor & Employment Relations and Department of Psychology, University of Illinois at Urbana-Champaign.

The authors would like to thank Stephen Stark for his comments on previous versions of the manuscript. A previous version of this article was presented in a symposium at the 31st annual conference of the Society for Industrial and Organizational Psychology. Mengyang Cao is now a People Research Scientist at Facebook, Inc. This article was completed as part of his doctoral dissertation at the University of Illinois at Urbana-Champaign, and has no relevance to his work with his current employer. Fritz Drasgow participated in the development and validation of the Tailored Adaptive Personality Assessment System (TAPAS).

Correspondence concerning this article should be addressed to Mengyang Cao, who is now at Facebook, Inc., 1 Hacker Way, Menlo Park, CA 94025. E-mail: pkucmy@gmail.com

ducted to reveal what factors may potentially influence the fakability of FC measures.

In this study, we conducted a meta-analysis examining the score inflation effect size of forced-choice personality measures in simulated and actual high-stakes situations. Moreover, we examined two sets of moderators, FC scale characteristics and study designs, in order to explore the optimal conditions under which FC scales exhibit high resistance to faking.

Score Inflation on Forced-Choice Measures

In the literature, the term "forced-choice" has been inconsistently used to refer to scales constructed with a variety of formats. In order to determine the scope of this study, we define *forced-choice* (FC) measures as those scales consisting of blocks of multiple statements, with respondents instructed to choose the statement(s) that are most and/or least descriptive of themselves (Stark, Chernyshenko, & Drasgow, 2012). In other words, FC measures force respondents to make a choice or comparison among multiple alternatives, rather than ask respondents to provide ratings on each individual statement (Salgado, Anderson, & Tauriz, 2014). Based on this definition, scales forcing respondents to choose "yes" or "no" on single statements (e.g., the Minnesota Multiphasic Personality Inventory (MMPI)) are *not* considered as FC measures.

The FC format has been proposed as a faking prevention strategy when constructing personality scales for selection purpose (Dilchert & Ones, 2012). Central to this argument is the assumption that the FC format can prevent respondents from faking their responses to a statement solely because the content of the statement is socially desirable (Christiansen et al., 2005). With statements balanced in social desirability, unless respondents can distinguish which of the statements within an item block is more related to the purpose of selection, they may be unable to increase their scores as easily as on single-statement personality measures (Stark et al., 2012). Empirical results, however, do not consistently support this hypothesis. While some studies have found that FC measures result in less faking than single-statement measures (e.g., Christiansen et al., 2005; Jackson et al., 2000), others suggest no significant difference between the two formats (e.g., Heggestad et al., 2006).

With meta-analysis, we are able to obtain an overall estimate of the faking effect to synthesize the overall trend and address the discrepancies in empirical results. As with most faking research, we operationalize the faking effect as the level of *score inflation*, which is the standardized difference in personality scores between high-stakes situations and low-stakes situations. Two previous meta-analyses of single-statement personality measures have demonstrated that faking resulted in significant score inflation on most occasions (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Viswesvaran & Ones, 1999). In this study, we first examined whether FC personality measures, as designed to reduce faking, indeed shows smaller score inflation in high-stakes situations than single-statement personality measures.

Hypothesis 1: FC personality measures exhibit smaller score inflation than single-statement personality measures in high-stakes situations.

Previous meta-analyses of single statement measures found mixed results on how different personality facets are inflated in high-stakes situations. While one meta-analysis found that conscientiousness and emotional stability exhibit significant score inflation among studies using an applicant-incumbent design (Birkeland et al., 2006), another meta-analysis showed that all personality facets were inflated in studies with an induced faking design (Viswesvaran & Ones, 1999). For FC scales, there is dependency between personality facets displayed within the same item. Previous research has suggested that although individuals tend to fake personality scales by responding in a socially desirable manner, they do not necessarily inflate their scores to the same extent for all personality facets (Smith & McDaniel, 2012). Instead, they tailor their responses so that they produce different personality profiles for different target jobs (Furnham, 1990; Mahar, Cologon, & Duck, 1995; Martin, Bowen, & Hunt, 2002). In this meta-analysis, primary studies include different types of target jobs, which may induce different levels of score inflation across personality traits. Therefore, we propose, as a research question, examining how the score inflation effect on FC scales varies across the Big Five factors.

Research Question 1: Does the score inflation effect on FC measures vary across the Big Five personality factors?

FC Scale Characteristics Moderators

Type of FC Scales

Despite the considerable variation in specific formats, FC measures can be generally constructed with one of the three formats. They are the PICK format, where respondents are instructed to pick the statement that is most descriptive of them; the MOLE (i.e., MOst and LEast like me) format, where respondents are asked to choose statements that are most and least descriptive of them; and the RANK format, where respondents rank the statements in terms of their descriptiveness (Hontangas et al., 2015).

Most FC scales constructed with the PICK format are presented as pairwise preference tasks, where two statements are presented side by side in an item block, and respondents are asked to choose which of the two statements is "more like me." Typical FC scales using this format are the Edwards Personal Preference Schedule (EPPS; Edwards, 1959), the People Orientation Inventory (POI; Shostrom, 1963), and the Tailored Adaptive Personality Assessment System (TAPAS; Stark, Drasgow, & Chernyshenko, 2008). Some FC scales may present more than two statements in an item block (e.g., Haaland, 2000), but the underlying response process is the same as in the pairwise preference format, that is, comparing the statements and choosing the one that is most descriptive.

FC scales constructed with the MOLE format mostly use tetrads of statements as item blocks. Specifically, each item tetrad contains four statements, often with two representing positive traits and two representing negative traits. Respondents are instructed to choose one statement that is "most like me," and one statement that is "least like me." Typical FC scales with the MOLE format are the Occupational Personality Questionnaire (OPQ; Saville, Holdsworth, Nyfield, Cramp, & Mabey, 1984), the Gordon Personal Profile (Gordon, 1963a), and the Gordon Personal Inventory (Gordon, 1963b). The MOLE is also known as the *partial ranking*

format, as the ranking information is unavailable between the two unchosen statements.

The RANK format often presents multiple statements in a block, and respondents are asked to sort the statements based on how well each of the statements describes themselves. The RANK format is different from the tetrad format in that respondents are required to provide full ranking information for all statements, rather than only picking the top- and bottom-ranked statements. A popular variation of the RANK format is known as the Q-sort (Block, 1961). With the Q-sort format, respondents are asked to assign the statements in one of several categories ranging from "least like me" to "most like me." The number of statements to be assigned to each category, however, is predefined. Thus, respondents still need to make comparisons among the statements and rank them, making the Q-sort format a special case of the RANK format that allows ties among statements. Popular FC measures constructed in a RANK format are the California Adult Q-Sort (Block, 1978) and the Riverside Situational Q-Sort (Wagerman & Funder, 2009).

All three FC formats attempt to reduce the probability of socially desirable responding, and thus all should theoretically contribute to the reduction of score inflation in high-stakes situations. However, according to the results of previous empirical studies, FC measures with different formats are not consistently resistant to faking. Besides sampling errors among studies, such inconsistency may be a result of the different response process of each FC format. For example, to respond to FC measures with the PICK format, individuals only need to compare two statements and determine which is more characteristic of themselves. While responding to MOLE FC measures, individuals need to make more comparisons to determine the best and worst options; clearly this requires more cognitive processing, which increases the effort it takes to fake and may thus reduce faking. On the other hand, MOLE FC measures often include two positive and two negative statements in a tetrad. Consequently, the statements are often more polarized in extremity, making it less difficult to fake. In this meta-analysis, we posit it as a research question to examine whether or not FC measures constructed with different formats differ in the effectiveness of faking prevention.

Research Question 2: Does the amount of score inflation vary across different formats of FC measures (i.e., PICK, MOLE, and RANK)?

Dimensionality of FC Scales

FC scales can be constructed as either unidimensional or multidimensional. Unidimensional FC scales are mostly constructed with the pairwise preference format, where the two statements within each item block measure the same personality trait, but represent different locations on the latent trait continuum. A typical unidimensional FC scale is the Myers-Briggs Type Indicator (MBTI). In multidimensional FC scales, item blocks consist of statements measuring different personality traits. Examples of multidimensional FC scales include the aforementioned TAPAS, POI, and OPQ.

The fakability of unidimensional and multidimensional FC scales has not been compared in previous research. For unidimensional FC scales, the amount of information provided by an item depends on the distance between the two statement locations on the latent trait continuum (Chernyshenko et al., 2009). As a result, unidimensional FC items often consist of two statements representing the opposite ends of a bipolar trait, making it similar to a single-statement item with a bipolar response option (i.e., yes/no). Such a format may make it easy for job applicants to detect which statement represents the more desirable side of the measured trait and consequently encourage faking. Therefore, we propose that unidimensional FC scales are not as faking resistant as multidimensional FC scales, where it is difficult for respondents to detect which of the two traits is more desirable for the target job.

Hypothesis 2: The dimensionality of FC scales moderates score inflation on FC scales, such that multidimensional FC scales exhibit smaller score inflation than unidimensional FC scales.

Statement Assembly of Multidimensional FC Scales

A common practice in constructing multidimensional FC measures is to balance the social desirability of the statements included in the same item block. For example, pairwise preference FC scales often have matched desirability of the two statements within an item block (e.g., Christiansen et al., 2005), while tetrad FC scales tend to have two equally socially desirable statements and two equally socially undesirable statements within an item block (e.g., Heggestad et al., 2006; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). As found in Krug (1958), the differences between the desirability ratings of two statements within an item block significantly correlate with the choice of statements under induced faking, indicating that social desirability could potentially affect responses to FC measures in high-stakes situations.

Although matching the social desirability of statements is recommended in the process of constructing FC measures, it is sometimes overlooked in the development of certain FC scales. Moreover, the method of obtaining social desirability ratings may also lead to different assemblies of FC scales. Traditionally, social desirability ratings of personality statements were obtained by subject matter experts' ratings on how socially desirable each statement is generally considered (e.g., Heggestad et al., 2006; Jackson et al., 2000). However, it has been argued that the desirability of personality statements varies across contexts, and that desirability should be rated in the context specific to the target job (Converse et al., 2010). For example, if the FC measure is to be used for hiring sales representatives, ratings should be made on how desirable each statement is in terms of a sales representative. Findings of Converse et al. (2010) suggest that FC scales developed based on job-specific desirability ratings are more resistant to faking than scales developed based on job-irrelevant desirability ratings. In this meta-analysis, we examine if such a conclusion generalizes to all primary studies.

Hypothesis 3.1: Social desirability balance in constructing FC scales moderates the score inflation on FC scales, such that FC scales constructed with balanced social desirability exhibit less faking than FC scales without social desirability balance.

Hypothesis 3.2: For FC scales constructed with balanced social desirability, score inflation is lower for scales balanced

with job-specific desirability ratings than for scales balanced with general desirability ratings.

Another characteristic that is often overlooked in the statements assembly process of multidimensional FC scales is the extremity of statements. Previous research has shown that subject matter experts' ratings of statement extremity can be used for developing FC measures (Stark, Chernyshenko, & Guenole, 2011). To date, there has been no empirical investigation on whether statement extremity balance leads to more faking-resistant FC measures. When taking personality measures in high-stakes situations, respondents may have the ability to look for any positive signals in the statements that make them appear to be a "better candidate." Without matching the extremity of the two statements in pairwise preference FC scales where both statements are in the same direction (e.g., positive), respondents can simply choose the statement with higher extremity on the trait level, which will generally produce higher scores (Stark et al., 2012). We propose the following hypothesis to examine the above assertion.

Hypothesis 4: Extremity balance (of statements in the same directions) moderates score inflation on FC scales, such that FC scales constructed with matched extremity exhibit less inflation than scales constructed without matched statement extremity.

Scoring Methods of Multidimensional FC Scales

A common problem associated with scoring multidimensional FC measures is ipsativity. For FC scales constructed with the PICK format, the classic scoring method is to assign "1" to the endorsed statement, and "0" to the unchosen statement. Then scores on all statements measuring the same dimension are summed to compute the score for that dimension (e.g., Kirchner, 1962). If such a scoring method is used, the total score across all dimensions will be the same number for every respondent. As a result, the scores of dimensions will be dependent on each other, as a high score on one dimension has to be compensated by a low score on one or more other dimensions. Similarly, a constant total score can also be found in FC scales with the MOLE format, where "1" is assigned to any statement chosen as "most like me" and "−1" is assigned to any statement chosen as "least like me" (e.g., Bowen, Martin, & Hunt, 2002). Scores obtained by the above method are labeled as "*ipsative scores.*"

The ipsativity issue can be somewhat mitigated for certain MOLE FC scales with both positive and negative statements within an item block. For example, in the Gordon Personal Profile (GPP; Gordon, 1993) scale, responding "most like me" to positive statements or "least like me" to negative statements will be given 1 point, whereas responding "least like me" to positive statements or "most like me" to negative statements will be given −1 point. Such a scoring method does not necessarily result in a constant total score across all dimensions. However, it still generates a similar dependency problem among dimensions as in ipsative scoring. Thus, scores obtained by this scoring method are labeled as "*partially ipsative scores*" (Hicks, 1970).

Both ipsative and partially ipsative scores suffer from substantial psychometric problems when analyzed in the same way as scores obtained from single-statement Likert scales (i.e., *normative scores*). For example, dimension scores tend to correlate negatively with each other, and internal consistency is often lower for ipsative scales (Meade, 2004). A recent meta-analysis found that ipsative scoring of FC personality measures generally showed lower criterion-related validity than single-statement personality measures. Interestingly, partially ipsative scores were found to have higher criterion-related validity than ipsative scores, and even scores of single-statement measures (Salgado et al., 2014).

A remedy to the ipsativity issue is to use item response theory (IRT) to obtain normative scores for FC measures. Such models include the multi-unidimensional pairwise-preference (MUPP) model (Stark, Chernyshenko, & Drasgow, 2005), the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011), and the McCloy-Heggestad-Reeve unfolding model (McCloy, Heggestad, & Reeve, 2005). Results from empirical studies have consistently demonstrated that normative scores obtained from FC measures through IRT modeling are comparable to the scores obtained from single-statement measures, and no significant discrepancies have been found in terms of factor structure and criterion-related validity (e.g., Chernyshenko et al., 2009; Hontangas et al., 2015; Joubert, Inceoglu, Bartram, Dowdeswell, & Lin, 2015; Zhang et al., in press). In the context of personnel selection, the application of normative scoring is particularly critical for multidimensional FC scales, because ipsative scores are only appropriate for comparisons across personality dimensions *within* individuals, but they do not provide meaningful comparisons *between* individuals (Hicks, 1970). This is a major limitation of ipsative scores, as personnel selection practices always require between-person comparisons.

In empirical research on faking, some studies report the results of all personality facets included in the FC scale (Braun & Farrell, 1974; Heggestad et al., 2006), while other scholars tend to report only the personality facets that they believed are desirable for the target job (e.g., Christiansen et al., 2005; Converse et al., 2010). Due to the nature of ipsative scoring, the score inflation effect across all personality facets should sum to zero. As a result, there will be an overall inflation effect for ipsatively scored scales if some studies only report desirable traits. Additionally, ipsative scoring is based on classical testing theory (CTT), which does not account for item characteristics when scoring personality traits. In other words, with ipsative scoring individuals are scored in the same approach as long as they respond to the same items, regardless of whether they take the measure in a faking or honest condition. However, empirical research has shown that individuals actually use different response processes when they respond to personality measures between low- and high-stakes situations (Klehe et al., 2012), leading to a change in item parameters or even a shift of response models (O'Brien & LaHuis, 2011). As a result, the score inflation we observe in faking conditions could reflect a mix of item parameter drift and personality trait inflation, which ipsative scoring is unable to disentangle.

Normative scoring, on the other hand, is based on IRT and always requires a linking procedure that places the estimated trait values of faking and honest groups on the same underlying continuum (Tay, Meade, & Cao, 2015). For example, to compare the standardized test scores of two groups that took a test at different times, one would first need to perform a linking procedure to place the item parameters of the two question sets on the same discrimination and difficulty scale. IRT software typically assumes that the distribution of latent trait for each group being analyzed is standard normal, which conceals true differences. Similarly, the

normative scoring approach implicitly links scores by using the same set of item parameters to estimate trait values, thus yielding scores that are comparable across people and groups (Brown & Maydeu-Olivares, 2011; Stark, Chernyshenko, Drasgow, & White, 2012). This controls for the artifact of potential item parameter changes across low- and high-stakes situations, so that score inflation reflects true differences in trait estimates between the two groups, and consequently results in lower score inflation than ipsative scoring. Additionally, with ipsative scoring, the range of scores is constrained in both honest and faking conditions, whereas normative scoring is often based on maximum likelihood estimation, which can lead to a wide range of scores. Thus, the pooled within group standard deviation may be relatively smaller for ipsative scoring, which may result in a larger effect size. Empirical research has consistently suggested that normative scoring successfully reduces the amount of score inflation in selection scenarios compared with ipsative scoring, even for traits that are desirable for the target job (Guan, 2015; Luo, Liu, Zhang, & Wang, 2013). Thus, we proposed the following hypothesis.

Hypothesis 5: The scoring method moderates the level of score inflation on multidimensional FC personality measures, such that FC measures using normative scoring exhibit smaller inflation than FC measures using ipsative or partially ipsative scoring.

Study Design Moderators

Two research designs are typically adopted in empirical studies that examine the faking effect—the induced faking design and the applicant-incumbent design. In the induced faking design, participants are often asked to respond to personality scales as if they were applying for a much desired job. Responses in the induced faking condition are then compared with the responses from an honest condition, where participants are asked to respond honestly. In the applicant-incumbent design, comparisons are made between responses from a group of job applicants, who are assumed to engage in faking to get hired, and responses from a group of job incumbents, who are not motivated to fake.

Previous research using different faking study designs has generated different conclusions on how faking influences the validity of personality measures (e.g., Ellingson et al., 2001; Topping & O'Gorman, 1997). Moreover, two meta-analyses have suggested that studies using an induced-faking design generally reported more score inflation on single-statement personality measures than studies using an applicant-incumbent design (Birkeland et al., 2006; Viswesvaran & Ones, 1999). This is consistent with the hypothesis that respondents in induced faking conditions may have the tendency to please the researcher and exaggerate the extent to which they fake personality measures. Hence, results generated from the induced faking design should be considered as the upper bound of faking behaviors (Smith & Ellingson, 2002). Therefore, we proposed the following hypothesis on the type of study design.

Hypothesis 6: The type of design moderates the level of score inflation on FC personality measures, such that inflation is smaller in studies with an applicant-incumbent design than in studies with an induced faking design.

Although the applicant-incumbent design almost always requires between-subjects comparisons, induced faking studies can be conducted in either a between-subjects approach or a within-subjects approach. In the within-subjects design, comparisons are made between responses of the same sample of participants under two conditions, whereas in the between-subjects design, comparisons are made between two groups of participants (Cook, Campbell, & Day, 1979). An advantage of the within-subjects design is that it is effective in eliminating preexisting differences inherent in between-subjects applicant-incumbent designs. However, within-subject designs are also more easily affected by artifacts such as history and maturation effects. According to previous meta-analytic results, studies using within-subjects designs in general report larger faking effects on single-statement personality scales than those using between-subjects designs. As within-subjects design requires individuals to take the personality inventory at least twice while under different instructions, it is likely that the respondents become more sensitized to researcher's demand and more skillful in faking (Viswesvaran & Ones, 1999). In this meta-analysis, we examined the source of variance (within-subjects vs. between-subjects) as a potential moderator.

Hypothesis 7: The source of variance in study design moderates the level of score inflation on FC personality measures, such that inflation is smaller in between-subjects studies than in within-subjects studies.

Another potential moderator for studies using the induced faking design is the type of faking instruction used to induce faking. Two types of instructions are commonly adopted in induced faking studies—a "good impression" instruction, where respondents are instructed to intentionally inflate their personality and leave a good impression (e.g., Braun & LaFaro, 1967), and a "respond as applicants" instruction, where respondents are instructed to complete the FC scales as if they were applying for a job, either a specific job (e.g., Bowen et al., 2002) or a desired job in general (e.g., Anderson, Sison, & Wester, 1984). The "good impression" instruction may induce a higher level of score inflation as it directly appeals to inflation on personality traits, whereas the "respond as applicants" instruction provides a closer approximation to the actual selection scenario by implicitly suggesting faking. Therefore, we proposed the following hypothesis.

Hypothesis 8: The type of faking instruction moderates the level of score inflation on FC personality measures, such that inflation is smaller in studies with "respond as applicants" instructions than studies with "good impression" instructions.

For multidimensional FC scales, as the comparison is made between at least two personality dimensions, a faking strategy individuals may adopt is that they only choose the statement that seems to be more relevant to the target job. Such a response strategy will consequently lead to a particular personality profile where facets of high relevance to the target job exhibit higher score inflation than facets of low relevance to the target job. Thus, we also examined the moderation effect of personality facets in terms of their relevance to the target job.

Hypothesis 9: The relevance to the target job moderates the level of score inflation on multidimensional FC personality measures, such that inflation is smaller on personality facets of low relevance to the target job than on personality facets of high relevance to the target job.

Method

Literature Search and Inclusion Criteria

To locate primary studies for the meta-analysis, the authors along with six trained research assistants conducted a comprehensive literature search using the following approaches.

Keyword searching. Three sets of keywords used for the literature search are displayed in Table 1, with the first set related to faking, the second set related to forced-choice, and the third set related to personality. All possible combinations of keywords from at least two different sets were chosen to perform keyword searching in the *PsycINFO* and the *Business Resource Complete* databases.

Forced-choice scale searching. We searched the names of established forced-choice personality scales (e.g., "*Gordon Personal profile*"; "*Occupational Personality Questionnaire*"; "*Edwards Personal Preference Schedule*") in *Google Scholar* to locate studies that used those forced-choice scales in high-stakes situations.

Unpublished article searching. Keyword searching in the *ProQuest Dissertations and Theses* database was performed to locate unpublished theses and dissertations. We also conducted keyword searching in programs of the annual conference of Society for Industrial and Organizational Psychology (SIOP), the annual conference of Academy of Management (AOM), and the annual conference of International Personnel Assessment Council (IPAC) from 2011–2016 for unpublished conference presentations. A call for unpublished articles was posted on the listserv of the Human Resource division of the AOM as well as in the LinkedIn group of SIOP.

Reference searching. The reference lists of all qualified studies found in the above procedures were examined for more primary studies. We also examined the reference lists of previous meta-analyses on faking (Adair, 2014; Birkeland et al., 2006; Viswesvaran & Ones, 1999) and recent studies citing the above meta-analyses.

In order to be included in the meta-analysis, the study needed to use an FC personality measure both in a low-stakes situation (e.g., honest instruction, job incumbents sample) and in a simulated (i.e., faking instructions about a desirable job) or actual (i.e., a job applicants sample) high-stakes situation, and report necessary information (e.g., *M* and *SD*; *t* value) to compute an effect size for score inflation. As a result, 43 primary studies were identified, with 74 independent substudies (i.e., paired samples).

Table 1
Keywords for Literature Search

Set 1	Set 2	Set 3
Faking	Forced-choice	Personality
Fake	Paired comparison	
Response distortion	Pairwise preference	
Impression management	Rank ordering	
Social desirability	Ranking format	
Self-presentation	Q-sort	
Intentional distortion		
Score inflation		

Coding of Study Characteristics

Coding of primary studies was performed by Mengyang Cao and three experienced research assistants. Specifically, Mengyang Cao coded all the primary articles to serve as the benchmark. The three research assistants independently coded the first 10 articles for training purposes, and met with Mengyang Cao to resolve discrepancies through discussion. They then split the remaining articles to crosscheck with the coding results of Mengyang Cao. The average percentage of agreement between the three coders and Mengyang Cao was 95%.

The mean and standard deviation of the FC measures were recorded for both low-stakes and high-stakes situations to compute the standardized mean differences (i.e., *d*). For studies using a within-subjects design, we used Formula 8.12 in Schmidt and Hunter (2014, p. 351) to calculate effect sizes. Effect sizes for studies using a between-subjects design were computed with Formulas 7.5 and 7.6 in Schmidt and Hunter (2014, pp. 284–285). For consistency in the meta-analysis, all effect sizes were computed so that a positive value indicates a higher score in the high-stakes sample than in the low-stakes sample. Besides effect sizes, study characteristics were coded to examine moderation effects. Specifically, potential moderators included FC scale characteristics moderators, study design moderators, and personality facet moderators. Coding results of all primary studies are presented in Appendix A.

FC scale moderators. Characteristics of the FC scales used in each primary study were coded, including type of FC scale (PICK, MOLE, or RANK), dimensionality (unidimensional or multidimensional), social desirability balance (yes or no), extremity balance (yes or no), and scoring method (ipsative, partially ipsative, or normative). For FC scales that balanced the social desirability of statements within an item block, we also coded whether that balance was specific to the target job or irrelevant to the target job.

Study design moderators. Several study design features were coded for each independent substudy. Specifically, we recorded the type of design used in the study, comparing the studies using an induced faking design with those using an applicant-incumbent design. For induced faking studies, we also coded source of variance (within-subjects or between-subjects) and type of instruction (to leave a "good impression" or to "respond as job applicants").

Personality facet moderators. Many FC scales were not developed based on the five factor model (FFM). For those FC scales, we adopted several approaches to establish the correspondence between personality facets included in FC scales and the Big Five factors. First, we referred to the manual of the FC scale and examined whether a clear connection was provided between the personality facets and the FFM. Thirty-six percent of the primary studies fell into this category. Second, for FC scales that did not specify a connection with the FFM, we employed the taxonomy developed by previous meta-analyses of personality (e.g., Adair, 2014; Birkeland et al., 2006) to convert the personality facets to the FFM. 14% of the primary studies fell into this category. Third, for the remaining 50% primary studies without any of the above information, Mengyang Cao, along with three well-trained research assistants, made subjective ratings to code the personality facets into the FFM. Specifically, we used the descriptions of FFM facets provided by Costa and McCrae (1992) as the standard definitions, read the facet descriptions provided in the FC scale

manuals, and independently decided onto which Big Five factor each personality facet should be mapped. Personality facets that did not clearly represent a single FFM facet (e.g., *humility*) were coded as "other" and thus excluded from the personality facet moderator analysis. A personality facet was successfully determined if three out of four coders agreed on the categorization. Using this criterion, agreement was reached in 73.4% of all the undetermined personality facets. For other facets, disagreement was resolved through discussion among the coders. If agreement was still not reached through discussion, the personality facets were coded as "other."

For primary studies that provided a specific target job, we also coded the extent to which each personality facet was relevant to the target job. First, we retrieved the Top 5 work tasks for each occupation from O*NET. Second, independent ratings of job relevance were performed by Mengyang Cao and the same three research assistants as in the personality facet categorization. Specifically, each coder was asked to read the work tasks of each occupation as well as the definition of each personality facet, and decide how strongly each personality trait was relevant to perform the work tasks. Ratings were made on a 5-point Likert-type scale, with 1 = *not relevant at all*, 3 = *neutral*, and 5 = *very much relevant*. The average interrater agreement (Fleiss' κ) between each pair of coders was .72 (Fleiss, 1971), which indicates substantial agreement (Landis & Koch, 1977). To further mitigate the potential bias of subjective ratings, we computed the average ratings among four coders, and dichotomized the moderator such that average ratings above 3.0 were coded as "high relevance," and average ratings equal to or below 3.0 were coded as "low relevance." As a result, 71.4% of the personality facets were coded as "high relevance" to the target job. Dichotomized individual ratings of job relevance were consistent with the final dichotomous ratings in 95.5% of the cases. Coding results of personality facet moderators were presented in Appendix B.

Analysis

Meta-analysis was performed using the following steps proposed by Schmidt and Hunter (2014). First, following the suggestions of Schmidt and Hunter (2014, p. 336), all the *d*s were converted to point-biserial *r*s. Second, composites of *r*s were computed at the substudy level to avoid introducing potential dependency among effect sizes within each substudy. Personality facets describing a socially undesirable trait (e.g., *aggression*) were reversed coded before composites were computed. Third, weighted mean correlations were computed for each moderator using the formula provided by Schmidt and Hunter (2014). As some of the primary studies have unequal sample sizes between the two groups, simply using sample sizes as weights to calculate meta-analytic effect sizes can greatly underestimate sampling error and lead to inappropriate weighting (Laczo, Sackett, Bobko, & Cortina, 2005). To address this issue, we used Formula 7.23a from Schmidt and Hunter (2014, p. 293) to obtain an accurate estimate of the sampling error variance of *d*, then used the inverse of sampling error variance as the corrected weight of each substudy to calculate meta-analytic effect sizes and variances. Additionally, consistent with previous meta-analyses on faking (e.g., Birkeland et al., 2006; Viswesvaran & Ones, 1999), a bare-bone analysis was conducted without correcting for unreliability and range restriction to reflect the operational effect sizes. Fourth, meta-analytic results were converted back to the *d*-scale for the purpose of interpretation. Eighty percent credibility intervals and 95% confidence intervals were also computed for each meta-analysis. A wide credibility interval indicates a high level of heterogeneity and that potential moderators should be considered. A 95% confidence interval excluding zero indicates that the effect size *d* is significant at the .05 level. Fifth, the meta-analytic effect sizes were compared across different levels of each moderator to examine the moderation effect. Although no formal testing was recommended by Schmidt and Hunter (2014), we followed the suggestions in Aguinis, Sturman, and Pierce (2008) to compute the *t*-value between two meta-analytic effect sizes. The *t*-value was then compared to the critical value of a *t* distribution with $(k_1 + k_2 - 1)$ degree of freedom (k_1, k_2 denote the number of primary substudies used to compute the meta-analytic effect sizes) to determine the significance level.

Two of the primary studies (Drasgow et al., 2012; Griffith, Peterson, Quist, Benda, & Evans, 2008) have much larger corrected sample sizes than the other primary studies, which may strongly influence the meta-analytic results when comparing the effect sizes between moderators, as the differences may be primarily driven by the large samples. To examine the impact, we conducted all the moderator analyses with and without these two samples to demonstrate how these large samples influence the meta-analytic results.

Results

Table 2 presents the overall meta-analytic results for all primary studies. Note that all the meta-analytic effect sizes reported in this section and the tables are on the *d* scale. The score inflation effect size between high-stakes and low-stakes samples is 0.06 across all studies. In response to Research Question 1, the 95% confidence interval (0.02, 0.09) excludes zero, suggesting that the overall inflation effect is statistically significant at the .05 level. Nevertheless, $d = 0.06$ is smaller than a small effect size (0.20 as proposed by Cohen, 1992), and is much smaller than the most of the effect sizes reported in previous meta-analyses on the fakability of single-statement personality measures (Birkeland et al.,

Table 2
Meta-Analytic Results for Overall Effect Size

Variable	N	N_c	k	d	SD_d	CV_{10}	CV_{90}	CI_L	CI_U
Overall	267,586	38,328	74	.06	.15	−.13	.24	.02	.09
N	124,994	24,412	32	.00	.08	−.09	.10	−.02	.03
E	169,069	34,559	40	.16	.22	−.12	.43	.09	.22
O	217,100	26,022	36	.00	.10	−.13	.13	−.03	.03
A	260,288	35,943	38	.19	.35	−.26	.65	.08	.31
C	262,748	36,606	53	.23	.35	−.22	.67	.13	.32

Note. N = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; k = total number of effect sizes included in the meta-analysis; N = total sample size across all effect sizes; N_c = total corrected sample size (inverse of sampling error variances); d = sample size weighted mean effect size; SD_d = sample size-weighted observed standard deviation of effect size; CV_{10} and CV_{90} = 10% and 90% credibility values, respectively; CI_L and CI_U = lower and upper bounds, respectively, of the 95% confidence interval around the corrected mean effect size.

2006; Viswesvaran & Ones, 1999). Therefore, Hypothesis 1 is supported at the overall level. The 80% credibility interval [−0.13, 0.24] is wide and includes zero, suggesting that there is considerable heterogeneity to be examined by the moderator analyses.

To address Research Question 1, we coded the personality facets included in FC scales into the FFM, and examined the differences among the Big Five factors. As presented in Table 2, meta-analytic results suggest different levels of score inflation across the Big Five factors. Significant inflation was found in extraversion (d = 0.16, 95% CI [0.09, 0.22]), agreeableness (d = 0.19, 95% CI [0.08, 0.31]), and conscientiousness (d = 0.23, 95% CI [0.13, 0.32]), while no significant inflation was found in emotional stability (d = 0.00, 95% CI [−0.02, 0.03]) and openness to experience (d = 0.00, 95% CI [−0.03, 0.03]).

FC Scale Moderators

Table 3 summarizes the meta-analytic effect sizes for all FC scale moderators. Only one primary study was found using the RANK format, thus the FC scale moderator analysis was only conducted between the PICK format and the MOLE. Overall results suggested that the score inflation effect of the PICK format is 0.05, which is a very small effect albeit significant at the .05 level (95% CI [0.01, 0.09]), whereas the overall effect size d of studies using the MOLE format is 0.37, which is significant at the .05 level (95% CI [0.25, 0.49]). In response to Research Question 2, FC scales constructed with the PICK format had lower score inflation effect than FC scales with the MOLE format, $t(71) = 5.06$, $p < .05$. At the trait level, PICK scales were also found to exhibit lower score inflation than MOLE scales, except for agreeableness where no significant difference was found. Excluding the two large samples, the same findings held at the overall level and for the personality facets, except there was not a statistically significant difference in emotional stability.

For the dimensionality moderator, meta-analytic results showed that unidimensional FC scales generates significant score inflation (d = 0.13, 95% CI [0.08, 0.18]), whereas multidimensional FC scales led to trivial score inflation that was not statistically significant at the .05 level (d = 0.03, 95% CI [−0.01, 0.07]). In general, the score inflation effect was smaller for multidimensional FC scales than for unidimensional FC scales, $t(72) = -3.32$, $p < .05$. Such results also held at the trait level, except for openness to experience where unidimensional scales actually exhibited score suppression (d = −0.64). Hence, Hypothesis 2 is mostly supported with the full sample. However, meta-analytic results changed substantially when large samples were removed, where unidimensional scales exhibited less faking than multidimensional scales at the overall level as well as in agreeableness. Given the discrepancies and the relatively small number of studies examining unidimensional FC scales, results on the dimensionality moderator should be interpreted with cautions.

For multidimensional FC scales, we also examined desirability balance, extremity balance, and scoring method as potential moderators. Meta-analytic results suggest that desirability balance is effective, such that it resulted in nonsignificant score inflation (d = 0.02, 95% CI [−0.02, 0.07]) and smaller inflation than FC scales without desirability balance, $t(55) = -3.22$, $p < .05$. Similar results were also found at the trait level, except for openness to experience where the difference was not statistically significant. When large samples were removed, although the overall effect became nonsignificant, scales with desirability balance still exhibited smaller score inflation at the personality facet level except for openness to experience. Thus, Hypothesis 3.1 is mostly supported. Regarding Hypothesis 3.2, social desirability balance was found to be more effective if it was specific to the target job, $t(43) = 2.36$, $p < .05$, while at the trait level such differences were only significant for emotional stability and openness to experience. Thus, Hypothesis 3.2 is only partially supported with the full sample. With the large samples excluded, results became less consistent, such that job-specific balance only exhibited significantly smaller score inflation on openness to experience, while significantly larger score inflation was found for emotional stability and extraversion facets.

Consistent with Hypothesis 4, extremity balance was found to be a significant moderator, $t(57) = 2.00$, $p < .05$. Moreover, FC scales with extremity balance had a nonsignificant score inflation effect (d = 0.01, 95% CI [−0.06, 0.09]), while FC scales without extremity balance had a small but significant overall effect size (d = 0.14, 95% CI [0.04, 0.24]). At the trait level, however, extremity balance was found significantly effective only on emotional stability and conscientiousness scales. With the large samples removed, results were completely reversed, such that extremity balance exhibited even larger score inflation at both overall and facet levels.

Among the three scoring methods, the normative scoring method produced the lowest score inflation effect (d = 0.00) and was not statistically significant (95% CI [−0.02, 0.02]), supporting Hypothesis 5. Both ipsative (d = 0.31) and partially ipsative (d = 0.72) methods led to significant score inflation. Interestingly, the ipsative method resulted in a smaller score inflation than the partially ipsative method, $t(48) = -3.23$, $p < .05$. Results were consistent across personality facets, as well as when larger samples were removed with only two exceptions—normative scoring did not lead to significantly smaller score inflation than ipsative scoring on emotional stability and openness to experience.

Study Design Moderators

Table 4 presents the results of the study design moderator analyses. Consistent with Hypothesis 6, studies using the induced faking design generally reported larger score inflation than studies using the applicant-incumbent design, $t(73) = 4.81$, $p < .05$. Results were also consistent across personality facets. When large samples were excluded, though, significant results were only found at the overall level as well as on emotional stability and openness to experience, whereas the applicant-incumbent design exhibited larger score inflation on extraversion. Thus, Hypothesis 6 is only supported with the full sample. We also compared the trait level results for applicant-incumbent studies with the meta-analytic results for single-statement measures in Birkeland et al. (2006), and found that score inflation in FC measures is significantly lower on all personality facets than in single-statement measures, lending further support to Hypothesis 1.

Among induced faking studies, the between-subjects design generated significantly larger score inflation than within-subjects design, $t(58) = -2.07$, $p < .05$. At the trait level, however, this only applied to emotional stability scales. On the other four personality facets, within-subject design studies reported larger score

Table 3
Meta-Analytic Results for the Forced-Choice Scale Moderators

Variable	Full sample										No large sample				
	N	N_c	k	d	SD_d	CV_{10}	CV_{90}	CI_L	CI_U	t-value	N_c	k	d	SD_d	t-value
Type of FC scales															
PICK	263,096	37,304	40	.05	.13	−.12	.22	.01	.09		4,040	38	.12	.35	
N	121,857	23,712	12	.00	.05	−.07	.06	−.03	.03		1,049	11	.22	.09	
E	165,739	33,748	18	.15	.22	−.13	.44	.05	.26		1,306	16	.03	.34	
O	214,638	25,418	21	.00	.09	−.12	.12	−.04	.04		2,755	20	−.03	.28	
A	257,398	35,294	17	.20	.36	−.26	.66	.03	.37		2,853	15	−.01	.00	
C	259,070	35,664	26	.22	.35	−.23	.67	.09	.35		3,177	24	.14	.34	
MOLE	3,852	918	32	.37	.34	−.06	.81	.25	.49	5.06**	918	32	.37	.34	3.1**
N	2,499	595	18	.28	.29	−.09	.66	.15	.42	4.06**	595	18	.28	.29	.84
E	2,968	706	22	.32	.00	.32	.32	.32	.32	3.24**	706	22	.32	.00	3.53**
O	1,880	448	14	.26	.04	.20	.31	.23	.28	11.19**	448	14	.26	.04	4.46**
A	2,354	547	21	.08	.27	−.26	.42	−.04	.19	−1.15	547	21	.08	.27	1.55
C	3,446	789	28	.51	.38	.02	.99	.37	.65	2.9**	789	28	.51	.38	3.72**
Dimensionality															
Unidimensional	45,453	11,292	17	.13	.10	.00	.26	.08	.18		696	16	−.13	.40	
N	692	172	2	.25	.00	.25	.25	.25	.25		172	2	.25	.00	
E	42,834	9,854	6	.49	.00	.49	.49	.49	.49		75	5	.19	.00	
O	470	104	10	−.64	.00	−.64	−.64	−.64	−.64		104	10	−.64	.00	
A	43,230	9,956	6	.81	.18	.58	1.03	.67	.95		177	5	−.32	.00	
C	43,498	10,058	8	.79	.09	.68	.91	.73	.86		234	7	.30	.11	
Multidimensional	221,930	26,985	56	.03	.15	−.17	.22	−.01	.07	−3.32**	4,318	55	.16	.35	2.66**
N	124,099	24,190	29	.00	.08	−.10	.10	−.03	.03	−17.42**	1,528	28	.23	.21	−.57
E	126,308	24,656	35	.03	.08	−.08	.14	.00	.06	−32.49**	1,993	34	.12	.29	−1.47
O	216,483	25,817	26	.01	.09	−.11	.12	−.03	.04	35.47**	3,154	25	.03	.26	12.67**
A	216,957	25,940	33	−.02	.03	−.06	.02	−.03	−.01	−11.45**	3,278	32	.02	.09	21.44**
C	219,453	26,451	47	.03	.16	−.17	.23	−.01	.07	−19.48**	3,788	46	.20	.37	−1.4
Desirability balance															
No	2,243	424	12	.25	.23	−.05	.54	.12	.37		424	12	.25	.23	
N	819	145	6	.40	.18	.17	.62	.26	.54		145	6	.40	.18	
E	819	149	6	.30	.00	.30	.30	.30	.30		149	6	.30	.00	
O	647	106	2	.03	.00	.03	.03	.03	.03		106	2	.03	.00	
A	819	139	6	.21	.00	.21	.21	.21	.21		139	6	.21	.00	
C	819	140	6	.43	.15	.24	.63	.31	.55		140	6	.43	.15	
Yes	219,687	26,560	44	.02	.15	−.17	.21	−.02	.07	−3.22**	3,893	43	.16	.36	−1.05
N	123,280	24,046	23	.00	.07	−.09	.09	−.03	.03	−5.42**	1,383	22	.21	.21	−2.18*
E	125,489	24,507	29	.03	.09	−.08	.13	−.01	.06	−17.5**	1,844	28	.11	.30	−3.42**
O	215,836	25,711	24	.01	.09	−.11	.13	−.03	.04	−1.22	3,048	23	.03	.27	.09
A	216,138	25,801	27	−.02	.04	−.07	.03	−.03	−.01	−32**	3,138	26	.01	.11	−9.56**
C	218,634	26,311	41	.03	.15	−.17	.22	−.02	.07	−6.13**	3,648	40	.19	.38	−2.8**
Job-related rating															
Yes	211,701	24,652	8	.01	.12	−.14	.16	−.07	.09		1,985	7	.14	.39	
N	117,900	22,731	3	−.01	.01	−.03	.01	−.03	.00		68	2	.46	.00	
E	118,776	22,915	5	.02	.03	−.01	.06	.00	.05		253	4	.40	.00	
O	210,039	24,341	4	−.01	.05	−.08	.06	−.07	.04		1,679	3	−.19	.11	
A	210,039	24,352	4	−.02	.00	−.02	−.02	−.02	−.02		1,690	3	.03	.00	
C	211,123	24,572	8	.02	.07	−.07	.11	−.03	.07		1,909	7	.25	.08	
No	7,986	1,908	36	.17	.33	−.25	.59	.07	.28	2.36*	1,908	36	.17	.33	.2
N	5,164	1,261	19	.18	.22	−.09	.46	.09	.28	3.91**	1,261	19	.18	.22	−5.52**
E	6,497	1,538	23	.05	.32	−.36	.46	−.08	.18	.39	1,538	23	.05	.32	−5.24**
O	5,581	1,317	19	.32	.14	.15	.50	.26	.39	7.96**	1,317	19	.32	.14	7.1**
A	5,883	1,397	22	−.03	.17	−.25	.19	−.10	.04	−.23	1,397	22	−.03	.17	−1.57
C	7,295	1,686	32	.12	.53	−.56	.80	−.06	.30	1.02	1,686	32	.12	.53	−1.33
Extremity balance															
Yes	119,891	23,107	10	.01	.12	−.15	.17	−.06	.09		440	9	.79	.67	
N	118,367	22,836	6	−.01	.05	−.07	.06	−.05	.03		173	5	.70	.00	
E	118,997	22,969	8	.02	.02	.00	.05	.01	.04		306	7	.39	.00	
O	118,195	22,805	2	.00	.00	.00	.00	.00	.00		143	1	.13	.00	
A	118,367	22,841	6	−.02	.00	−.02	−.02	−.02	−.02		178	5	.13	.00	
C	119,209	22,995	9	.01	.08	−.09	.12	−.04	.06		333	8	.72	.25	
No	102,951	3,991	48	.14	.35	−.31	.59	.04	.24	2.00*	3,991	48	.14	.35	−2.85**
N	5,732	1,354	23	.17	.18	−.06	.40	.10	.25	4.23**	1,354	23	.17	.18	−14.27**
E	7,311	1,687	27	.07	.31	−.32	.47	−.04	.19	.8	1,687	27	.07	.31	−5.42**
O	98,288	3,011	24	.03	.27	−.32	.38	−.08	.14	.48	3,011	24	.03	.27	−1.85*
A	98,590	3,099	27	.01	.11	−.14	.16	−.03	.05	1.55	3,099	27	.01	.11	−5.52**
C	100,244	3,455	38	.16	.35	−.29	.61	.04	.27	2.29*	3,455	38	.16	.35	−5.32**

(*table continues*)

Table 3 (*continued*)

Variable	Full sample										No large sample				
	N	N_c	*k*	*d*	SD_d	CV_{10}	CV_{90}	CI_L	CI_U	*t*-value	N_c	*k*	*d*	SD_d	*t*-value
							Scoring method								
Ipsative	7,874	1,659	31	.31	.44	−.25	.86	.15	.46	3.23[a**]	1,659	31	.31	.44	3.23[a**]
N	2,965	664	14	.21	.28	−.15	.56	.06	.35	6.69[a**]	664	14	.21	.28	6.69[a**]
E	4,921	1,084	19	.31	.00	.31	.31	.31	.31	—	1,084	19	.31	.00	—
O	3,274	720	11	.15	.00	.15	.15	.15	.15	2.61[a**]	720	11	.15	.00	2.61[a**]
A	3,490	766	14	−.11	.00	−.11	−.11	−.11	−.11	11.22[a**]	766	14	−.11	.00	11.22[a**]
C	5,341	1,166	21	.38	.25	.06	.70	.27	.49	4.75[a**]	1,166	21	.38	.25	4.75[a**]
Partially ipsative	1,329	299	18	.72	.38	.23	1.21	.54	.90	−3.90[b**]	299	18	.72	.38	−3.59[b**]
N	650	148	10	.71	.00	.71	.71	.71	.71	−69.54[b**]	148	10	.71	.00	—
E	903	207	11	.42	.00	.42	.42	.42	.42	−13.5[b**]	207	11	.42	.00	−4.45[b**]
O	482	104	8	.47	.34	.03	.90	.23	.70	−3.71[b**]	104	8	.47	.34	−2.87[b**]
A	740	156	12	.53	.20	.28	.78	.42	.64	−9.71[b**]	156	12	.53	.20	−8.93[b**]
C	1,385	305	19	.73	.21	.46	1.00	.63	.83	−11.37[b**]	305	19	.73	.21	−4.45[b**]
Normative	212,984	25,092	8	.00	.03	−.04	.04	−.02	.02	−7.97[c**]	2,425	7	.00	.09	−7.48[c**]
N	120,741	23,442	6	−.01	.03	−.04	.02	−.03	.01	−2.84[c**]	780	5	.15	.00	−.81[c]
E	120,484	23,365	5	.01	.07	−.08	.10	−.05	.07	−9.9[c**]	703	4	−.25	.30	−3.73[c**]
O	212,727	24,992	7	.00	.09	−.11	.11	−.07	.06	−4.5[c**]	2,330	6	−.02	.29	−1.43[c]
A	212,727	25,018	7	−.02	.00	−.02	−.02	−.02	−.02	—	2,355	6	.02	.00	—
C	212,727	24,980	7	.01	.11	−.13	.14	−.07	.09	−5.45[c**]	2,317	6	.05	.36	−2.13[c*]

Note. *N* = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; *k* = total number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; N_c = total corrected sample size (inverse of sampling error variances); *d* = sample size weighted mean effect size; SD_d = sample size-weighted observed standard deviation of effect size; CV_{10} and CV_{90} = 10% and 90% credibility values, respectively; CI_L and CI_U = lower and upper bounds, respectively, of the 95% confidence interval around the corrected mean effect size. [a] (Partially ipsative − Ipsative); [b] (Normative − Partially ipsative); [c] (Normative − Ipsative).
[*] $p < .05$. [**] $p < .01$.

inflation than between-subjects studies. When large samples were excluded, no significant difference was found at the overall level, while mixed results were found for different personality traits. Hence, Hypothesis 7 is only partially supported with the full sample at the personality trait level. Comparing the effect sizes to the meta-analytic results for single-statement measures in Viswesvaran and Ones (1999), we found that score inflation in FC measures is only significantly lower on emotional stability (d_{FC} = 0.37 vs. d_{SS} = 0.64), openness to experience (d_{FC} = 0.11 vs. d_{SS} = 0.65), and agreeableness (d_{FC} = 0.00 vs. d_{SS} = 0.48) among between-subjects studies, and on emotional stability (d_{FC} = 0.20 vs. d_{SS} = 0.93) and openness to experience (d_{FC} = 0.29 vs. d_{SS} = 0.76) among within-subjects studies. Indeed, larger score inflation in FC measures was found on extraversion (d_{FC} = 0.31 vs. d_{SS} = 0.16) among between-subjects studies, and on agreeableness (d_{FC} = 0.75 vs. d_{SS} = 0.47) among within-subjects studies. This leads to mixed support for Hypothesis 1.

Another moderator examined among studies using an induced faking design was the type of faking instructions. Overall, meta-analytic results supported Hypothesis 8 such that when individuals were instructed to fake as job applicants, they inflated their scores on FC scales to a smaller extent than when they were instructed to leave a good impression, $t(63) = -2.16$, $p < .05$. At the personality trait level, such results only held on emotional stability, extraversion, and conscientiousness, whereas reversed results were found on the other two personality facets. No significant difference was found at the overall level when large samples were removed, whereas "respond as applicants" instructions led to smaller score inflation on all personality facets but openness to experience.

Table 5 presents the meta-analytic results on the level of relevance of the personality facet to the target job. Meta-analytic results indicated that individuals tended to inflate to a larger extent on personality facets that are of high relevance to the target job than those of low relevance to the target job, $t(50) = 4.13$, $p < .05$. Thus, Hypothesis 9 is supported. Moreover, although a significant inflation effect was found in personality facets highly related to the target job (d = 0.12, 95% CI [0.06, 0.18]), a weak but significant suppression effect was found in personality facets that are of low relevance to the target job (d = −0.16, 95% CI [−0.29, −0.04]).[1]

Unique Effect of Each Moderator

To address the potential overlap between moderators and detect the unique effect of each moderator, we used the "lm" function in R to conduct weighted linear regression with the *d* effect sizes of each study as the outcome, dummy-coded FC scale and study design moderators as predictors, and corrected sample sizes as weights (Nye, Su, Rounds, & Drasgow, 2012). As some moderators are nested within each other, we performed three separate regression models using only the studies that are applicable to each moderator. Results are presented in Table 6, where the β coefficients indicate the unique effect of each moderator after controlling for the effects of other moderators in the model. In general, most moderation effects are consistent with the results from independent moderator analyses, suggesting that the overlap between moderators does not confound the meta-analytic results when examining each moderator independently. The only inconsistent

[1] Although the dichotomization of the job relevance ratings could mitigate the impact of inaccuracy and discrepancies in inter-rater subjective ratings, it led to a loss of information by artificially transforming a continuous variable to a categorical variable. Thus, we also examined the correlation between the mean of original ratings across four raters and the inflation effect size of each personality facet. The correlation turned out to be .16, which is statistically significant at .05 level. This further supported Hypothesis 9 that the relevance of personality facet to the target job is positively associated with the level of score inflation.

Table 4
Meta-Analytic Results for the Study Design Moderators

Variable	Full sample										No large sample				
	N	N_c	*k*	*d*	SD_d	CV_{10}	CV_{90}	CI_L	CI_U	*t*-value	N_c	*k*	*d*	SD_d	*t*-value
Type of design															
Applicant-incumbent	213,951	25,256	15	.01	.07	−.08	.09	−.03	.04		2,589	14	.06	.21	
N	118,476	22,871	6	−.01	.00	−.01	−.01	−.01	−.01		208	5	−.01	.08	
E	119,879	23,181	9	.03	.03	−.01	.07	.01	.05		519	8	.30	.00	
O	211,671	24,710	9	−.01	.05	−.08	.06	−.04	.03		2,047	8	−.13	.14	
A	211,671	24,719	9	−.02	.00	−.02	−.02	−.02	−.02		2,057	8	−.01	.00	
C	212,242	24,835	12	.02	.08	−.08	.12	−.02	.07		2,173	11	.24	.16	
Induced faking	53,635	13,072	59	.16	.20	−.10	.41	.11	.21	4.81**	2,475	58	.26	.45	2.58*
N	6,518	1,541	26	.27	.17	.06	.49	.21	.34	8.64**	1,541	26	.27	.17	6.01**
E	49,466	11,379	33	.43	.20	.18	.69	.36	.50	11.17**	1,600	32	.06	.30	−4.72**
O	5,485	1,262	28	.24	.29	−.13	.60	.13	.34	4.34**	1,262	28	.24	.29	5.01**
A	48,719	11,224	31	.71	.33	.29	1.13	.59	.83	12.36**	1,445	30	−.01	.22	.06
C	50,912	11,721	44	.70	.35	.25	1.14	.59	.80	11.73**	1,896	43	.18	.50	−.64
Source of variance															
Between-person	5,331	1,169	22	.28	.27	−.07	.63	.17	.39		1,169	22	.28	.27	
N	2,822	631	11	.37	.13	.21	.53	.29	.45		631	11	.37	.13	
E	2,830	605	9	.31	.00	.31	.31	.31	.31		605	9	.31	.00	
O	1,771	384	8	.11	.00	.11	.11	.11	.11		384	8	.11	.00	
A	2,365	514	7	.00	.37	−.48	.48	−.28	.27		514	7	.00	.37	
C	4,026	864	16	.54	.24	.23	.85	.42	.66		864	16	.54	.24	
Within-person	48,304	11,903	37	.14	.19	−.09	.38	.08	.20	−2.07*	1,306	36	.25	.56	−.26
N	3,696	910	15	.20	.16	.00	.41	.12	.28	−2.94**	910	15	.20	.16	−2.94**
E	46,636	10,774	24	.44	.20	.18	.70	.36	.52	3.17**	995	23	−.09	.30	−6.49**
O	3,714	878	20	.29	.34	−.15	.73	.14	.44	2.45*	878	20	.29	.34	2.45*
A	46,354	10,710	24	.75	.27	.40	1.10	.64	.86	4.95**	931	23	−.01	.04	−.09
C	46,886	10,857	28	.71	.35	.26	1.16	.58	.84	1.85*	1,033	27	−.10	.49	−5.74**
Faking instruction															
Good impression	740	170	16	.50	.63	−.31	1.30	.19	.80		170	16	.50	.63	
N	310	68	7	.99	.00	.99	.99	.99	.99		68	7	.99	.00	
E	484	112	11	.59	.00	.59	.59	.59	.59		112	11	.59	.00	
O	274	62	7	−.58	.00	−.58	−.58	−.58	−.58		62	7	−.58	.00	
A	346	78	8	.19	.00	.19	.19	.19	.19		78	8	.19	.00	
C	484	108	11	.84	.00	.84	.84	.84	.84		108	11	.84	.00	
Respond as applicants	53,187	12,969	48	.15	.19	−.09	.39	.10	.21	−2.16*	2,373	47	.25	.43	−1.46
N	6,208	1,473	19	.24	.16	.03	.45	.17	.32	−19.87**	1,473	19	.24	.16	−19.87**
E	48,888	11,275	24	.43	.21	.17	.70	.35	.52	−3.7**	1,496	23	.03	.33	−8.17**
O	5,291	1,222	23	.26	.28	−.09	.62	.15	.38	14.43**	1,222	23	.26	.28	14.43**
A	48,453	11,166	25	.71	.33	.29	1.14	.58	.84	7.82**	1,387	24	−.02	.26	−3.98**
C	50,278	11,605	34	.70	.35	.25	1.15	.58	.82	−2.33*	1,781	33	.17	.52	−7.41**

Note. *N* = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; *k* = total number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; N_c = total corrected sample size (inverse of sampling error variances); *d* = sample size weighted mean effect size; SD_d = sample size-weighted observed standard deviation of effect size; CV_{10} and CV_{90} = 10% and 90% credibility values, respectively; CI_L and CI_U = lower and upper bounds, respectively, of the 95% confidence interval around the corrected mean effect size.
* $p < .05$. ** $p < .01$.

result is that unidimensional scales were found to exhibit nonsignificantly less score inflation when controlling for scale type and type of design ($\beta = -0.07$). The weighted regression results also suggested that type of design is more influential on score inflation than the other overall moderators, and that scoring method is the most influential moderator among all moderators examined within multidimensional scales. Although the PICK format still demonstrated less score inflation than the MOLE format, the difference was not significant after controlling for dimensionality and type of design. From the R^2 statistics, we also learned that source of variance and faking instruction only accounted for a small portion of variance among induced faking studies ($R^2 = .02$), possibility due to the large variance within each moderator category. On the contrary, the three moderators examined explained more than half of the variance among studies using multidimensional scales ($R^2 = .59$), demonstrating that these three factors substantially influence how fakable the FC measures are.

With the concern that the ratio between the number of predictors (i.e., five) and the number of primary studies (i.e., 56) in Model 3 might be too low for regression analysis (Green, 1991), we conducted two follow-up models, Model 3a and Model 3b, to separately examine the effects of social desirability/extremity balance and the effects of scoring methods. As shown in Table 6, although the coefficients of social desirability/extremity bal-

Table 5
Meta-Analytic Results for Job Relevance and Publication Bias

Variable	N	N_c	k	d	SD_d	CV_{10}	CV_{90}	CI_L	CI_U	t-value
Job relevance										
High	259,626	36,646	33	.12	.17	−.10	.33	.06	.18	
Low	257,758	36,208	18	−.16	.26	−.50	.17	−.29	−.04	−4.13**
Publication bias										
Published	10,459	2,302	60	.30	.46	−.29	.90	.19	.42	
Unpublished	257,127	36,026	14	.04	.08	−.06	.14	.00	.08	−4.15**

Note. k = total number of effect sizes included in the meta-analysis; N = total sample size across all effect sizes; N_c = total corrected sample size (inverse of sampling error variances); d = sample size weighted mean effect size; SD_d = sample size-weighted observed standard deviation of effect size; CV_{10} and CV_{90} = 10% and 90% credibility values, respectively; CI_L and CI_U = lower and upper bounds, respectively, of the 95% confidence interval around the corrected mean effect size.
** $p < .01$.

ance are not statistically significant, these moderators still explain a decent amount of variance of the score inflation effect sizes ($R^2 = .21$).

Publication Bias

To examine the potential bias of the publication status of primary studies, we conducted separate meta-analyses for published journal articles and unpublished studies, including dissertations, conference presentations, and technical reports. As shown in Table 5, published studies generally reported larger score inflation than unpublished studies, $t(73) = 4.15$, $p < .05$. Although most cases of publication bias assume that studies with significant results are more likely to be published, scholars in favor of FC scales as a faking prevention strategy should be more motivated to publish nonsignificant results, which indicate that FC scales are faking resistant. A possible explanation of the results is that unlike published studies that tended to use preexisting FC scales, unpublished technical reports often adopted a rigorous approach to develop new FC scales, which were often equipped with more faking resistant characteristics.

A major concern of publication bias is that studies with small effect sizes are likely to be suppressed from publication. This can be addressed by inspecting a funnel plot, where effect sizes on the horizontal axis are plotted against sample sizes on the vertical axis (Sterne, Becker, & Egger, 2005). As a few large samples were included in this meta-analysis, the logarithm of sample sizes was taken before creating the funnel plot. As illustrated in Figure 1, many studies reported near zero effect sizes. We examined the rank order correlation between standardized effect sizes and sampling error variances to detect if effect sizes are symmetrically distributed. Kendall's rank correlation $\tau = -0.03$, which is not statistically significant ($z = -0.34$, $p > .05$), indicating that the meta-analytic results are not biased by large sample sizes.

Table 6
Weighted Linear Regression Results for Forced-Choice Scale and Study Design Moderators

Model	b	SE	β	t	R^2
Model 1 (All effect sizes)					.22
Induced faking = 1 (vs. applicant-incumbent = 0)	.25	.10	.15	2.58*	
PICK = 1 (vs. MOLE = 0)	−.20	.16	−.14	−1.22	
Unidimensional = 1 (vs. multidimensional = 0)	−.11	.10	−.07	−1.06	
Model 2 (Induced faking only)					.02
Between-person = 1 (vs. within-person = 0)	.14	.14	.09	.96	
Good impression = 1 (vs. respond as applicants = 0)	.24	.46	.13	.51	
Model 3 (Multidimensional only)					.59
Nonspecific desirability balance = 1 (vs. no desirability balance = 0)	−.03	.10	−.03	−.25	
Job-specific desirability balance = 1 (vs. no desirability balance = 0)	−.08	.12	−.06	−.70	
Extremity balance = 1 (vs. no Extremity balance = 0)	−.05	.04	−.04	−1.03	
Partially ipsative scoring = 1 (vs. ipsative scoring = 0)	.54	.12	.56	4.40**	
Normative scoring = 1 (vs. ipsative scoring = 0)	−.29	.08	−.21	−3.67**	
Model 3a (Multidimensional only)					.21
Nonspecific desirability balance = 1 (vs. no desirability balance = 0)	−.07	.14	−.07	−.51	
Job-specific desirability balance = 1 (vs. no desirability balance = 0)	−.22	.14	−.15	−1.62	
Extremity balance = 1 (vs. no extremity balance = 0)	−.04	.06	−.03	−.72	
Model 3b (Multidimensional only)					.31
Partially ipsative scoring = 1 (vs. ipsative scoring = 0)	.34	.26	.25	1.28	
Normative scoring = 1 (vs. ipsative scoring = 0)	−.42	.11	−.22	−3.84**	

* $p < .05$. ** $p < .01$.

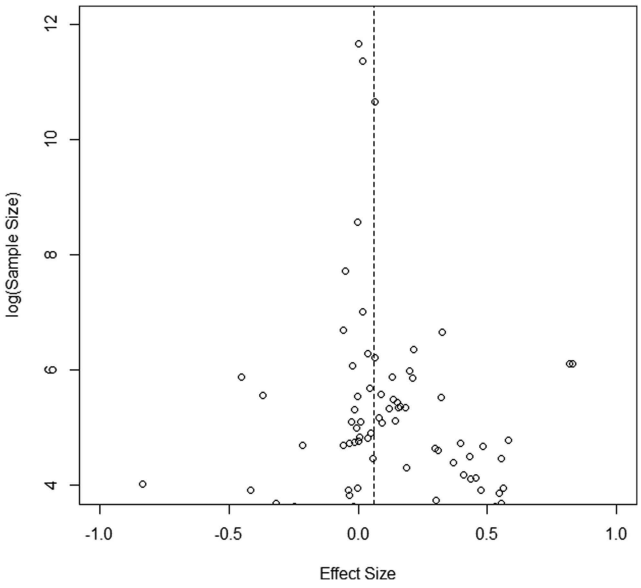

Figure 1. Funnel plot for examining publication bias.

Discussion

Despite the growing interest in using forced-choice (FC) measures as a faking prevention strategy in personnel selection contexts (e.g., Boyce, Conway, & Caputo, 2015; Stark et al., 2008; Underhill, Bearden, & Chen, 2008), empirical research has shown inconsistent results regarding whether or not the FC format successfully reduces the magnitude of faking in personality assessment (Heggestad et al., 2006; Jackson et al., 2000). This meta-analysis provides a comprehensive investigation of the overall score inflation on FC personality measures, and how the magnitude of score inflation varies across FC scales with different characteristics and studies with different designs. The overall score inflation effect across all FC personality measures is 0.06, which is a very tiny effect. Although mixed results were found across personality traits among induced-faking studies, the score inflation effect of FC measures was consistently smaller than the effect of single-statement personality measures among job applicants (Birkeland et al., 2006; Viswesvaran & Ones, 1999). We can conclude that compared to the single-statement format, the FC format reduces faking on personality measures in real-life selection process.

Across personality traits, significant score inflation, albeit small in effect sizes, was only found in extraversion, agreeableness, and conscientiousness. This is inconsistent with the previous meta-analysis where conscientiousness and emotional stability were found to exhibit higher score inflation than the other three dimensions (Birkeland et al., 2006). We found that the score inflation effects of personality dimensions are generally comparable among applicant-incumbents studies, while the differences were mostly driven by studies using induced faking designs. A possible reason for this result is that the majority of the simulated target jobs provided in the primary studies were top executive managers, sales representatives, and customer service representatives. Participants were likely to believe that agreeableness and extraversion are more desirable traits for those jobs, as people high in those facets are more inclined to cooperation, socialization, and exerting influence. Although emotional stability is also relevant to those jobs, it might have been suppressed due to the ipsative nature of FC scales. Another important finding of this meta-analysis is that personality facets that are of high relevance to the target job exhibited larger score inflation than facets of low relevance to the target job. This is consistent with results for single-statement measures in that respondents are more likely to fake the personality facets that, they believe, are more desirable for the target job (Furnham, 1990). Moreover, FC measures in general exhibit significant inflation on job-relevant personality facets, and significant suppression on facets of low-relevance to the target job.

Our meta-analytic results also revealed considerable variability in the magnitude of score inflation across FC scales. Among the FC scale characteristics moderators we examined, consistent results were found in scale type and scoring method moderators—statistically significant results were found in both full-sample and without-large-sample results, in overall results and across most personality traits, as well as in weighted regression models controlling for the effects of other moderators. Specifically, FC scales with the PICK format are more faking-resistant than scales with the MOLE format, and normative scoring leads to smaller score inflation than ipsative or partially ipsative scoring. Previously, the MOLE format was claimed to be superior to the PICK format in that MOLE FC scales are less likely to suffer from ipsativity issues due to the availability of partially ipsative scoring (Hicks, 1970). However, partially ipsative scores were found to exhibit even larger score inflation effect than ipsative scores. Moreover, normative scores can now be more easily obtained from PICK FC scales with the development of IRT normative scoring (Brown & Maydeu-Olivares, 2011; Stark et al., 2005). Normative scoring also overcomes the issue of response process change by linking all individuals' trait estimates to the same scale, so that score inflation will purely reflect the trait differences between faking and honest groups.

We would like to reiterate that although normative scoring leads to reduced score inflation compared with ipsative scoring, it does not *prevent* faking. Nevertheless, linking metrics may compensate for the possible response process shift when people take personality measures for selection purposes, such that the computed score differences more accurately capture score inflation between faking and honest conditions. We also note that weighted regression results showed that the effect of scale type was not statistically significant after controlling for type of design. A closer examination of primary studies revealed that only five samples used the applicant-incumbent design to study MOLE scales, while the sample size weighted effect size is only 0.12, suggesting that the MOLE format may be potentially faking-resistant at least in actual selection scenarios. Given that all the five samples were collected in the 1950s and 1960s, we would like to see more field studies using the MOLE format for selection purposes before we conclude on the fakability of MOLE FC scales. Thus, the PICK format and normative scoring are recommended based on the results of the current meta-analysis, though we call for more empirical studies using the applicant-incumbent design to further examine the fakability of the MOLE format.

Mixed results were found on the other FC scale moderators: dimensionality, social desirability balance, and extremity balance. Multidimensional scales exhibited less score inflation than unidi-

mensional scales with the full sample, but results did not hold when large samples were excluded, or when controlled for scale type and type of design. Among the few studies that used unidimensional scales and could be mapped to the Big Five model, most of them used the Myers-Briggs Type Indicator (MBTI) scale, the reliability and construct validity of which have been challenged by many scholars (e.g., Barbuto, 1997; Boyle, 1995). It is possible that the inconsistent results could be attributed to the lack of reliability and construct validity of unidimensional scales included in this meta-analysis. For social desirability balance, we found that balancing the social desirability across statements helps reduce score inflation on most personality facets. Regarding whether the desirability balance is specific to the job, however, no consistent support was found, especially when large samples were removed. The same findings applied to extremity balance. Admittedly, both moderator analyses were based on a small number of studies and, moreover, there appeared to be a considerable lack of agreement in the methods used to obtain extremity and job-specific desirability ratings of statements. For example, Jackson et al. (2000) used proportion of endorsement as the index of extremity, whereas Heggestad et al. (2006) obtained extremity through factor loadings. It is possible that the effect of extremity and job-specific balance varied across studies using different methods. Nevertheless, social desirability balance, especially job-specific desirability balance, combined with extremity balance still explained substantial amount of variance in the overall effect size.

For study design moderators, smaller score inflation was found in applicant-incumbent studies than induced faking studies as hypothesized. Weighted regression analysis also indicated that applicant-incumbent design led to smaller score inflation, even after controlling for scale type and dimensionality. This suggests that for FC personality measures, faking is less of a problem in actual selection scenarios than in induced situations. Among induced faking studies, no consistent results were found on source of variance or faking instructions. Weighted regression results also showed that both moderators explained minimal variance in the overall effect size. This suggests that these study design factors do not consistently bias the performance of FC personality measures in a systematic way.

Practical Implications

Besides the contributions to the literature on FC measures, the current study also highlights several implications for practitioners who are interested in leveraging FC measures as faking-resistant personality assessment tools. For instance, the meta-analytic results provide general guidance on the steps of constructing faking-resistant FC personality measures. First, one needs to choose a format for the FC scale. Meta-analytic results recommend the use of the PICK format, as it consistently exhibited very small score inflation. Besides, the MOLE is also more cognitively loaded, which may cause potential problems of adverse impact when used in selection scenarios. Second, although it remains unclear whether multidimensional scales outperform unidimensional scales, when multidimensional scales are chosen, statements within an item block should always be balanced in social desirability. Ideally, social desirability should be rated concerning the specific target job, though it may limit the generalizability of FC measures to other jobs. Third, responses to multidimensional FC measures should be scored in a normative way. Although scoring method does not affect how people respond to FC scales, normative scoring does lead to reduced score inflation by linking the metrics, even for traits that are job relevant. More importantly, it also allows comparisons among individuals, which is essential when making hiring decisions.

Finally, the meta-analytic results are also informative for the choice of utilizing personality scores to make personnel selection decisions. In personnel selection practices, personality assessment results are often considered either by only using the personality facets that are related to a job as predictors, or by taking the composite across all personality facets (Hogan, Hogan, & Roberts, 1996). For the former approach, FC measures are advantageous to single-statement measures because of the reduced level of score inflation in selection scenarios. For the composite approach, one needs to be cautious about the weight assigned to each personality facet to compute the composites. As shown in the meta-analysis, FC measures can lead to different levels of inflation and even suppression effects on different personality traits in high-stakes situations. Thus, it might be necessary to reexamine the weights obtained from a low-stakes situation (e.g., through multiple regression; Sackett & Lievens, 2008) to test whether they are still applicable to FC measures in high-stakes situations.

Limitations and Future Directions

Despite the implications described in previous sections, results from the current meta-analysis are limited on several aspects. First, the focus of this meta-analysis is the score inflation effect, which is only one of the several consequences of faking. Due to the limited number of primary studies and the incomplete information reported in those studies, we were unable to examine other faking consequences, such as reduction in criterion-related validity and change in rank orders. Moreover, the interpretation of the score inflation effect should be cautious, as it only represents the overall effect, which may not necessarily apply to each individual study (Viswesvaran & Ones, 1999). Future empirical studies are still needed to rigorously examine the effect of the FC scale moderators by experimentally controlling confounding factors.

Second, as we noted in the Method section, two of the primary studies (Drasgow et al., 2012; Griffith et al., 2008) have very large corrected sample sizes, which appeared to influence some of the meta-analytic results as shown in the comparison between full-sample and no-large-sample results. For example, without the two large samples, multidimensional scales ($d = 0.16$) were no longer found to be more faking resistant than unidimensional scales ($d = 0.13$), and scales with extremity balance ($d = 0.79$) exhibited higher score inflation than scales without extremity balance ($d = 0.14$). We want to acknowledge that the scales used in both large-sample studies went through rigorous scale development process and have demonstrated excellent performance in practice, thus we are still confident that the meta-analytic results with the two large samples are reliable. Nevertheless, we presented the meta-analytic results with and without large samples side by side so that researchers are aware of the discrepancies and to what extent results are influenced by the large samples.

Third, unlike previous faking meta-analyses that primarily focused on moderator analyses at each Big Five personality facet level (Birkeland et al., 2006; Viswesvaran & Ones, 1999), this

meta-analysis also reports the overall effect size for each level of moderator across all personality facets. This is mainly because most FC scales were constructed prior to the time when the FFM became popular; hence, subjective coding was necessary to map the facets onto the Big Five. In fact, only a subset of personality facets can be clearly mapped to one of the FFM, resulting in a reduced set of primary studies at the facet level and consequently lower power in detecting moderators. In addition, from a practical perspective, composite personality scores are often used in making selection decisions, suggesting that there is still value reporting the pooled results. Therefore, we presented both overall and facet level meta-analytic results for moderator analyses in Tables 3 and 4. In general, results of most moderator analyses at the facet level are consistent with the overall results in terms of directions (i.e., positive or negative), though statistical significance levels may vary across facets. We would like to raise a caveat on the significant testing of moderators at the facet level, as power is substantially limited by the reduced number of primary studies. In this case, effect sizes are more appropriate for interpreting the results than significance testing.

Fourth, due to the limited number of primary studies, several potential moderators could not be examined in this meta-analysis. For example, only one study used the RANK format, and only one study adopted a within-subject applicant-incumbent design, making it impossible to compute a meta-analytic effect for those moderators. There were also not enough primary studies to compare the different methods of balancing the extremity of statements. Therefore, this meta-analysis should be considered as a call for more primary studies investigating the potential factors that may affect the effectiveness of FC measure in faking reduction. Additionally, the number of primary studies also limited further examinations of the interaction between moderators. Meta-analytic results exhibited wide credibility intervals for some moderator analysis, suggesting that there is still substantial heterogeneity within each level of the moderator. We attempted to address the heterogeneity issue by conducting weighted regression models with multiple moderators. Results suggested that the heterogeneity within studies using multidimensional scales can be effectively explained by the three moderators examined (i.e., social desirability balance, extremity balance, and scoring method), whereas only little variance is accounted for among induced faking studies with source of variance and faking instruction moderators.

Conclusion

In conclusion, meta-analytic results of this study suggest that the forced-choice format generally reduces the score inflation on personality measures in personnel selection scenarios, though many factors can potentially temper the fakability of FC measures, making them more susceptible to socially desirable responding. To construct a faking-resistant FC personality scale, scholars are recommended to consider using the PICK format, balance the social desirability of statements, and adopt a normative scoring approach. There is some evidence supporting the use of multidimensional scales (over unidimensional) and extremity balance, but results remain inconclusive. We believe the current meta-analysis serves to debunk the myth of inconsistent results regarding the performance of FC personality scales in selection scenarios, and we expect to see more empirical research dedicated to further explore the fakability of FC measures.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Adair, C. (2014). *Interventions for addressing faking on personality assessments for employee selection: A meta-analysis* (Unpublished doctoral dissertation). DePaul University, Chicago, IL.

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods, 11,* 9–34. http://dx.doi.org/10.1177/1094428106292896

*Anderson, H. N., Sison, G., & Wester, S. (1984). Intelligence and dissimulation on the personal orientation inventory. *Journal of Clinical Psychology, 40,* 1394–1398. http://dx.doi.org/10.1002/1097-4679(198411)40:6<1394::AID-JCLP2270400620>3.0.CO;2-R

Barbuto, J. E., Jr. (1997). A critique of the Myers-Briggs Type Indicator and its operationalization of Carl Jung's psychological types. *Psychological Reports, 80,* 611–625. http://dx.doi.org/10.2466/pr0.1997.80.2.611

Berry, C. M., & Sackett, P. R. (2009). Faking in personnel selection: Tradeoffs in performance versus fairness resulting from two cut-score strategies. *Personnel Psychology, 62,* 835–863. http://dx.doi.org/10.1111/j.1744-6570.2009.01159.x

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

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Thomas. http://dx.doi.org/10.1037/13141-000

Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.

*Bowen, C. C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis, 10,* 240–259. http://dx.doi.org/10.1108/eb028952

*Boyce, A., Conway, J. S., & Caputo, P. M. (2015). *Development and validation of Aon Hewitt's personality model and Adaptive Employee Personality Test (ADEPT-15)*. Unpublished technical report, Aon Hewitt Consulting, New York, NY.

Boyle, G. J. (1995). Myers-Briggs Type Indicator (MBTI): Some psychometric limitations. *Australian Psychologist, 30,* 71–74. http://dx.doi.org/10.1111/j.1742-9544.1995.tb01750.x

*Braun, J. R. (1962). Effects of a top management faking set on the Gordon Personal Inventory. *Psychological Reports, 10,* 611–614. http://dx.doi.org/10.2466/pr0.1962.10.3.611

*Braun, J. R. (1963a). Effects of positive and negative faking sets on the survey of interpersonal values. *Psychological Reports, 13,* 171–173. http://dx.doi.org/10.2466/pr0.1963.13.1.171

*Braun, J. R. (1963b). Fakability of the Gordon Personal Inventory: Replication and extension. *The Journal of Psychology, 55,* 441–444. http://dx.doi.org/10.1080/00223980.1963.9916638

*Braun, J. R. (1965). Effects of specific instructions to fake on Gordon Personal Profile Scores. *Psychological Reports, 17,* 847–850. http://dx.doi.org/10.2466/pr0.1965.17.3.847

*Braun, J. R., & Farrell, R. M. (1974). Re-examination of the fakability of the Gordon Personal inventory and profile: A reply to Schwab. *Psychological Reports, 34,* 247–250. http://dx.doi.org/10.2466/pr0.1974.34.1.247

*Braun, J. R., & LaFaro, D. (1967). Effects of a good impression set on the Thorndike dimensions of temperament. *Journal of Educational Measurement, 4,* 237–240. http://dx.doi.org/10.1111/j.1745-3984.1967.tb00592.x

*Braun, J. R., & La Faro, D. (1969). A further study of the fakability of the Personal Orientation Inventory. *Journal of Clinical Psychology, 25,* 296–299. http://dx.doi.org/10.1002/1097-4679(196907)25:3<296::AID-JCLP2270250323>3.0.CO;2-V

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

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22,* 105–127. http://dx.doi.org/10.1080/08959280902743303

*Christiansen, N. D. (1997). *The development and validation of a job-related choice method of personality assessment* (Unpublished doctoral dissertation). Northern Illinois University, Dekalb, IL.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

*Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance, 23,* 323–342. http://dx.doi.org/10.1080/08959285.2010.501047

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653–665. http://dx.doi.org/10.1016/0191-8869(92)90236-I

*Dicken, C. F. (1959). Simulated patterns on the Edwards Personal Preference Schedule. *Journal of Applied Psychology, 43,* 372–378. http://dx.doi.org/10.1037/h0044779

Dilchert, S., & Ones, D. S. (2012). Application of preventative strategies. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 177–200). Oxford, UK: Oxford University Press.

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

*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 selection and classification decisions* (Technical Report 1311). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

*Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15,* 13–24. http://dx.doi.org/10.1111/j.1744-6570.1962.tb01843.x

Edwards, A. L. (1959). *Personal preference schedule*. New York, NY: Psychological Corporation.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86,* 122–133. http://dx.doi.org/10.1037/0021-9010.86.1.122

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382. http://dx.doi.org/10.1037/h0031619

*Fluckinger, C. D. (2010). *Measurement of Big Five Personality via Q-Sort: Comparison with a Likert measure and test-taker perceptions and reactions* (Unpublished doctoral dissertation). University of Akron, Akron, OH.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology, 9,* 46–55. http://dx.doi.org/10.1007/BF02686767

Gordon, L. V. (1963a). *Gordon Personal Inventory*. San Diego, CA: Harcourt, Brace & World.

Gordon, L. V. (1963b). *Gordon Personal Profile: Manual*. San Diego, CA: Harcourt, Brace & World.

Gordon, L. V. (1993). *Gordon Personal Profile-Inventory: Manual* (revised). New York, NY: Psychological Corporation.

*Gordon, L. V., & Stapleton, E. S. (1956). Fakability of a forced-choice personality test under realistic high school employment conditions. *Journal of Applied Psychology, 40,* 258–262. http://dx.doi.org/10.1037/h0043595

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499–510. http://dx.doi.org/10.1207/s15327906mbr2603_7

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341–355. http://dx.doi.org/10.1108/00483480710731310

*Griffith, R. L., Peterson, M. H., Quist, J., Benda, A., & Evans, A. L. (2008, April). *Faking the personality profile: Easier said than done*. Symposium presented at the 23rd annual conference of Society for Industrial and Organizational Psychology, San Francisco, CA.

*Guan, L. (2015). *Personality, faking, and the ability of identify criteria: Can forced choice formats untangle their relationships?* Unpublished master thesis, University of Virginia, Charlottesville, VA.

*Haaland, D. E. (2000). *Self-assessment of interpersonal competency: Development and validation of a forced-choice method to minimize response distortion in job applicant contexts* (Unpublished doctoral dissertation). Central Michigan University, Mount Pleasant, MI.

*Hedberg, R. (1962). More on forced-choice test fakability. *Journal of Applied Psychology, 46,* 125–127. http://dx.doi.org/10.1037/h0038453

*Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91,* 9–24. http://dx.doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167–184. http://dx.doi.org/10.1037/h0029780

*Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality, 42,* 1323–1333. http://dx.doi.org/10.1016/j.jrp.2008.04.006

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist, 51,* 469–477. http://dx.doi.org/10.1037/0003-066X.51.5.469

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39,* 598–612. http://dx.doi.org/10.1177/0146621615585851

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

*Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13,* 371–388. http://dx.doi.org/10.1207/S15327043HUP1304_3

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and Likert scale versions of a personality instrument. *International Journal of Selection and Assessment, 23,* 92–97. http://dx.doi.org/10.1111/ijsa.12098

*Kaess, W. A., & Witryol, S. L. (1957). Positive and negative faking on a forced-choice authoritarian scale. *Journal of Applied Psychology, 41,* 333–339. http://dx.doi.org/10.1037/h0043451

*Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology, 15,* 241–261. http://dx.doi.org/10.1080/13594320600625872

*Kirchner, W. K. (1962). "Real-life" faking on the Edwards personal preference schedule by sales applicants. *Journal of Applied Psychology, 46,* 128–130. http://dx.doi.org/10.1037/h0039528

*Kirchner, W. K., Dunnette, M. D., & Mousley, N. (1960). Use of the Edwards personal preference schedule in the selection of salesmen. *Personnel Psychology, 13,* 421–424. http://dx.doi.org/10.1111/j.1744-6570.1960.tb02099.x

Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25,* 273–302. http://dx.doi.org/10.1080/08959285.2012.703733

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93,* 140–154. http://dx.doi.org/10.1037/0021-9010.93.1.140

Krug, R. E. (1958). A selection set preference index. *Journal of Applied Psychology, 42,* 168–170. http://dx.doi.org/10.1037/h0046461

Laczo, R. M., Sackett, P. R., Bobko, P., & Cortina, J. M. (2005). A comment on sampling error in the standardized mean difference with unequal sample sizes: Avoiding potential errors in meta-analytic and primary research. *Journal of Applied Psychology, 90,* 758–764. http://dx.doi.org/10.1037/0021-9010.90.4.758

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. http://dx.doi.org/10.2307/2529310

*Larson, N. L., Lewis, R. J., O'Neill, T. A., & Carswell, J. J. (2013, April). *Are forced choice personality measures contaminated by general mental ability?* Poster presented at the 28th annual conference of the Society for Industrial and Organizational Psychology, Houston, TX.

*Longstaff, H. P., & Jurgensen, C. E. (1953). Fakability of the Jurgensen classification inventory. *Journal of Applied Psychology, 37,* 86–89. http://dx.doi.org/10.1037/h0057806

*Luo, F., Liu, H., Zhang, D., & Wang, S. (2013). The development of forced-choice personality scale for neuroticism and its effect of faking resistance. *Experimental Psychology, 33,* 460–464.

*Mahar, D., Coburn, B., Griffin, N., Hemeter, F., Potappel, C., Turton, M., & Mulgrew, K. (2006). Stereotyping as a response strategy when faking Personality questionnaires. *Personality and Individual Differences, 40,* 1375–1386. http://dx.doi.org/10.1016/j.paid.2005.11.018

*Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. *Personality and Individual Differences, 18,* 605–609. http://dx.doi.org/10.1016/0191-8869(94)00200-C

Marcus, B. (2006). Relationships between faking, validity, and decision criteria in personnel selection. *Psychological Science, 48,* 226–246.

*Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32,* 247–256. http://dx.doi.org/10.1016/S0191-8869(01)00021-6

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8,* 222–248. http://dx.doi.org/10.1177/1094428105275374

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77,* 531–552. http://dx.doi.org/10.1348/0963179042596504

*Mudd, J. (2005). *Revealing socially undesirable information: A comparison of bipolar adjective scaling methods*. Unpublished master thesis, Western Kentucky University, Bowling Green, KY.

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

*Norman, W. T. (1963). Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology, 47,* 225–241. http://dx.doi.org/10.1037/h0042106

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7,* 384–403.

O'Brien, E., & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment, 19,* 109–118. http://dx.doi.org/10.1111/j.1468-2389.2011.00539.x

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

*Rusmore, J. T. (1956). Fakability of the Gordon Personal Profile. *Journal of Applied Psychology, 40,* 175–177. http://dx.doi.org/10.1037/h0042524

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59,* 419–450. http://dx.doi.org/10.1146/annurev.psych.59.103006.093716

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, 88,* 797–834. http://dx.doi.org/10.1111/joop.12098

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L., & Mabey, W. (1984). *The Occupational Personality Questionnaires (OPQ)*. London, UK: Saville & Holdsworth (UK) Ltd.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

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,* 966–974. http://dx.doi.org/10.1037/0021-9010.78.6.966

*Schwab, D. P. (1971). Issues in response distortion studies of Personality inventories: A critique and replicated study. *Personnel Psychology, 24,* 637–647. http://dx.doi.org/10.1111/j.1744-6570.1971.tb00377.x

*Shipley, W. C., Gray, F. E., & Newbert, N. (1946). The personal inventory; its derivation and validation. *Journal of Clinical Psychology, 2,* 318–322. http://dx.doi.org/10.1002/1097-4679(194610)2:4<318::AID-JCLP2270020403>3.0.CO;2-I

Shostrom, E. L. (1963). *The personal orientation inventory*. San Diego, CA: EdITS.

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87,* 211–219. http://dx.doi.org/10.1037/0021-9010.87.2.211

Smith, D. B., & McDaniel, M. (2012). Questioning old assumptions: Faking and the personality–performance relationship. In R. Griffith & M. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 53–69). Greenwich, CT: Information Age Publishing.

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2012). Constructing fake-resistant personality tests using item response theory. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 214–239). Oxford, UK: Oxford University Press.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15,* 463–487. http://dx.doi.org/10.1177/1094428112444611

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts' ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods, 14,* 256–278. http://dx.doi.org/10.1177/1094428109356712

Stark, S., Drasgow, F., & Chernyshenko, O. S. (2008, September). *Update on Tailored Adaptive Personality Assessment System (TAPAS): The next generation of personality assessment systems to support personnel selection and classification decisions*. Paper presented in 50th annual conference of the International Military Testing Association, Amsterdam, the Netherlands.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). West Sussex, UK: Wiley.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18,* 3–46. http://dx.doi.org/10.1177/1094428114553062

Topping, G. D., & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23,* 117–124. http://dx.doi.org/10.1016/S0191-8869(97)00006-8

Underhill, C. M., Bearden, R. M., & Chen, H. T. (2008). *Evaluation of the fake resistance of a forced-choice paired-comparison computer adaptive personality measure*. Millington, TN: Navy Personnel Research Studies and Technology. http://dx.doi.org/10.21236/ADA484384

*Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19,* 175–199. http://dx.doi.org/10.1207/s15327043hup1903_1

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

Wagerman, S. A., & Funder, D. C. (2009). Personality psychology of situations. In P. J. Corr & G. Matthews (Eds.), *Cambridge handbook of personality*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511596544.005

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

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84,* 551–563. http://dx.doi.org/10.1037/0021-9010.84.4.551

Appendix A

List of Primary Studies and Coding Results

Study	Type of design	Source of variance	Faking instruction	Type of scale	Dimensionality	SD balance	Extremity balance	Scoring method	*d*	*N*
Anderson, Sison, & Wester, 1984 (1)	Induced faking	Between	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	−.26	18
Anderson et al., 1984 (2)	Induced faking	Between	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	−.42	18
Boyce, Conway, & Caputo, 2015 (1)	Induced faking	Within	Respond as applicants	PICK	Multidimensional	Job-specific	No	Normative	.16	176
Boyce et al., 2015 (2)	Induced faking	Within	Respond as applicants	PICK	Multidimensional	General	No	Normative	.33	216
Boyce et al., 2015 (3)	Induced faking	Within	Respond as applicants	PICK	Multidimensional	No	No	Normative	.37	212
Boyce et al., 2015 (4)	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	Job-specific	No	Normative	.03	86974
Boyce et al., 2015 (5)	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	Job-specific	No	Normative	−.01	5269
Braun & Farrell, 1974 (1)	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.97	61
Braun & Farrell, 1974 (2)	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.96	90
Braun & La Faro, 1969 (1)	Induced faking	Within	Good impression	PICK	Unidimensional	N/A	N/A	N/A	−1.32	22
Braun & La Faro, 1969 (2)	Induced faking	Within	Good impression	PICK	Unidimensional	N/A	N/A	N/A	−.68	40
Braun & La Faro, 1969 (3)	Induced faking	Within	Good impression	PICK	Unidimensional	N/A	N/A	N/A	−.51	38
Braun & LaFaro, 1967 (1)	Induced faking	Within	Good impression	MOLE	Multidimensional	No	Yes	Partially ipsative	.68	30
Braun & LaFaro, 1967 (2)	Induced faking	Within	Good impression	MOLE	Multidimensional	No	Yes	Partially ipsative	.63	42
Braun & LaFaro, 1967 (3)	Induced faking	Within	Good impression	MOLE	Multidimensional	No	Yes	Partially ipsative	.6	34
Braun & LaFaro, 1967 (4)	Induced faking	Within	Good impression	MOLE	Multidimensional	No	Yes	Partially ipsative	.9	66
Braun, 1962	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	1.25	38

(*Appendices continue*)

Appendix A (*continued*)

Study	Type of design	Source of variance	Faking instruction	Type of scale	Dimensionality	SD balance	Extremity balance	Scoring method	*d*	*N*
Braun, 1963a	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Ipsative	0	52
Braun, 1963b (1)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	1.31	48
Braun, 1963b (2)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	1.07	50
Braun, 1965 (1)	Induced faking	Within	Good impression	MOLE	Multidimensional	General	No	Partially ipsative	1.31	24
Braun, 1965 (2)	Induced faking	Within	Good impression	MOLE	Multidimensional	General	No	Partially ipsative	1.35	52
Braun, 1965 (3)	Induced faking	Within	Good impression	MOLE	Multidimensional	General	No	Partially ipsative	1.02	62
Braun, 1965 (4)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	1.34	88
Christiansen, 1997	Induced faking	Between	Respond as applicants	PICK	Multidimensional	Job-specific	Yes	Ipsative	.41	400
Christiansen, Burns, & Montgomery, 2005	Induced faking	Between	Respond as applicants	PICK	Multidimensional	Job-specific	No	Ipsative	.43	350
Converse et al., 2010 (1)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	General	No	Ipsative	1.1	107
Converse et al., 2010 (2)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	General	No	Ipsative	.87	113
Converse et al., 2010 (3)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	General	No	Ipsative	.65	100
Converse et al., 2010 (4)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	General	No	Ipsative	.63	104
Dicken, 1959	Induced faking	Within	Good impression	PICK	Multidimensional	General	No	Ipsative	−.04	38
Drasgow et al., 2012	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	Job-specific	Yes	Normative	0	117620
Dunnette, McCartney, Carlson, & Kirchner, 1962 (1)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Ipsative	.07	124
Dunnette et al., 1962 (2)	Applicant-incumbent	N/A	N/A	MOLE	Multidimensional	General	No	Ipsative	.01	126
Dunnette et al., 1962 (3)	Applicant-incumbent	N/A	N/A	MOLE	Multidimensional	General	No	Ipsative	−.07	113
Dunnette et al., 1962 (4)	Applicant-incumbent	N/A	N/A	MOLE	Multidimensional	General	No	Ipsative	−.06	166
Fluckinger, 2010	Induced faking	Between	Respond as applicants	RANK	Multidimensional	No	No	Ipsative	−.04	435
Gordon & Stapleton, 1956 (1)	Applicant-incumbent	N/A	N/A	MOLE	Multidimensional	General	No	Ipsative	.28	242
Gordon & Stapleton, 1956 (2)	Applicant-incumbent	N/A	N/A	MOLE	Multidimensional	General	No	Ipsative	.24	209
Griffith, Peterson, Quist, Benda, & Evans, 2008	Induced faking	Within	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	.13	42500
Guan, 2015	Induced faking	Within	Respond as applicants	PICK	Multidimensional	General	No	Normative	−.1	2260
Haaland, 2000	Induced faking	Between	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	.08	544
Hedberg, 1962	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Ipsative	0	118
Heggestad, Morrison, Reeve, & McCloy, 2006	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	Yes	Ipsative	.44	575
Hirsh & Peterson, 2008	Induced faking	Between	Respond as applicants	PICK & RANK	N/A	N/A	N/A	N/A	−.03	203
Jackson, Wroblewski, & Ashton, 2000	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	Yes	Ipsative	.32	212
Kaess & Witryol, 1957 (1)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	No	No	Ipsative	.13	507
Kaess & Witryol, 1957 (2)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	No	No	Ipsative	.18	264
Kaess & Witryol, 1957 (3)	Induced faking	Between	Respond as applicants	PICK	Multidimensional	No	No	Ipsative	.26	361
Kanning & Kuhne, 2006 (1)	Induced faking	Between	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	−.11	110
Kanning & Kuhne, 2006 (2)	Applicant-incumbent	N/A	N/A	PICK	Unidimensional	N/A	N/A	N/A	−.44	110
Kirchner, 1962 (1)	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	General	No	Ipsative	.02	166
Kirchner, 1962 (2)	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	General	No	Ipsative	−.03	115
Kirchner, Dunnette, & Mousley, 1960	Applicant-incumbent	N/A	N/A	PICK	Multidimensional	General	No	Ipsative	.04	1122
Larson, Lewis, O'Neill, & Carswell, 2013	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.68	253
Longstaff & Jurgensen, 1953 (1)	Induced faking	Within	Both	MOLE	Multidimensional	No	No	Ipsative	.79	82
Longstaff & Jurgensen, 1953 (2)	Induced faking	Within	Both	MOLE	Multidimensional	No	No	Ipsative	.38	74
Longstaff & Jurgensen, 1953 (3)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	No	No	Ipsative	.1	136
Luo, Liu, Zhang, & Wang, 2013	Induced faking	Between	Respond as applicants	PICK	Multidimensional	General	No	Ipsative & Normative	0	257

(*Appendices continue*)

Appendix A (*continued*)

Study	Type of design	Source of variance	Faking instruction	Type of scale	Dimensionality	SD balance	Extremity balance	Scoring method	*d*	*N*
Mahar et al., 2006 (1)	Induced faking	Within	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	−.08	50
Mahar et al., 2006 (2)	Induced faking	Within	both	PICK	Unidimensional	N/A	N/A	N/A	−.02	24
Mahar et al., 2006 (3)	Induced faking	Within	both	PICK	Unidimensional	N/A	N/A	N/A	−.02	24
Mahar, Cologon, & Duck, 1995	Induced faking	Within	Both	PICK	Unidimensional	N/A	N/A	N/A	.11	88
Martin, Bowen, & Hunt, 2002	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	No	Ipsative	.09	294
Mudd, 2005	Applicant-incumbent	N/A	N/A	PICK	Unidimensional	N/A	N/A	N/A	1.42	120
Norman, 1963 (1)	Induced faking	Within	Respond as applicants	PICK	Multidimensional	Job-specific	Both	Ipsative	2.96	456
Norman, 1963 (2)	Induced faking	Within	Respond as applicants	PICK	Multidimensional	Job-specific	Both	Ipsative	2.84	456
Rusmore, 1956	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.19	162
Schwab (1)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.63	22
Schwab (2)	Induced faking	Within	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	1.33	40
Shipley, Gray, & Newbert, 1946 (1)	Applicant-incumbent	N/A	N/A	PICK	Unidimensional	N/A	N/A	N/A	.69	784
Shipley et al., 1946 (2)	Applicant-incumbent	N/A	N/A	PICK	Unidimensional	N/A	N/A	N/A	−.11	815
Underhill, Bearden, & Chen, 2008	Induced faking	Within	Respond as applicants	PICK	Unidimensional	N/A	N/A	N/A	−.01	148
Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006	Induced faking	Between	Respond as applicants	MOLE	Multidimensional	General	No	Partially ipsative	.29	167

Note. PICK = pick one format; MOLE = "most like me/ least like me" format; RANK = ranking format; SD = social desirability. Numbers in parentheses indicate the sample number of each primary study.

Appendix B

List of Studies With Specific Occupations and Job Relevance Coding

Study	Occupation	Personality facet	Big five	Job relevance
Bowen, Martin, & Hunt, 2002	Junior manager	Conscientiousness	C	High
		Socially Confident	—	High
		Forward Planning	—	High
		Optimistic	N	Low
		Achieving	C	High
		Innovative	O	Low
		Relaxed	N	Low
		Democratic	—	Low
		Detail Conscious	C	High
		Caring	A	High
Boyce et al., 2015 (1)	Junior manager	Cooperativeness	A	High
		Sensitivity	A	Low
		Humility	—	Low
		Composure	N	High
		Positivity	N	High
		Awareness	—	High
		Assertiveness	E	High
		Liveliness	E	High
		Drive	C	High
		Structure	C	High
		Conceptual	O	Low
		Flexibility	O	Low
		Mastery	—	High
		Ambition	—	High
		Power	—	High

(*Appendices continue*)

Appendix B (*continued*)

Study	Occupation	Personality facet	Big five	Job relevance
Boyce et al., 2015 (2)	Assembly line worker	Drive	C	High
		Flexibility	O	Low
		Cooperation	A	High
		Sensitivity	A	Low
		Humility	—	Low
Boyce et al., 2015 (3)	Junior manager	Drive	C	High
		Flexibility	O	Low
		Cooperation	A	High
		Sensitivity	A	Low
		Humility	—	Low
Braun & Farrell, 1974	Top management executive	Cautiousness	C	High
		Original Thinking	O	High
		Personal Relations	A	High
		Vigor	—	High
		Ascendancy	E	High
		Responsibility	C	High
		Emotional Stability	N	High
		Sociability	E	High
Braun, 1962	Top management executive	Cautiousness	C	High
		Original Thinking	O	High
		Personal Relations	A	High
		Vigor	—	High
Braun, 1963a	Top management executive position	Support	A	High
		Conformity	—	Low
		Recognition	—	High
		Independence	—	High
		Benevolence	A	Low
		Leadership	E	High
Braun, 1963b	Top management executive position	Cautiousness	C	High
		Original Thinking	O	High
		Personal Relations	A	High
		Vigor	—	High
Braun, 1965	Top management executive position	Ascendancy	E	High
		Responsibility	C	High
		Emotional Stability	N	High
		Sociability	E	High
Christiansen, 1997	Sales representative	Conscientiousness	C	High
		Extraversion	E	High
Christiansen et al., 2005	Sales representative	Conscientiousness	C	High
		Extraversion	E	High
Converse et al., 2010	Police officer	Emotional Stability	N	High
		Conscientiousness	C	High
Drasgow et al., 2012	Army soldier	Achievement	C	High
		Adjustment	N	High
		Cooperation	A	High
		Dominance	E	Low
		Even Tempered	N	High
		Attention Seeking	E	Low
		Selflessness	A	High
		Intellectual Efficiency	O	Low
		Nondelinquency	C	High
		Order	C	High
		Physical Conditioning	E	High
		Self Control	C	High
		Sociability	E	Low
		Tolerance	O	Low
		Optimism	N	Low
Dunnette et al., 1962	Sales representative	Reasoning	—	Low
		Sales Effectiveness	—	High
		Assertiveness	E	High
		Cooperativeness	A	High

(*Appendices continue*)

Appendix B (*continued*)

Study	Occupation	Personality facet	Big five	Job relevance
		Conscientiousness	C	High
		Calmness	N	High
		Imaginativeness	O	Low
Griffith et al., 2008	Sales representative	Goal-Orientation	—	High
		Need For Control	E	High
		Social Confidence	E	High
		Social Drive	E	High
		Detail-Orientation	C	High
		Good Impression	—	High
		Need To Nurture	—	Low
		Skepticism	—	Low
Guan, 2015	Sales representative	Agreeableness	A	High
		Conscientiousness	C	High
		Extraversion	E	High
		Emotional Stability	N	High
		Openness	O	Low
Haaland, 2000	Customer service	Self-Management Competency	—	High
		Relationship Management Competency	—	High
		Stress Management Competency	—	High
Larson et al., 2013	Gardener	Agreeableness	A	Low
		Independence	—	Low
		Methodicalness	C	High
		Industriousness	C	High
		Extraversion	E	Low
		Dependability	C	High
Mahar et al., 2006	Accountant	Extraversion	E	Low
		Introversion	E	Low
		Sensing	O	Low
		Intuition	O	Low
		Thinking	A	Low
		Feeling	A	Low
		Judging	C	High
		Perceiving	C	High
Mahar et al., 1995	Psychiatric nurse	Extraversion-Introversion	E	High
		Sensation-Intuition	O	Low
		Thinking-Feeling	A	High
		Judging-Perceiving	C	High
Mudd, 2005	Police sergeant	Conscientiousness	C	High
Schwab, 1971	Top manager executive	Cautiousness	C	High
		Original Thinking	O	High
		Personal Relations	A	High
		Vigor	—	High

Note. N = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness. Numbers in parentheses indicate the sample number of each primary study.

Received September 9, 2017
Revision received March 12, 2019
Accepted March 18, 2019 ■