FAKE IT TO MAKE IT: A FURTHER EXAMINATION OF THE SUSCEPTIBILITY OF CONSTRUCT-LADEN SITUATIONAL JUDGMENT TESTS TO SOCIALLY DESIRABLE RESPONDING

A Thesis

by

FELIX GEORGE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Winfred Arthur, Jr.
Committee Members,	Stephanie C. Payne
	Murray R. Barrick
Head of Department,	Heather C. Lench

December 2020

Major Subject: Psychological Sciences

ABSTRACT

Faking on personality measures has been a concern for practitioners and academics alike due to the potential resultant reduction in the utility of personality assessments in operational contexts. Researchers have asserted that faking occurs to some extent on most Likert-based noncognitive measures, although the issue of which method of assessment is the most resilient to faking has yet to be settled. A promising option to mitigate concerns regarding faking is the situational judgment test (SJT). SJTs make use of a predetermined scoring key with correct and incorrect answers, making the issue of faking technically moot. Using a 2 (response instruction: faking vs. honest) × 2 (assessment method: SJT vs. single-statement) experimental design with response format nested within the SJT (rate vs. rank) and single-statement (Likert vs. true-false) methods, the comparative susceptibility of SJTs and single-statement measures to faking was examined. It was hypothesized that the SJT would be more resilient to attempts to "fake good" than single-statement measures of the same constructs when test-takers are given explicit instructions to fake. It was also hypothesized that the rank SJT response format would be more resilient to faking compared to the rate SJT format. In a sample of 583 participants recruited from Amazon Mechanical Turk (MTurk), the results indicated that across response formats, the standardized mean difference between the honest and faking conditions on tests of agreeableness and conscientiousness were significantly larger for the single-statement measures compared to the SJT. Both the rate and rank

SJTs were more resilient to faking than the Likert and true-false single-statement measures. In addition, the rank SJT was more resilient to faking than the rate SJT; however, this effect was obtained for the conscientiousness SJT but not agreeableness. Using supervisor-perspective job performance ratings, no support was found for any of the hypotheses positing that the SJT measures would have higher criterion-related validity than the single-statement measures under faking conditions. Overall, the results indicated that the SJTs were effective at preventing mean shifts due to faking, but the higher resiliency to faking did not result in higher criterion-related validity compared to the single-statement measures. The implications for the science and practice of SJTs, and faking on noncognitive measures, specifically personality, are discussed.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Arthur, and my committee members, Dr. Payne and Dr. Barrick, for their time and support throughout the process.

Thanks also goes to my wife, Jasmine George, daughter, Isabel George, as well as friends and colleagues for supporting me along the way.

Finally, thanks to my parents, Dorothy and Felix George, for their encouragement and love, which undoubtedly helped me to succeed.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by the student’s dissertation committee chair, Professor Winfred Arthur, Jr., with Professor Stephanie C. Payne of the Department of Psychological and Brain Sciences, and Professor Murray Barrick of the Department of Management as committee members.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

1. INTRODUCTION

The present study examined the extent to which a construct-focused situational judgment test (SJT) can mitigate concerns about faking on noncognitive measures (i.e., those for which there are no correct or incorrect answers to items). Historically, faking (also referred to as intentional response distortion) on personality measures has been a concern for practitioners and academics alike. Despite the ubiquity of self-report measures in organizational research and operational employment contexts, researchers have long held that response distortion represents potential harm to the value of these measures (Arnold, Feldman, & Purbhoo, 1985; Ellingson, Sackett, & Hough, 1999). In the context of self-report personality measures, faking in employment contexts has been defined as “a deliberate attempt to match one's own personality profile to one's perception of what management sees as the ideal personality for a specific job” (Martin, Bowen, & Hunt, 2002, p. 248). This definition not only indicates that faking is an intentional act, it also highlights the perspective of management, and in so doing, helps to establish the current study’s focus on faking in operational employment contexts.

It is important to contrast faking with lazy or careless responding. Central to the issues caused by careless responding is the notion that random responding or responding with a predetermined plan rather than making judgments related to the content of items, can threaten the utility of noncognitive measures (Nichols, Greene, & Schmolck, 1989). In this way, careless responding is defined as inattentiveness, arbitrary response patterns, and even an “unwillingness to comply with the testing demands” (Nichols et al., 1989, p.

240). Conversely, social desirability responding (SDR) is primarily concerned with attempts by test-takers to present themselves in a more favorable light.

Researchers have historically observed differences in personality test scores depending on the context, such that those completing personality tests in operational employment contexts have higher test means compared to incumbents. This is ostensibly due to deliberate changes in test-takers' response patterns, as individuals' responses are influenced by the level of importance they attribute to score outcomes. Thus, an objective indicator of social desirability responding is observed differences in personality test scores in low-stakes versus high-stakes testing situations. That scores shift depending on the context is rooted in the notion that there is some true score on a specified noncognitive trait, such that under low-stakes, participants' scores are closer to the true score. That is, when motivation to fake is low, test-takers do not engage in deliberate distortion, leading to lower test means. Conversely, in high-stakes situations, participants are motivated to respond with more socially desirable responses, resulting in higher test means.

The aforementioned observed difference in scores has been described using a variety of terms. Some have referred to this phenomenon as social desirability responding or intentional response distortion, on one hand, or faking, on the other hand. Regardless of the nomenclature used, these terms represent or describe the phenomena of observed increases in scores when stakes are high versus low. Notably, there has been much debate as to whether these observed score differences can be characterized as true or error variance. One school of thought is that this observed increase in scores is true variance, emerging as a result of the impression management behaviors associated with well-adjusted individuals (Hogan, Barrett, & Hogan, 2007). An alternative view is that it represents error variance (Reeve, Heggestad, & George, 2005; Stewart, Darnold, Zimmerman, Parks, & Dustin, 2010), and thus, should be mitigated or controlled.

Given the above, and the concerns associated with score increases in high-stakes testing, the prevailing view is that it is a problem, particularly a measurement problem. Under these conditions, the objective of the present paper is to explore the viability of SJTs as a means of addressing this problem.

Is there support for the notion that faking is indeed a concern in operational employment contexts? Critically, in 2002, 68% of members surveyed from the Society for Human Resource Management thought faking concerns rendered self-report integrity tests useless (Rynes, Colbert, & Brown, 2002). In spite of this perception, organizations have continued to use personality tests, if not increasingly so. Consequently, there has remained an interest in determining the extent to which faking is a threat to selection programs. Indeed, faking on self-report measures has the potential to reduce the utility of personality assessments in operational contexts. On the Big Five facet-level scales, Rosse, Stecher, Miller, and Levin (1998) demonstrated that applicant faking in the field was comparable to levels found in directed faking studies, emphasizing that faking is a phenomenon that is not only observed in research settings but one that is also prevalent in organizational settings. Faking was also more prominent among applicants than incumbents (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006), furthering the idea that faking can compromise selection outcomes.

In addition to determining whether faking occurs (e.g., Griffith, Chmielowski, & Yoshita, 2007), past research has focused on determining the extent to which specific testing formats are susceptible to faking. To this end, researchers have confirmed that faking occurs to some extent on most Likert-based personality measures. In fact, some have asserted that "all [noncognitive] tests are fakable" (Snell, Sydell, & Lueke, 1999, p. 223). Empirical investigations that compare mean differences between applicants and non-applicants on personality dimensions provide additional evidence that response distortion takes place (Birkeland et al., 2006; Weekley, Ployhart, & Harold, 2003). It is worth noting that overall, score inflation is greater when participants are given explicit instructions to fake (i.e., directed faking study designs) than in applicant-incumbent designs (Cao & Drasgow, 2019). In applicant-incumbent designs, faking is inferred from the higher scores of applicants relative to incumbents (Barrick & Mount, 1996; Birkeland et al., 2006; Griffith et al., 2007; Schmit & Ryan, 1993).

Furthermore, rank-order changes in means across personality dimensions depend on the job context, suggesting that applicants distort their responses to appear more suitable for specific positions. The extent to which faking improves scores has also been quantified. Ones, Viswesvaran, and Korbin (1995) demonstrated that instructions to fake good on personality tests resulted in score increases of almost half a standard deviation. However, score increases as high as 1.50 standard deviations have also been reported (Drasgow, Chernyshenko, & Stark, 2010; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Viswesvaran and Ones' (1999) meta-analysis demonstrated significant mean differences between honest and fake good conditions (*d* = 1.06). Similarly, Hurd, Barrett, Miguel, Tan, and Lueke (2001) obtained significantly higher social desirability scores for those instructed to fake good.

Detecting and eliminating faking is of great concern due to the potential impact faking has on test scores and selection programs. These concerns include rank-order changes, differential item functioning, decreased construct-related validity, and other threats to the psychometric properties of personality measures. Fluckinger, McDaniel, and Whetzel (2008) have asserted that “attempts to fake can show up in a number of statistical indicators, including test means … criterion-related validity, actual or simulated hiring decisions, and construct validity” (p. 92).

2. IMPACT OF FAKING

Psychometric Properties and Validity

Donovan, Dwight, and Schneider's (2014) investigation of faking included administrations of personality tests at the time of a job application and again five months after being hired. The results revealed that faking was indeed prevalent, and fakers performed worse on the job compared to non-fakers. In addition, there was evidence that the psychometric properties—both internal consistency and factor structure—of the personality measures were compromised by faking. However, Viswesvaran and Ones (1999) warn against overstating the impact of faking on selection programs, asserting that "effect sizes reflect neither changes in rank ordering of individuals nor distortion in linear relationship with other measures (i.e., correlations). Effect sizes merely indicate that the mean response changes according to the instruction provided" (p. 205).

Nonetheless, there is evidence that faking can have negative effects on construct-related validity. Specifically, Schmit and Ryan (1993) found that when respondents attempted to appear as the ideal employee, the factor structure of Big Five personality measures was compromised. This "ideal-employee" factor has been attributed to respondents' attempts to maximize their chance of being selected (Rosse et al., 1998). Whereas early empirical work revealed that the validities of self-report noncognitive measures remained relatively stable despite intentional response distortion (Hough et al., 1990), Douglas, McDaniel, and Snell (1996) found that under faking conditions, criterion-related and construct-related validity decreased. In fact, there is empirical evidence (e.g., Cellar, Miller, Doverspike, & Klawsky, 1996; Fluckinger et al., 2008;

Zickar & Robie, 1999) of reduced construct-related validity when applicants fake. Specifically, factor analysis reveals that an additional factor emerges that is distinct from the focal factor of interest (i.e., Big Five). Multi-trait multi-method analyses have also revealed similar findings (Douglas et al., 1996).

There is also evidence that the correlational structure of Big Five measures is compromised when test-takers are given explicit instructions to fake. Specifically, measures are positively correlated under fake good instructions (Paulhus, Bruce, & Trapnell, 1995), with the intercorrelations among big five traits being higher when participants are instructed to fake.

Griffin, Hesketh, and Grayson (2004) revealed differential item functioning in items measuring conscientiousness and openness to experience in a comparison of applicant and student responses. That is, applicants with the same trait score performed differently on items due to differences in an unassessed underlying trait. Reeve et al. (2005) also demonstrated that faked responses are associated with an increase in transient error (i.e., variance across test occasions attributable to unassessed factors) compared to honest responses.

There are also concerns regarding the influence of cognitive ability on respondents' ability to fake on measures of noncognitive constructs. Empirical evidence (Kasten, Freund, & Staufenbiel, 2020) supports the "smart faker" hypothesis, or the notion that cognitive ability is associated with the ability to fake on noncognitive measures. Specifically, faking ability appears to be more strongly related to respondents' comprehension-knowledge (Gc) than fluid reasoning (Gf; MacCann, 2013). Cognitive ability was also explicitly linked to the ability to fake in empirical work by Pauls and Crost (2005). Specifically, cognitive ability was positively associated with the amount of faking as well as profile-specific response distortion patterns. That is, those with higher cognitive ability were better able to fake a specific personality profile. Thus, efforts to reduce faking should have the beneficial effect of neutralizing the association between cognitive ability and scores on personality measures.

Rank-Order Changes

Although Barrick and Mount (1996) demonstrated that impression management and self-deception did not affect the predictive validity of personality measures of emotional stability and conscientiousness, there are other harmful effects of faking to be considered, particularly rank-order changes. For instance, in a top-down selection context, rank-order changes at the individual level make it likely that some individuals who are faking-good will be selected who otherwise would have been screened out. That is, those with distorted scores can improve their chances over those who respond honestly. Members comprising the top of the applicant pool change as a result of individual differences in the extent to which individuals exhibit response distortion (Morgeson et al., 2007). According to Morgeson et al. (2007) "this research has shown that faking will be more problematic as selection ratios decrease and if top-down selection is used. That is, different people will be hired due to faking" (p. 686). Griffith et al. (2007) demonstrated extreme rank-order changes after administering personality tests to applicants and obtaining honest scores one month later. Some participants' rank-order dropped as much as 48 positions in the honest condition. This is not surprising as

Paulhus and Bruce (1991) asserted that applicants are able to fake good given information on the specific job application to be simulated (cf. Hogan et al., 2007). Furthermore, Winkelspecht, Lewis, and Thomas (2006) demonstrated that those who fake more are likely to appear near the top of the score distribution. In a selection context, “individuals would be selected solely because they elevated their scores not only more than others but also more than can be explained by measurement error” (Stewart et al., 2010, p. 628).

3. DETECTING AND CORRECTING FOR FAKING

To mitigate concerns over faking, researchers have addressed faking with a variety of detection and correction methods—a sampling of which is reviewed here—that can be implemented during or after the administration of noncognitive measures.

Blatant Extreme Responding

Pertaining to detection, blatant extreme responding is one of several techniques used by researchers, and it is typically calculated as the proportion of items endorsed with a 1 or 5 on a 1 to 5 scale (Levashina, Weekley, Roulin, & Hauck, 2014). A primary drawback associated with this approach is that those responding honestly may still endorse a high proportion of items at the extreme ends, potentially leading to false positives. Also, there has been no empirical evidence to date that this method improves the amount of explained variance of personality-performance associations. This is an issue, as extreme responding is ubiquitous and can occur in both research and organizational contexts.

Social Desirability Scales for Detection and Correction

A frequently-used method within the domain of detection and correction techniques involves the administration of social desirability scales. Examples of these scales are the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1988, 1991) and Marlowe–Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960) with the former providing separate measures of impression management (i.e., intentional inflation of self-descriptions) and self-deceptive enhancement (i.e., unconscious bias in self-descriptions). Instead of detecting faking in real time, these scales are used to correct scores after the administration of noncognitive measures. However, this technique is not without drawbacks. Ones, Viswesvaran, and Reiss (1996) demonstrated that social desirability scales fail to add meaningful variance when predicting performance from personality measures. Schmitt and Oswald (2006) similarly found that applying corrections based on scores from social desirability scales does not appear to improve predictive validity. Stark, Chernyshenko, and Drasgow (2005) assert that "efforts to correct for score inflation and changes in the rank ordering of respondents post hoc have been generally ineffective" (p. 184). Ellingson et al. (1999) assert that "social desirability correction is ineffective and fails to produce a corrected score that approximates an honest score" (p. 155). According to Morgeson et al. (2007), for corrections to be effective, "the faking measure that you are using has to be correlated with the outcome, the predictor, or both. And, in most cases, they do not or those correlations are relatively small" (p. 709).

An additional issue with social desirability scales is that applicants may employ a faking strategy that is specific to the job of interest, making faking far more difficult to detect and address (Mahar, Cologon, & Duck, 1995). Finally, there is the concern that partialling lie scale scores from noncognitive test scores could potentially depress the rankings of honest individuals (Stewart et al., 2010). In line with Cronbach (1990), once lies are told in the testing context, it appears that the harm cannot be undone.

There is utility in framing the issue of social desirability responding under a socioanalytic perspective, highlighting that personality and behavior are generally directed towards the aims of getting along (i.e., social acceptance) and getting ahead (i.e., individual achievement; Hogan & Holland, 2003). It follows, then, and has been asserted by others (e.g., Ones et al., 1996; Uziel, 2010), that the partialling of lie scores may involve the removal of true variance that could be relevant to the personality dimension of interest (e.g., conscientiousness, agreeableness). In line with this view, faking can be misconstrued as aberrant behavior that leads to measurement error when it is actually a central aspect of socialized behavior (i.e., the impression management behaviors associated with well-adjusted individuals). According to Hogan et al. (2007), "the larger point here is that it is almost impossible to distinguish faking from socialized behavior. And this means that it is very hard to assign a clear meaning to the claim that some people fake when they respond to personality measures" (p. 1282). In line with the views of Hogan et al. (2007) and Barrick and Mount (1996), Hough and Oswald (2000) assert that "distortion does not tend to moderate, mediate, suppress, or attenuate the criterion-related validities of personality scales" (p. 634). Proponents of this view call for faking to be investigated under the larger umbrella of self-presentation and establishing reputation rather than studied in isolation (Morgeson et al., 2007). Moreover, Smith and Ellingson (2002) found empirical support for the notion that in real-world contexts (e.g., job application scenarios), response distortion does little to impact the construct validity of personality measures.

The debate about the need to detect and correct for faking has been ongoing for quite some time and remains largely unsettled. Nonetheless, efforts to address faking continue to be made, and the perspective adopted herein is that addressing testing procedures or test characteristics that could prevent faking before it occurs appears to be a much more fruitful endeavor than detection or correction.

4. PREVENTING FAKING

As noted above, it has been proposed that instead of detecting faking and correcting scores post hoc, a more effective strategy would be to prevent faking altogether. Early empirical work by Hough et al. (1990) detailed a strategy for preventing faking with the use of subtle items that assessed constructs that are not apparent to the test-taker. In addition, more recent empirical work comparing ipsative and normative versions of personality scales has revealed that ipsative measures are more resistant to faking (Bowen, Martin, Caroll, & Hunt, 2002). To this end, researchers have used a variety of techniques, most notably forced-choice formats. Before detailing why situational judgment tests (SJTs) may be a more effective approach, the various alternatives are examined.

Overt and Covert Tests

Early work by Alliger, Lilienfeld, and Mitchell (1996) explored the susceptibility of overt and covert tests to faking using explicit instructions to fake good. On a covert test, the focal construct being assessed is not apparent to the test-taker, while the construct is more obvious to respondents on overt tests. The authors demonstrated that a covert integrity test was resilient to faking and coaching while performance on the overt test could be successfully coached.

Forced-Choice Format

Both the U.S. Army’s Assessment of Individual Motivation (AIM) and the Navy’s Navy Computer Adaptive Personality System (NCAPS) use forced-choice tests in an effort to reduce faking in high-stakes testing contexts (Hough & Oswald, 2008).

Unidimensional forced-choice tests are designed to force respondents to choose from amongst a fixed set of desirable options corresponding to a single construct (e.g., conscientiousness). Multidimensional forced-choice measures build on this by using multidimensional statements. That is, researchers typically use pairs (multidimensional pairwise preference items; MDPP)—although triads or quads may also be used—of response items with similar levels of social desirability; each option represents a different dimension (Stark et al., 2005). Hybrid items such as these, with two or more constructs used as response options, tend to correlate (.72 - .82) with unidimensional measures but are less susceptible to response distortion (Bernal, 1998). The tailored Adaptive Personality Assessment System (TAPAS), a forced-choice personality test developed by Drasgow Consulting Group (Stark et al., 2014), is a specific example of a forced-choice test designed for use in a high-stakes testing environment. Created for and used by the United States military, TAPAS uses MDPP items in an effort to mitigate the threat of faking when conducting personality assessments of military personnel.

Overall, forced-choice scales are considered to be less fakeable and have had early promise (e.g., Christiansen, Edelstein, & Fleming, 1998). Meta-analytic results demonstrated that forced-choice formats have lower standardized mean differences between honest and fake good conditions than Likert scales (Stanush, 1997). More recent empirical evidence by Cao and Drasgow (2019) comparing low-stakes and high-stakes testing using forced-choice scales demonstrated a *d* of 0.06. The magnitude of this effect, particularly when compared to single-statement personality measures, provides evidence that the forced-choice format stands as one of the most effective methods of preventing faking. Empirical work by Jackson, Wroblewski, and Ashton (2000) demonstrated that the forced-choice format was less susceptible to response distortion than a single-stimulus format. Specifically, respondents were asked to complete measures frankly and again while making a good impression for a job application. The mean shift for the forced-choice format was one-third of the mean shift for the single-stimulus format. NCAPS (Houston, Borman, Farmer, & Bearden, 2006), a testing system similar in design and function to TAPAS, was employed by Underhill, Bearden, and Chen (2008) with some success, providing further evidence that pairing the forced-choice format with the computer adaptive testing (CAT) format further reduces susceptibility to faking.

Problems inherent in the forced-choice format include the lack of interval-level scaling. Although approaches based on item response theory (IRT) methods (Heggestad, Morrison, Reeve, & McCloy, 2006) have been used to overcome this shortcoming, both CAT and IRT introduce monetary and experience requirements which create additional burdens not inherent in other approaches. Also, empirical evidence regarding its benefits have been inconclusive. In fact, some studies have demonstrated that forced-choice scales can exhibit higher and lower means than Likert scales (Fluckinger et al., 2008).

Implicit Association Tests

Mixed evidence exists for the notion that implicit association tests are viable alternatives to other self-report measures when response distortion is a concern. In addition, unique concerns exist with implicit association tests, most notably that they have weak correlations with explicit measures of the same personality traits (Vecchione, Dentale, Alessandri, & Barbaranelli, 2014). While a meta-analysis by Greenwald, Poehlman, Uhlmann, and Banaji (2009) found that implicit association tests had stronger predictive validity than self-report measures, Oswald, Mitchell, Blanton, Jaccard, and Tetlock's (2015) meta-analysis demonstrated that implicit association tests were poor predictors in a variety of criterion categories, including interpersonal behavior.

Option-Keying

Different keying or answer key creation procedures have also been examined. According to Kluger, Reilly, and Russell (1991), "with an option-keying (OK) strategy, each item response option (alternative) is analyzed separately and contributes to the score only if it correlates significantly with the criterion" (p. 890). Thus, option-keying has been used to score item response options based on their relationship with performance on a criterion (Snell et al., 1999). The primary drawback of option-keying is the lack of a conceptual or theoretical basis to support the procedure. Recent empirical evidence by Cucina et al. (2018) found that option keying and item-level empirical keying positively impacted criterion-related validities. Furthermore, empirical keying reduced the impact of faking.

Elaboration

Elaboration or written self-reports of behavior corresponding to a trait of interest has also been proposed as a method to reduce faking and remains a viable option. Test designers have combined traditional item formats (e.g., biodata items) with elaboration with inconclusive results. Schmitt et al. (2003) demonstrated that elaboration had no effect on the correlations between a biodata measure and a social desirability measure. In contrast, Lievens, Peeters, and Schollaert (2008) found that elaboration, when combined with SJT items, effectively reduced the number of fakers at the top of the score distribution.

Ideal Point Model

Another alternative to traditional scoring procedures is the ideal point model. In contrast to scoring methods commonly used for Likert measures (i.e., dominance model) in which positive and negative items are summed, the ideal point model contains positive, negative, and intermediate options. Unlike traditional Likert measures, reverse scoring is prohibited, and trait scores consist of “the mean item location of the items endorsed or the ideal point trait estimate” (Drasgow et al., 2010, p. 470). Proponents of the ideal point model assert that items may represent an intermediate level of the trait of interest, leading to a lack of endorsement by those very low or very high on the trait. The summing process in dominance models obscures intermediate levels of traits and fails to differentiate between those trait levels that fall between the extreme low and high ranges. Ideal point models can be used in conjunction with the forced-choice item format. Researchers have asserted that because the most socially desirable item responses are not transparent to the test user, items are less susceptible to faking. The multidimensional pairwise preference model (MUPPM; Stark, 2002; Stark et al., 2005) is one specific application of the ideal point model and forced-choice format that has been used with success. A drawback of ideal point models is the requirement of large sample sizes which are “needed to estimate the item parameters of ideal point models” (Drasgow et al., 2010, p. 519).

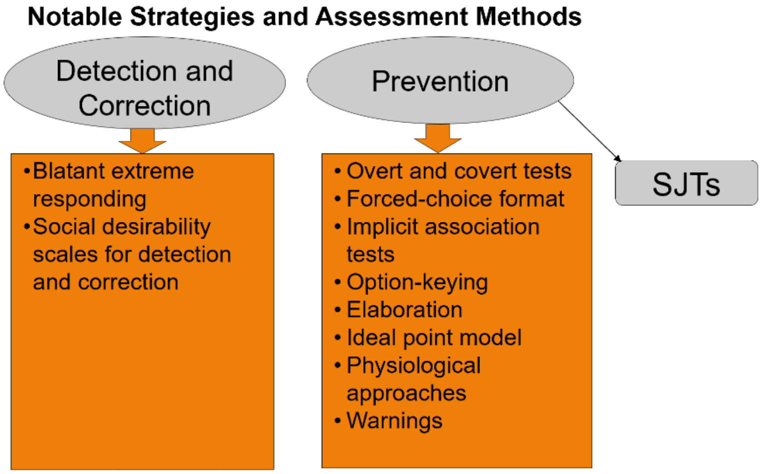
Physiological Approaches

There has also been scholarly interest in using physiological data to assess personality, which can be considered a passive assessment method in that objective biological/physiological data are collected from individuals, and these data are subsequently used to make inferences about personality. Critically, because physiological measures obviate the need for self-reported information, they avoid the issue of socially desirable responding altogether. Empirical work by DeYoung et al. (2010) detailed an investigation of the neurological correlates of personality. Specifically, the authors used functional magnetic resonance imaging (fMRI) to correlate the volume of specific brain regions with scores on the Big Five personality dimensions. Finding support for the biological model of personality, specified brain regions were significantly correlated with specified personality dimensions for four out of five of the personality dimensions analyzed. However, in a selection context, cost and privacy concerns render physiological measures impractical.

Warnings

Simple warnings of the consequences of distorting responses have been used on personality measures (Hough et al., 1990). Warnings have often included notifications to test-takers that faked responses can and will be verified. There is some evidence of success as meta-analytic results (Dwight & Donovan, 2003) have demonstrated that warnings lead to lower test means on noncognitive measures. In contrast, some investigations have found small effects of warnings (e.g., Robson, Jones, & Abraham, 2007). Therefore, there is no conclusive evidence that warnings are a meaningful way to reduce faking.

Novel procedures that attempt to actively identify potential fakers and provide warnings during test administration have also been proposed by Landers, Sackett, and Tuzinski (2011) and Fan et al. (2012). The test-warning-retest procedure calls for potential fakers (i.e., those identified due to their extreme responding) to retest or re-attempt previously completed items. While empirical results have been promising, a potential issue is false positives. When honest responders are misidentified as potential fakers, they may then distort their scores during a retest to appear less extreme. Similar to corrections based on social desirability scale scores, honest responders may ultimately be penalized despite making no attempt to fake. Figure 1 presents an illustrative summary of the strategies that have been reviewed in the preceding sections.


Notable Strategies and Assessment Methods

Detection and Correction

- Blatant extreme responding
- Social desirability scales for detection and correction

Prevention

- Overt and covert tests
- Forced-choice format
- Implicit association tests
- Option-keying
- Elaboration
- Ideal point model
- Physiological approaches
- Warnings

SJTs

Figure 1. Notable strategies for prevention, detection, and correction of faking on personality measures. SJTs = situational judgment tests.

5. PRESENT STUDY

Situational Judgment Tests

Snell et al. (1999) proposed a model of applicant faking that includes factors affecting applicants' ability to fake (dispositional factors, experiential factors, and test characteristics) and motivation to fake (demographic factors, dispositional factors, and perceptual factors). Of the various factors, testing method and test characteristics remain the elements that are the most controllable by those involved in the assessment of noncognitive factors. Despite a multitude of available options regarding the measurement of personality dimensions, the issue of which method is the most resistant to faking and practical has yet to be settled. While computer-adaptive, forced-choice tests hold considerable promise, there are several drawbacks preventing their widespread use, including significant overhead in terms of the required financial resources and required expertise of testing personnel for effective implementation. Computer-adaptive, forced-choice testing data analyzed using IRT methods also have the added requirement of requiring large sample sizes (de la Torre & Hong, 2010).

Another option, and the focus of the present study is the SJT, a testing method that has received empirical support as a promising alternative to the previously mentioned methods. SJTs have traditionally been defined as low-fidelity simulations of work roles and situations (Motowidlo, Dunnette, & Carter, 1990) and there has been continued interest in their use in high-stakes selection contexts. Research by Arthur et al. (2014), Arthur (2017a), and Kasten et al. (2020) suggests that SJTs compare favorably to other methods that suffer from high susceptibility to faking (e.g., Likert scales), undesirable consequences on psychometric properties (e.g., corrections based on social desirability scale scores), high costs (e.g., fMRI-based methods, computer-adaptive testing), and uncertainties regarding resistance to faking (e.g., forced-choice, subtle items). SJTs may help alleviate many of the concerns of using personality tests in high-stakes assessment situations, and by so doing, potentially increase the criterion-related validity of personality scores. In fact, Fluckinger et al. (2008) have boldly proclaimed that SJTs "with knowledge instructions appear to be the sole available method of assessing personality in a manner that is faking resistant" (p. 103). Empirical evidence has supported this assessment (Nguyen, Biderman, & McDaniel, 2005). However, before assessing the susceptibility of SJTs to faking, a discussion of their utility as a method of measuring unidimensional personality traits is germane.

Although SJTs were originally viewed as work samples or multi-dimensional tests of procedural and work-related knowledge, they have since evolved for use as a measure of unidimensional traits or constructs. And thus, it is unsurprising that the validity of SJTs varies as a function of the construct being measured (Christian, Edwards, & Bradley, 2010). Researchers have previously highlighted the need to make a clear distinction between the method and construct being measured and this is particularly germane to the study of SJTs (e.g., Arthur & Villado, 2008). Historically, SJTs, like assessment centers, have been characterized by their multidimensionality (Ployhart, 2006), and empirical debates ensued regarding whether SJT scores, regardless of the focal construct being tested, are capturing global, cognitive skills (e.g., tacit knowledge, practical intelligence, procedural knowledge; Sternberg, Wagner, Williams,

& Horvath, 1995). Since then, a construct-focused approach to SJT development has resulted in the creation of construct-laden SJTs. By utilizing theory, specifying the construct domain, and linking critical behaviors to relevant situations, SJTs *can* be constructed to be unidimensional.

Christian et al. (2010) detailed the benefit of a construct-focused approach, including reducing contamination and construct-irrelevant variance. Construct-laden SJTs of integrity (Arthur et al., 2014), personal initiative (Bledow & Frese, 2009) and prosocial implicit trait policy (Motowidlo, Ghosh, Mendoza, Buchanan, & Lerma, 2016) represent successful attempts to create construct-laden SJTs. Additional evidence includes an SJT to measure the HEXACO personality dimensions (Ashton & Lee, 2007) by Oostrom, de Vries, and de Wit (2019), which demonstrated adequate construct-related validity and criterion-related validity. Becker (2005) developed and validated an SJT measure of integrity, finding that the SJT predicted a range of employee outcomes (e.g., job performance, career progress). Similarly, de Meijer, Born, van Zielst, and van der Molen (2010) detailed the development of a video-based SJT of integrity, which was used in a law enforcement setting. The authors found evidence of construct-related validity and no significant subgroup differences between different ethnic groups. And Teng, Brannick, and Borman (2020) recently published an SJT measuring resilience.

Advantages. With regard to validity, SJTs compare favorably with other methods, as meta-analytic evidence has demonstrated a mean criterion-related validity of

.26[1] (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Unlike other self-report methods, SJTs have a scoring approach that is particularly relevant to the issue of faking. That is, the key differentiating feature of SJTs in comparison to Likert scales and other self-report personality measures is the use of a predetermined scoring key. Scoring keys may be created by relying on (a) expert-based scoring (i.e., subject matter experts arrive at a consensus on correct answers), (b) theoretical scoring (i.e., correct answers are informed by relevant theory), (c) empirical scoring (i.e., answers are selected based on interrelationships with other measures), (d) normative scoring (i.e., answers are selected based on the most commonly selected responses of test-takers), and (e) hybridized scoring (i.e., combinations of other scoring strategies; Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). It is worth noting that with empirical scoring—scoring in which the answer key is determined by correlating response options with a criterion of interest (e.g., task performance)—the correct answer becomes even less apparent to test-takers.

Whether a rank (rank ordering each option from best to worst) or rate (rating each option individually) response format is used, each response contains a dedicated right answer. Given the test characteristics, response options, and the use of scoring keys of SJTs, it is not surprising that researchers have begun to investigate whether construct-laden SJTs can be faked (e.g., Arthur, 2017a; Arthur et al., 2014; Kasten et al., 2020).

[1] The meta-analytic effect size was collapsed across constructs.

To the extent that the characteristics of a construct-laden SJT make it less susceptible to faking, scores on the SJT should correlate weakly with an independent measure of SDR. Comparatively, a Likert scale measuring the same construct would be expected to have significantly larger correlations with a measure of SDR. This is exactly what Arthur (2017a) found using SJTs of agreeableness and conscientiousness with a sample of 692 participants. Specifically, Arthur (2017a) observed an SJT-SDR correlation of .07 for agreeableness and .01 for conscientiousness. Comparatively, a 10-item International Personality Item Pool Likert measure of the same constructs had correlations of .20 and .28 with the SDR measure. Furthermore, the factor structure of the SJTs supported their use as unidimensional personality measures. These findings serve as a foundation for the current study, as they demonstrate that not only can SJTs effectively serve as measures of unidimensional personality constructs, but they can do so without substantial influence from social desirability responding.

It is important to note that efforts to extend Arthur's (2017a) findings call for a stronger test of the susceptibility of SJTs to faking. That is, while correlations with SDR measures serve as an important metric when examining the susceptibility of a testing method to faking, a more rigorous approach is to examine these effects in situations where test-takers are motivated to either respond honestly or to engage in faking. One way of doing so is to test job applicants before they are hired and again after they begin to work in their new position to examine score differences. Presumably, due to their need to present themselves in the best possible light, job applicants would be more motivated to fake than incumbents, resulting in mean differences between the two conditions (e.g., Tsaousis & Nikolaou, 2001; Weekley et al., 2003; Zickar, Gibby, & Robie, 2004). In lieu of this, an alternative, and maybe even more rigorous approach is to have two distinct test administration procedures whereby participants are explicitly instructed to either respond honestly or fake good. This approach is the one used in the current study. Recent empirical work by Kasten et al. (2020) also used this technique with promising results. Indeed, because the current study was in the design phase when their study was published, this presents the opportunity to replicate and extend Kasten et al.'s (2020) study. Consequently, their findings are highlighted below while discussing how the present study aims to address additional research questions not addressed in their study.

As previously noted, while the susceptibility of SJTs to faking was investigated by Arthur et al. (2014) and Arthur (2017a), study designs that use explicit instructions to "fake good" can be considered a stronger approach to assessing a testing method's susceptibility to faking. Using two independent samples, Kasten et al. (2020) employed a study design with honest responding and "fake good" conditions. Specifically, participants "were asked to present themselves in a favorable light in order to maximize their chances of being hired" (Kasten et al., 2020, p. 139). In the honest condition, participants received standard instructions. In Study 1, participants were tested in both conditions (i.e., a within-subjects design). In Study 2, participants were assigned to either the honest condition or faking condition (i.e., a between-subjects design).

In Study 1, using difference scores from the honest and faking conditions to quantify the extent to which participants were able to fake good, Kasten et al. (2020) found that an

SJT measure of conscientiousness had smaller differences (*d* = 0.78) than a Likert scale (*d* = 1.71) measuring the same construct. Results for Study 2 were similar, as the scores for the faking group were significantly higher than those for the honest group for both the SJT (*d* = 0.66) and Likert-based measure (*d* = 1.14). However, the score differences for the Likert measure were significantly higher than those for the SJT group, indicating that the SJT was less susceptible to instructions to fake good.

One important consideration regarding Kasten et al.'s (2020) study design was their decision to use both a within- and between-subjects design with independent samples. While the present study is replicating their study with respect to faking instructions (honest and fake good) and testing method (SJT and Likert), the choice of study design warrants further consideration for the current study. Indeed, the impact of faking in personality assessment appears at least partially dependent on whether a within- or between-subjects design is used. In a meta-analysis, Ones et al. (1995) demonstrated that between-subjects *d*s ranged from 0.48 to 0.65 while within-subject *d*s ranged from 0.47 to 0.93 when comparing honest and fake good conditions. Although Viswesvaran and Ones (1999) recommend within-subjects designs for more realistic estimates, there is empirical evidence indicating that the ordering of honest and faking conditions in within-subjects designs also impacts the observed effects. For example, Nguyen et al. (2005) obtained a *d* of 0.34 when the honest condition was ordered first compared to 0.15 when the faking condition was first.

So, to eliminate the effect of the ordering of measures inherent in a within-subject design, a between-subjects design is used with a Likert condition and SJT condition (see Figure 2 for a graphical overview). Overall, consonant with the extant literature and as a replication of past research (e.g., Kasten et al., 2020; Winkelspecht et al., 2006), score differences between fake good and honest conditions in the Likert group relative to the SJT group were expected.

Hypothesis 1: There will be smaller mean differences between the "fake good" and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness compared to the single-statement measure of the same constructs.

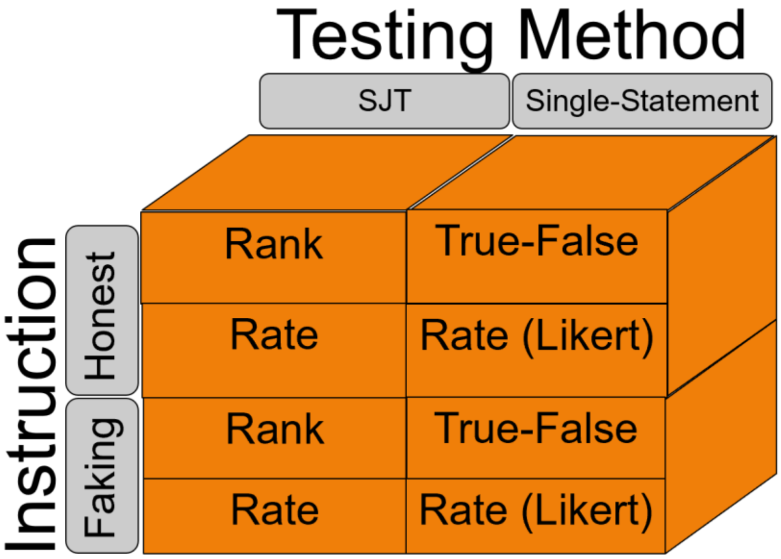


Figure 2. Study design: 2 (response instruction: faking vs. honest) × 2 (assessment method: SJT vs. single-statement) between-subjects with response format nested within testing method (SJT: rate vs. rank; Single-statement: rate [Likert] vs. true-false). SJT = situational judgment test.

Response format. As Ziegler and Buehner (2009) assert, "faking can be understood as a systematic measurement error resulting from the interaction between context (situational demand) and person" (p. 550). It follows that uncovering the SJT characteristics that make them more resistant to faking would lead to improved accuracy and effectiveness of organizational hiring programs. Previous research has been inconclusive regarding the extent to which different response formats translate into differences in susceptibility to faking. Nonetheless, response formats requiring test-takers to rate each option (rate) or rank each option from best to worst (rank) may be differentially susceptible to faking. In fact, there is empirical evidence that when holding the content of the SJT constant, SJTs differing in only response format demonstrate different criterion-related validities (Rasmussen, 2009). It follows, then, that a significant gap in the research literature is examining whether susceptibility to faking is substantially related to these criterion-related validity differences. That is, does the response format impact susceptibility to faking, which in turn impacts criterion-related validity?

It was previously noted that the forced-choice response format has well-documented potential, and some empirical success, with respect to reducing faking. It follows that the rank response format of the SJT, which forces participants to rank alternatives ordinally with no option for ties (i.e., no two options may have the same rank), should function similarly to a forced-choice response format. Thus, test-takers should be expected to find rank SJTs more difficult to fake than a Likert measure. Specifically, because the rank format is functionally similar to the forced-choice format, it is expected to have decreased susceptibility to faking compared to the rate format. To this end, the susceptibility to faking between the two most extreme response formats, the rank and the rate (which are nested within the SJT condition) are compared. Furthermore, researchers have previously compared forced-choice formats to true-false formats (e.g., Jackson, Neill, & Bevan, 1973), demonstrating that a true-false format compares favorably to the forced-choice with respect to convergent and discriminant validity. Early research on the true-false format has shown that when explicit instructions to fake are given, the magnitude of faking effects are not substantial (Braun & Costantini, 1970; Hoffmann & Nelson, 1971). Thus, it is expected that although a Likert scale using a true-false response format should be less susceptible to faking than a Likert scale using a traditional rate response format, an SJT using either a rate or rank response format should outperform both in terms of resiliency to faking.

> ***Hypothesis 2:*** There will be smaller mean differences between the "fake good" and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format compared to the single-statement measure of the same constructs using a Likert response format.
>
> ***Hypothesis 3:*** There will be smaller mean differences between the "fake good" and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format compared to the single-statement measure of the same constructs using a true-false response format.
>
> ***Hypothesis 4:*** There will be smaller mean differences between the "fake good" and honest conditions of the construct-laden SJT measure of (a) agreeableness

and (b) conscientiousness using a rank format compared to the construct-laden SJT measure of agreeableness and conscientiousness using a rate format.

Criterion-related validity. Although there is empirical evidence of SJTs' decreased susceptibility to faking compared to Likert measures, the criterion-related validity of SJTs could still be impacted by faking. In fact, this is what Peeters and Lievens (2005) demonstrated, finding that faking decreased the criterion-related validity of SJTs in an educational context. Specifically, psychology students were given explicit instructions to fake good or respond honestly on an SJT measuring various traits related to student performance (e.g., teamwork, communication, study habits). Not only did students in the "fake good" condition achieve higher scores, but correlations between the SJT and actual performance were lower for those in the faking condition (r = .09) compared to the honest condition (r = .33). Although the authors did not use a unidimensional SJT as is the case in the present study, their results are informative. That is, their results provide additional evidence that in directed faking studies, substantial differences between honest and faking conditions are associated with differences in criterion-related validity. It follows, then, that if a Likert measure of agreeableness and conscientiousness is more susceptible to faking than an SJT measuring the same constructs, the SJT will serve as a more valid predictor of performance.

The present study obtains criterion data to examine the comparative criterion-related validity of SJT and Likert measures of the same constructs using supervisor-perspective job performance ratings. Considerations regarding the accuracy of self-report measures are pertinent to criterion selection. On the one hand, there is empirical evidence that when individuals assume the perspective of their supervisors, such ratings of job performance are more strongly correlated with supervisor-ratings than traditional self-reported ratings (e.g., Schoorman & Mayer, 2008). On the other hand, however, Cho, Payne, Berry, & Lee, (2020) found that supervisor-perspective ratings of job performance are not a valid substitute for supervisor ratings given the low correlations between these two sources of job performance data (*r* = .34). However, because the present study did not have access to supervisor ratings, the choice of criteria was between either traditional self-ratings or supervisor-perspective ratings of job performance. Thus, for the present study, supervisor-perspective ratings of job performance are a feasible proxy of supervisor ratings and are used as the criterion. Specifically, it is expected that when respondents are explicitly instructed to fake, criterion-related validities will be higher in the SJT condition than the Likert condition.

> ***Hypothesis 5:*** When respondents are explicitly instructed to fake, the criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.

> ***Hypothesis 6:*** When respondents are explicitly instructed to fake, the criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format will be higher than the criterion-related

validity for the single-statement measure of the same constructs using a true-false response format.

Hypothesis 7: When respondents are explicitly instructed to fake, criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a rank format will be higher than the criterion-related validity for the SJT using a rate response format.

Cognitive ability. There is a vast literature, as well as meta-analytic evidence, detailing the positive association between cognitive ability and performance on SJTs (McDaniel, Hartman, Whetzel, & Grubb, 2007). For example, McDaniel et al.'s meta-analysis (2001) demonstrated that SJTs had correlations of .39[2] with general mental ability. Also, Kasten et al. (2020) found that the extent to which test-takers faked was positively associated with general mental ability. Although much of the research on SJTs and GMA does not make a distinction between the method and constructs assessed, it is widely accepted that the extent to which SJT performance is associated with cognitive ability is nontrivial (Whetzel, McDaniel, & Nguyen, 2008) because the completion of SJTs is more cognitively demanding than Likert measures. Consonant with this, higher correlations exist between SJT performance and GMA for SJTs that use the rank response format, regardless of SJT content (Arthur et al., 2014). Nonetheless, there appears to be no empirical evidence that the association between SJT scores and GMA scores suppresses or otherwise hinders the use of construct-laden SJTs in selection

[2] The meta-analytic effect size was collapsed across constructs.

programs. Although the present study offers no formal hypotheses regarding the correlation between the SJT personality measure and GMA measure, it is expected that the correlation will be in accordance with Kasten et al.’s (2020) findings. Table 1 presents a summary of all the hypotheses presented in this section.

Table 1
Summary of Hypotheses

Number	Hypothesis
1	There will be smaller mean differences between the “fake good” and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness compared to the single-statement measure of the same constructs.
2	There will be smaller mean differences between the “fake good” and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format compared to the single-statement measure of the same constructs using a Likert response format.
3	There will be smaller mean differences between the “fake good” and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format compared to the single-statement measure of the same constructs using a true-false response format.
4	There will be smaller mean differences between the “fake good” and honest conditions of the construct-laden SJT measure of (a) agreeableness and (b) conscientiousness using a rank format compared to the construct-laden SJT measure of agreeableness and conscientiousness using a rate format.
5	When respondents are explicitly instructed to fake, the criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.
6	When respondents are explicitly instructed to fake, the criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a (i) rank format, and (ii) rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a true-false response format.
7	When respondents are explicitly instructed to fake, criterion-related validity for the SJT of (a) agreeableness and (b) conscientiousness using a rank format will be higher than the criterion-related validity for the SJT using a rate response format.

Note. SJT = situational judgment test.

6. METHOD

Participants

Participants were recruited for the study using Amazon Mechanical Turk. Amazon Mechanical Turk allows participants to self-select into studies based on availability, required activities, and compensation. Inclusion criteria consisted of the following requirements: (a) 18 years of age or older, (b) employed with a part-time or full-time status, (c) a resident of the United States of America, and (d) proficient in the English language. Based on a power analysis, the initial goal was to recruit 800 participants. For the present study, the effect size, Cohen's *q*, represents the magnitude of the difference between standardized mean differences (*d*s); in this context, Cohen's *q* is computed by comparing (a) the honest and faking groups completing the SJT, and (b) the honest and faking groups completing the single-statement measure. With a sample size of 602, Kasten et al. (2020) obtained a *d* of 1.14 between honest and faking conditions for their Likert measure and a d of 0.66 for their SJT measure. Given that the present study is similar in design, using an alpha of .05 and power of .80, it was determined that a sample size of 730 was required to detect a comparable effect (q = 0.21; Likert d = 1.14; SJT d = 0.66). In addition, given the additional comparisons between conditions for the present study relative to Kasten et al., a larger sample size (N = 800) was sought.

Of the 1,311 initial responses to the inclusion criteria items, 131 participants did not meet the inclusion criteria, resulting in a total of 1,180 participants moving forward to the study measures. Computer-generated survey codes were provided to participants at the end of the measures to verify the authenticity of responses. The provided survey codes were then linked to participants' Mechanical Turk user IDs to ensure that each participant originated from Amazon Mechanical Turk. Participants exiting the measures before completion or those attempting the survey without authorization from Amazon Mechanical Turk were unable to be verified and were thus excluded. This resulted in the elimination of 326 such responses that were either incomplete or did not contain a valid survey code, yielding a total of 854 completed surveys. The 854 participants with valid survey codes were paid $0.50 for participating and those passing two out of three quality check items were paid $3.00 for completing the measures. A data quality check was performed to screen data for suspected duplicate responses (*n* = 46) and failure to pass at least 2 out of 3 quality check items (*n* = 26). This resulted in a total of 782 remaining responses. A final data quality check revealed that 74.56% of the 782 participants passed the manipulation check, resulting in 583 valid and usable responses.

Demographics. The following demographic information was obtained from participants: age, sex, ethnicity/race, position, education level, and employment status. Tenure was recorded as the length of time participants were employed at the same organization regardless of position. Of the 583 participants submitting valid responses, 289 (49.57%) were male, 292 (50.09%) were female, and two (0.34%) identified as "other." The mean age of participants was 38.22 years (*SD* = 12.49). Of the 583 participants, 500 (85.76%) were employed full-time and 83 (14.24%) part-time. Average tenure at the current place of employment, regardless of position or occupation, was 6.68 years (*SD* = 6.36). Counts and percentages for these demographic variables are presented in Table 2.

Table 2
Participant Demographics

Variable	*n*	%
Sex		
Male	289	49.57
Female	292	50.09
Other	2	0.34
Education		
High School	70	12.01
Technical/Vocational School	15	2.57
Associate's	59	10.12
Bachelor's	309	53.00
Master's Degree	115	19.73
PhD	15	2.57
Employment		
Full-time	500	85.76
Part-time	83	14.24
Race/Ethnicity		
African American or Black	52	8.92
American Indian or Alaska Native	4	0.69
Asian	39	6.69
Hispanic	34	5.83
Native Hawaiian or Pacific Islander	2	0.34
Two or More Races	8	1.37
White	443	75.99
Other	1	0.17

Note: N = 583

Measures

The study measures are described below along with information pertaining to the scoring procedures used and internal consistency reliability estimates. Sample items for these measures are presented in the appendix.

General mental ability. Cognitive ability was operationalized as scores on a general mental ability test (GMA$_{60}$; Arthur, 2017b). Participants were allotted 10 min to complete the 60-item (36 verbal, 24 numeric), 4-alternative multiple-choice assessment. Convergent validities of .42 – .55 have been obtained with ACT and SAT scores, along with criterion-related validities of .24 – .29 with GPA, and .32 with supervisor ratings of work performance (Arthur, 2017b). Retest reliabilities (7–10 days) of .76 and .70 have been reported for two alternate forms of the test (Naber, Arthur, Edwards, & Franco-Watkins, 2020). Test scores were computed as the total number of items answered correctly.

Social desirability responding. The Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1988, 1991) was administered to all participants. For each item, respondents were asked to rate how true a statement is descriptive of them on a seven-point Likert scale (1= not true, 7 = very true). Scores were computed as the total number of items on which the participant responded with a 6 or 7. Paulhus (1988) reported internal consistency reliability estimates between .75 and .86 for the impression management scale of the BIDR. Similar results were obtained in the present study with an internal consistency reliability estimate of .83 for the impression management scale.

Single-statement measures of agreeableness and conscientiousness. Agreeableness and conscientiousness were measured using the International Personality Item Pool (Goldberg, 1999) scale items (20 each) for each construct. Participants completing the single-statement measures used either a Likert response format (a traditional five-point scale [1 = very inaccurate, 5 = very accurate]) or true-false response format and responded to each statement in terms of the extent to which it was descriptive of them. The personality assessments were not timed. Crossing the instruction set (honest and faking) with response format (Likert and true-false) resulted in four distinct single-statement agreeableness measures and four distinct single-statement conscientiousness measures. Scores on the Likert and true-false measures were computed by summing the respondents' ratings for each item.

Internal consistency reliability estimates of .79 - .81 for conscientiousness and .80 - .85 for agreeableness have been reported (Donnellan, Oswald, Baird, & Lucas, 2006; Lim & Ployhart, 2006). In the present study, for the agreeableness measure using the rate response format with honest instructions, the internal consistency reliability estimate was .89, and with faking instructions, .92. For the agreeableness measure using the true-false response format, an internal consistency reliability estimate of .82 with honest instructions, and .89 with faking instructions were obtained.

For the conscientiousness measures using the rate response format with honest instructions, the internal consistency reliability estimate was .91 with honest, and .94 with faking instructions. For the conscientiousness measure using the true-false response format, with honest instructions an internal consistency reliability estimate of .84, and an estimate of .88 with faking instructions were obtained.

Situational judgment tests. SJTs of agreeableness and conscientiousness (Arthur, 2017a) were also administered. The SJT measure consists of 14 items (7 agreeableness, 7 conscientiousness). In the rate response format condition, for each item, the test-taker was presented with a scenario and 4 responses to the scenario which were then rated in terms of their effectiveness (1 = very ineffective, 5 = very effective) as responses to the scenario. In the rank condition, test-takers ranked each response to the scenario in terms of its effectiveness (1 = very ineffective, 4 = very effective) with no ties allowed. Scores were computed as the total number of correct items which was defined by whether the participant's responses matched the predetermined scoring key corresponding to participants' condition (rate or rank). Internal consistency reliability estimates of .75 and .80 have been reported (Arthur, 2017b) for the agreeableness and conscientiousness SJTs, respectively, using a rate response format.

For the agreeableness measures using the rate response format, an internal consistency reliability estimate of .85 was obtained for the honest condition, and .79 for the faking condition. For the rank response format, an internal consistency reliability estimate of .57 was obtained for the honest condition, and .68 for the faking condition.

For the conscientiousness measures using the rate response format, an internal consistency reliability estimate of .84 was obtained for the honest condition, and .84 for the faking condition. For the rank response format, an internal consistency reliability estimate of .60 was obtained for the honest condition, and .77 for the faking condition.

Organizational citizenship behaviors. Organizational citizenship behaviors (OCBs) were measured with 11 items compiled by Carpenter, Newman, and Arthur (2020). Using a supervisor-perspective, participants responded to each item on a five-point Likert scale (1 = very inaccurate, 5 = very accurate) to rate the extent to which they engage in helping behaviors that are not formally required in their work roles. The complete instructions can be found in the appendix. The assessment was not timed. An internal consistency reliability estimate of .90 has been reported (Carpenter et al., 2020) for the 11-item measure of OCBs; an estimate of .91 was obtained for the present study. Scores were computed as the sum of respondents' ratings to each item.

Counterproductive work behaviors. Counterproductive work behaviors (CWBs) were measured with 17 items compiled by Carpenter et al. (2020). Using a supervisor-perspective, participants responded to each item using a five-point Likert scale (1 = never, 5 = every day) to rate the extent to which they engage in counterproductive behaviors that are prohibited in the workplace. An internal consistency reliability estimate of .97 has been reported (Carpenter et al., 2020) for the 17 CWB items; an estimate of .98 was obtained for the present study. Scores were computed as the sum of respondents' ratings to each item.

Design and Procedure

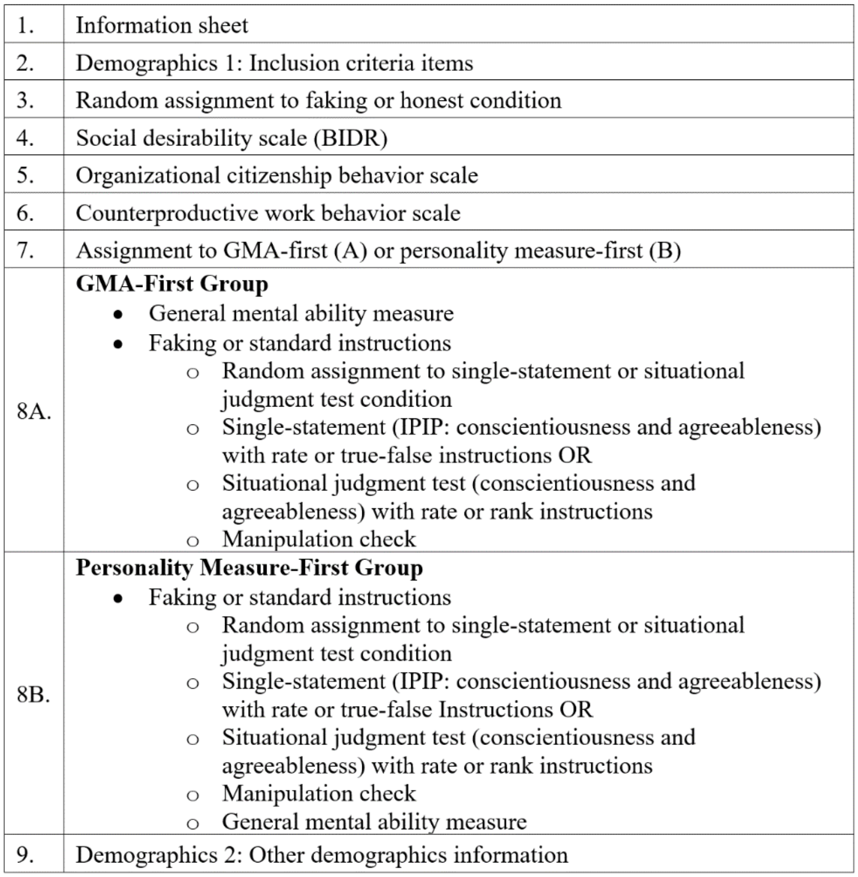
An experimental design was employed whereby participants were first randomly assigned to a condition (honest or faking) and then testing method (SJT or single-statement) and response format (rate, rank, Likert, true-false), as illustrated in Figure 2. Specifically, of the final sample of 583 participants, 336 were assigned to the faking condition and 247 to the honest condition. With respect to the testing method, participants were assigned to either the single-statement (*n* = 316) or SJT (*n* = 267) conditions. Those in the honest condition completing the single-statement measure were further split into Likert (*n* = 87) and true-false response format conditions (*n* = 86). Similarly, those in the faking condition and single-statement testing method were split into Likert (*n* = 71) and true-false conditions (*n* = 72). Those in the honest condition completing the SJT measure were further split into rate (*n* = 77) and rank conditions (*n* = 86). Finally, those in the faking condition and SJT method were split into rate (*n* = 51) and rank groups (*n* = 53). The difference in sample sizes across conditions was the result of unbalanced pass rates from the data quality checks. All assessments were administered remotely over the Internet.

After choosing to participate via the online system, participants read information sheets before electronically providing their consent to participate in the study. Subsequently, they were provided a single URL to complete all measures, which began with the BIDR. After completing the BIDR, participants completed the OCB and CWB measures. They then either completed the GMA measure or personality measure first. Presentation of the GMA or personality measure was counterbalanced to ensure that approximately half of the participants received the GMA measure first. Specifically, 307 participants received the GMA measure first and 276 received the personality measure first.

Faking instructions. Participants in the honest condition were given standard instructions corresponding to the specified testing method (i.e., single-statement or SJT). Those in the faking condition received modified instructions in which they were asked to imagine that they were applying for a job and to present themselves in the most favorable light to maximize their chances of being hired. The appendix provides the specific honest and faking instructions for each of the personality measures.

Manipulation check. A manipulation check (i.e., "When you took the personality test, which of the two instructions were you given?") was administered after completion of the personality measure. Participants selected from two instruction sets corresponding to either the honest or faking instructions for the specified personality measure. Of the 387 participants who were initially assigned to the honest condition, 51 did not pass, resulting in 85.82% (*n* = 336) who accurately recalled the manipulation they received. In the faking condition, 148 participants did not pass the manipulation check, resulting in 247 of 395 (62.53%) participants accurately recalling the manipulation they received. All participants failing the manipulation check (*n* = 199) were excluded, resulting in 583 participants being retained for the final statistical analyses. Table 3 presents a sequential outline and review of the activities presented within the current section.

Table 3
Sequence of Procedures



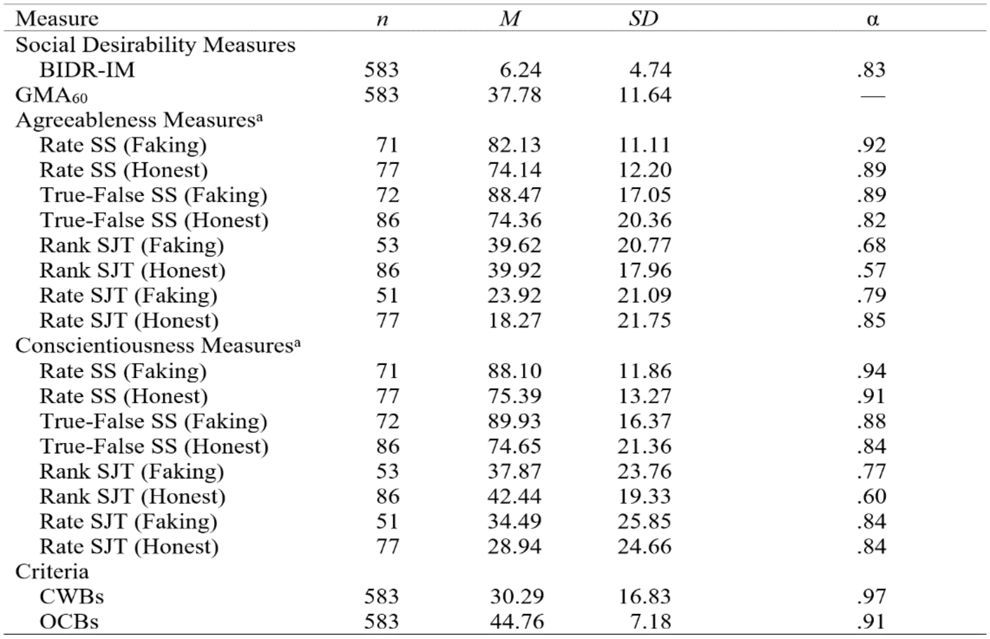
1.	Information sheet
2.	Demographics 1: Inclusion criteria items
3.	Random assignment to faking or honest condition
4.	Social desirability scale (BIDR)
5.	Organizational citizenship behavior scale
6.	Counterproductive work behavior scale
7.	Assignment to GMA-first (A) or personality measure-first (B)
8A.	**GMA-First Group** • General mental ability measure • Faking or standard instructions o Random assignment to single-statement or situational judgment test condition o Single-statement (IPIP: conscientiousness and agreeableness) with rate or true-false instructions OR o Situational judgment test (conscientiousness and agreeableness) with rate or rank instructions o Manipulation check
8B.	**Personality Measure-First Group** • Faking or standard instructions o Random assignment to single-statement or situational judgment test condition o Single-statement (IPIP: conscientiousness and agreeableness) with rate or true-false Instructions OR o Situational judgment test (conscientiousness and agreeableness) with rate or rank instructions o Manipulation check o General mental ability measure
9.	Demographics 2: Other demographics information

Note. BIDR = Balanced Inventory of Desirable Responding; GMA = general mental ability; IPIP = International Personality Item Pool.

7. RESULTS

Descriptive statistics and internal consistency reliability estimates for all the focal variables are presented in Table 4. For subsequent analyses, where warranted, *z*-tests, presented in Table 5, were used to compare the standardized mean differences (i.e., *ds*) between conditions and the differences between Fisher *z*-transformed independent Pearson correlations. The three focal independent variables of interest are condition (honest, faking), testing method (SJT, single-statement), and response format (rate, rank, and TF). Of note is that the terms *rate single-statement* and *Likert single-statement* are used interchangeably. The dependent variables are observed scores on the personality constructs (conscientiousness, agreeableness) in all instances.

Table 4
Descriptive Statistics and Internal Consistency Reliability Estimates for the Focal Measures



Measure	*n*	*M*	*SD*	α
Social Desirability Measures				
BIDR-IM	583	6.24	4.74	.83
GMA$_{60}$	583	37.78	11.64	—
Agreeableness Measures[a]				
Rate SS (Faking)	71	82.13	11.11	.92
Rate SS (Honest)	77	74.14	12.20	.89
True-False SS (Faking)	72	88.47	17.05	.89
True-False SS (Honest)	86	74.36	20.36	.82
Rank SJT (Faking)	53	39.62	20.77	.68
Rank SJT (Honest)	86	39.92	17.96	.57
Rate SJT (Faking)	51	23.92	21.09	.79
Rate SJT (Honest)	77	18.27	21.75	.85
Conscientiousness Measures[a]				
Rate SS (Faking)	71	88.10	11.86	.94
Rate SS (Honest)	77	75.39	13.27	.91
True-False SS (Faking)	72	89.93	16.37	.88
True-False SS (Honest)	86	74.65	21.36	.84
Rank SJT (Faking)	53	37.87	23.76	.77
Rank SJT (Honest)	86	42.44	19.33	.60
Rate SJT (Faking)	51	34.49	25.85	.84
Rate SJT (Honest)	77	28.94	24.66	.84
Criteria				
CWBs	583	30.29	16.83	.97
OCBs	583	44.76	7.18	.91

Note. [a]scores have been converted to a percentage ranging from 0 to 100; SS = single-statement; BIDR-IM = Balanced Inventory of Desirable Responding Impression Management scale; SJT = situational judgment test; GMA$_{60}$ = general mental ability test; CWBs = counterproductive work behavior; OCBs = organizational citizenship behaviors.

Table 5

Comparisons of Standardized Mean Differences Between Honest and Faking Conditions

				Statistical Comparison	
Measure	*d*	Measure	*d*	*z*-test	*q*
Agreeableness					
SJT (All Formats)	0.10	SS (All formats)	0.70	-3.50*	0.29
Rank SJT	-0.02	Rate SJT	0.26	-1.11	0.14
Rank SJT	-0.02	Rate SS	0.68	-2.91*	0.33
Rank SJT	-0.02	TF SS	0.75	-3.17*	0.36
Rate SJT	0.26	Rate SS	0.68	-1.71*	0.20
Rate SJT	0.26	TF SS	0.75	-1.97*	0.23
Conscientiousness					
SJT (All Formats)	0.01	SS (All formats)	0.86	-4.95*	0.41
Rank SJT	-0.22	Rate SJT	0.22	-1.75*	0.22
Rank SJT	-0.22	Rate SS	1.00	-5.00*	0.59
Rank SJT	-0.22	TF SS	0.79	-4.19*	0.49
Rate SJT	0.22	Rate SS	1.00	-3.14*	0.37
Rate SJT	0.22	TF SS	0.79	-2.32*	0.27

Note. SS = single-statement measure; SJT = situational judgment test; TF = true-false; Cohen's *q* is an effect size which represents the absolute value of the difference between the Fisher-*z* transformed standardized mean differences. **p* < .05 (one-tailed).

Differences Between Conditions

SJT vs. single-statement. Hypothesis 1 posited that there would be smaller mean differences between the fake good and honest conditions of the SJT measures of agreeableness and conscientiousness compared to single-statement measures of the same constructs. To facilitate the test of this hypothesis, a linear transformation was applied to participant scores on the rate and rank SJT measures, as well as the Likert and true-false measures. Specifically, scores on these measures were converted to a percentage of the maximum attainable score on the measures, which effectively converted all scores to a 100-point scale. The standardized mean differences between the conditions were then computed. As the results in Table 5 indicate, the difference between the faking and honest conditions on the agreeableness SJT across the rank and rate response formats (*d* = 0.10) was smaller than the difference between conditions for the agreeableness single-statement measure averaged across response formats (*d* = 0.70). Figure 3 illustrates these differences. A *z*-test revealed a statistically significant difference (*z* = -3.50, *p* < .05), and thus, Hypothesis 1(a) was supported. Table 5 presents all standardized mean differences, which are organized by the corresponding testing method and response format comparison; it also includes *z*-test results and Cohen's *q*, which is the magnitude of the effect.

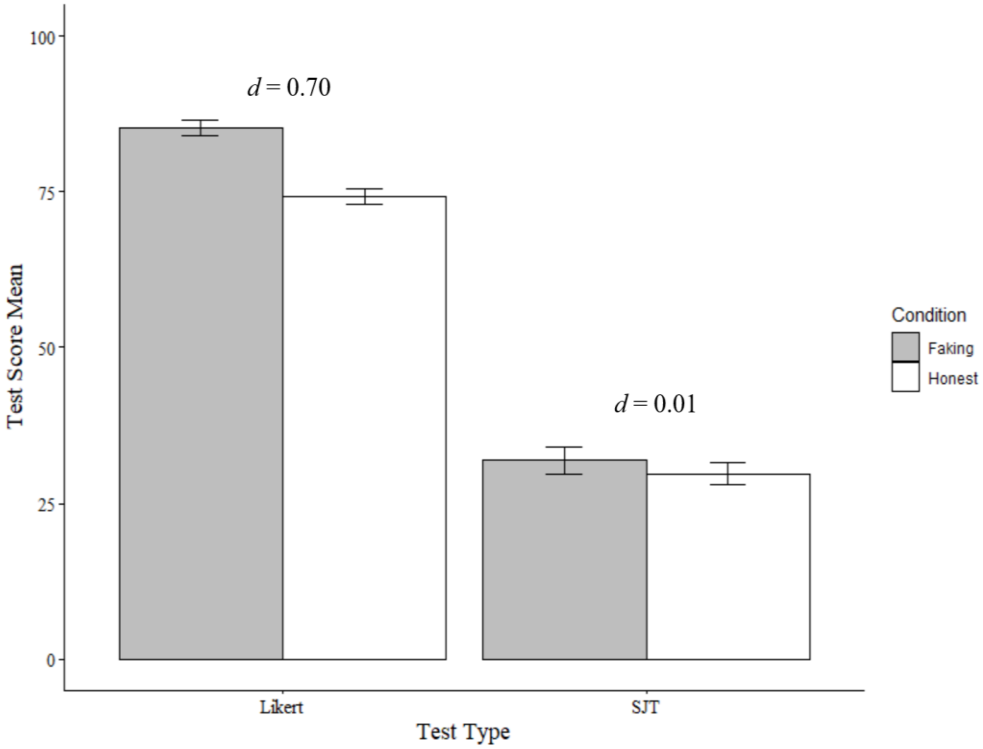


Figure 3. Mean agreeableness scores by testing method and condition, across response formats. Error bars represent standard error of the mean. SJT = situational judgment test.

Similarly, the difference between the faking and honest scores on the conscientiousness SJT across response formats (*d* = 0.01) was smaller than the difference between conditions for the conscientiousness single-statement measure across response formats (*d* = 0.86; see Figure 4). A *z*-test revealed a statistically significant difference (*z* = -4.95, *p* < .05), and thus, Hypothesis 1(b) was supported. As predicted, the SJTs were more resilient to intentional response distortion than the single-statement measures averaged across response formats.

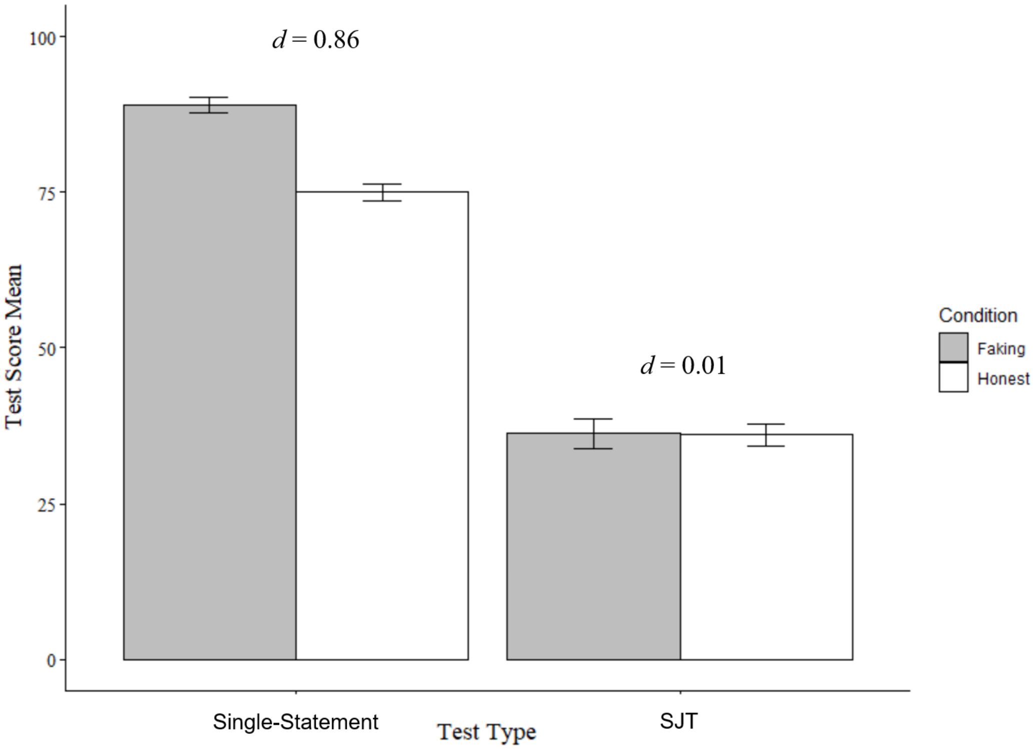


Figure 4. Mean conscientiousness scores by testing method and condition, across response formats. Error bars represent standard error of the mean. SJT = situational judgment test.

Rank and rate SJT vs. Likert. Hypothesis 2 predicted that the differences between the honest and faking conditions would be smaller for participants completing the rate and rank SJTs compared to those completing the Likert measures. The results, as illustrated in Figure 5, indicated that the difference between faking and honest scores on the agreeableness SJT using a rank format (*d* = -0.02) was significantly smaller than the difference between conditions for the agreeableness single-statement measure (*d* = 0.68; *z* = -2.91, *p* < .05), supporting Hypothesis 2(a)(i). In addition, the difference between the faking and honest scores on the agreeableness SJT using a rate format (*d* = 0.26) was significantly smaller than the difference between conditions for the agreeableness Likert measure (*d* = 0.68; *z* = -1.71, *p* < .05), and thus, Hypothesis 2(a)(ii) was supported.

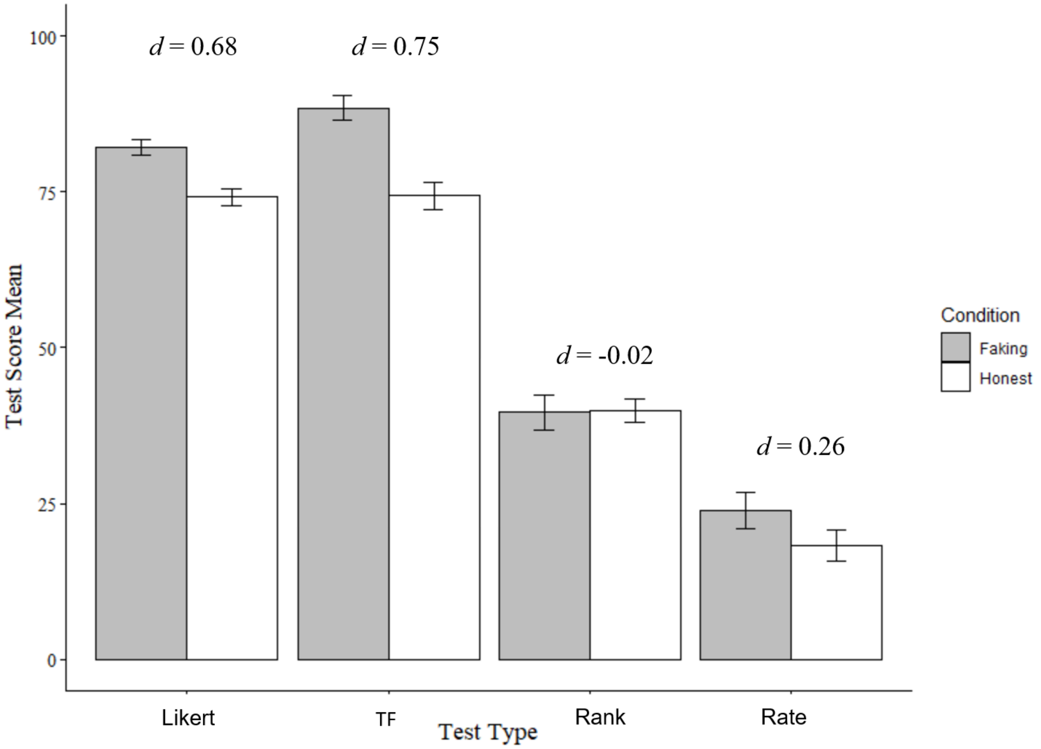


Figure 5. Mean agreeableness scores by condition, testing method, and response format. Error bars represent standard error of the mean. TF = true-false.

For conscientiousness, both the rank and rate SJT were more resilient to faking than the Likert measure. Specifically, the difference between the faking and honest scores on the conscientiousness SJT using a rank format (*d* = -0.22) were significantly smaller than the difference between conditions for the conscientiousness Likert measure (*d* = 1.00; *z* = -5.00, *p* < .05; see Figure 6), and thus Hypothesis 2(b)(i) was supported. In addition, the difference between conditions for the rate SJT of conscientiousness (*d* = 0.22) were smaller than the difference between conditions for the Likert measure of the same construct (*d* = 1.00; *z* = -3.14, *p* < .05). Thus, Hypotheses 2(b)(ii) was supported.

Rate and rank SJT vs. true-false. Hypothesis 3 predicted that the mean difference between the honest and faking conditions would be smaller for participants completing the rate and rank SJTs compared to those completing the single-statement measure using a true-false format. The results, as illustrated in Figure 5, indicated that the mean difference between faking and honest scores on the agreeableness SJT using a rank format (*d* = -0.02) was significantly smaller than the difference between conditions for the agreeableness true-false measure (*d* = 0.75; *z* = -3.17, *p* < .05), supporting Hypothesis 3(a)(i). The difference between faking and honest scores on the agreeableness SJT using a rate format (*d* = 0.26) was significantly different from the difference between conditions for the true-false measure of agreeableness (*d* = 0.75; *z* = -1.97, *p* < .05). Thus, Hypothesis 3(a)(ii) was supported.

As illustrated in Figure 6, for the conscientiousness SJT using a rank format, the mean difference between conditions (*d* = -0.22) was significantly different from that for the true-false measure of conscientiousness (*d* = 0.79; *z* = -4.19, *p* < .05), and so

Hypothesis 3(b)(i) was supported. Similar results were found when comparing the conscientiousness SJT using a rate format (*d* = 0.22) to the true-false measure of the same construct (*d* = 0.79; *z* = -2.32, *p* < .05). Thus, Hypothesis 3(b)(ii) was supported.

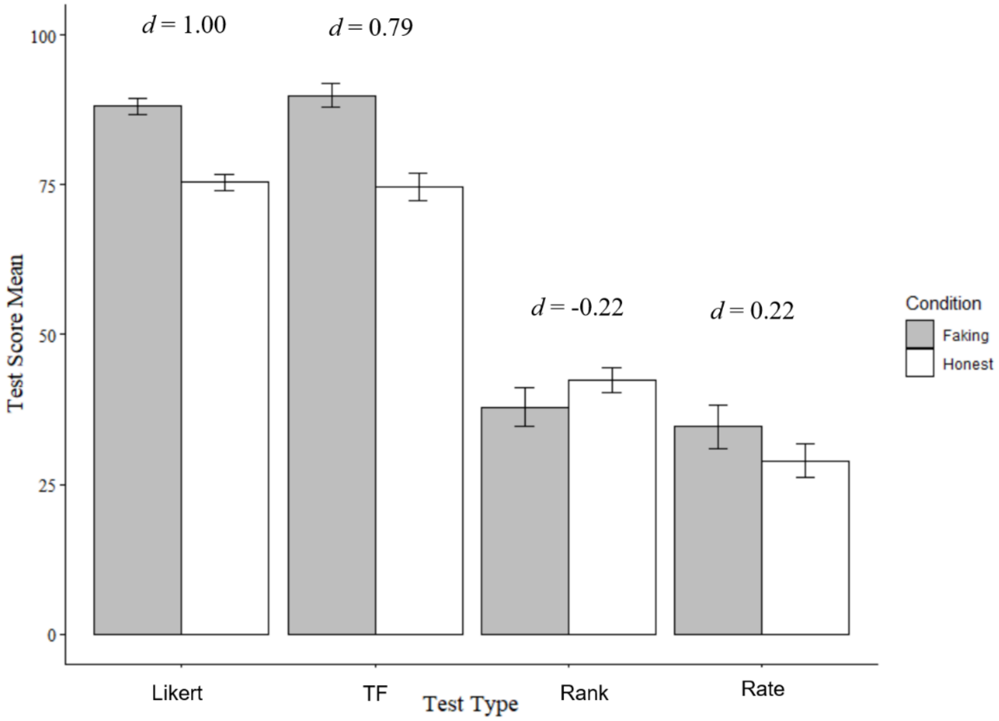


Figure 6. Mean conscientious scores by condition, testing method, and response format. Error bars represent standard error of the mean. TF = true-false.

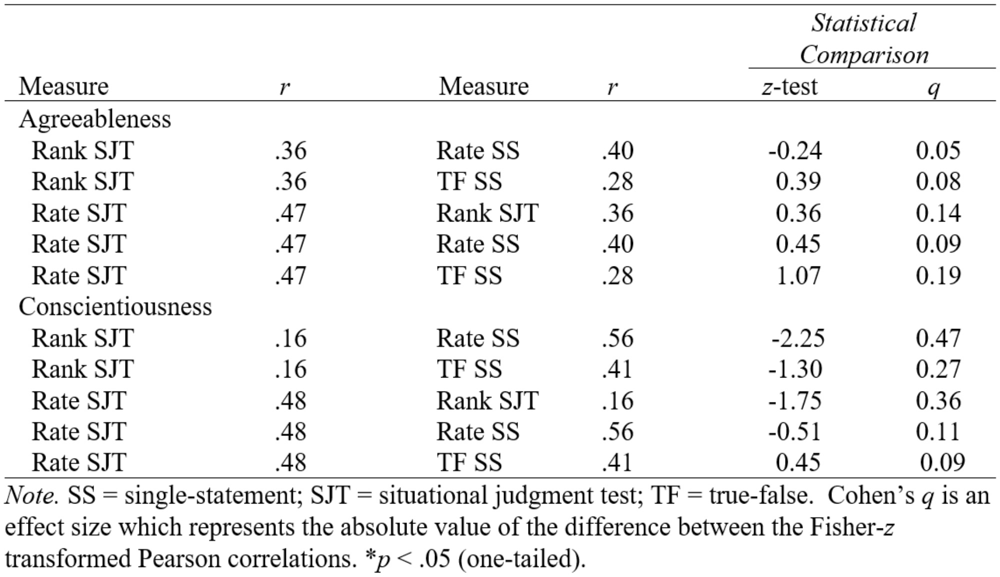
Rank SJT vs. rate SJT. Hypothesis 4 posited that the mean difference between the faking and honest conditions of the rank SJTs of agreeableness and conscientiousness would be smaller than the difference between conditions of the rate SJT of the same constructs. The mean difference between conditions for the agreeableness SJT using a rank response format (*d* = -0.02) was not significantly different from that for the rate SJT (*d* = 0.26; *z* = -1.11, *p* > .05), and thus Hypothesis

4(a) was not supported. In contrast to this finding, the difference between conditions for the conscientiousness SJT using a rank response format (*d* = -0.22) was significantly smaller than that for the rate SJT (*d* = 0.22; *z* = -1.75, *p* > .05), and thus Hypothesis 4(b) was supported.

Criterion-Related Validities

Agreeableness: Rank and rate SJT vs. Likert. Hypothesis 5 posited that when participants are explicitly instructed to fake, the association between scores on the personality measures and the criteria (OCBs and CWBs) will be higher for those completing SJT-based assessments using rank and rate response formats compared to those completing Likert measures. Comparisons between correlations were made by performing a *z*-test, which is a function of the difference between Fisher-*z* transformed correlations that takes into account the sampling variance. In addition, the magnitude of the simple difference between Fisher-*z* transformed correlations can be represented as Cohen's *q* (Ellis, 2010). Correlations and statistical comparisons between the personality measures and OCBs are presented in Table 6.

Table 6
Comparisons of Correlations Between Personality Measures and Organizational Citizenship Behaviors



				Statistical Comparison	
Measure	*r*	Measure	*r*	*z*-test	*q*
Agreeableness					
Rank SJT	.36	Rate SS	.40	-0.24	0.05
Rank SJT	.36	TF SS	.28	0.39	0.08
Rate SJT	.47	Rank SJT	.36	0.36	0.14
Rate SJT	.47	Rate SS	.40	0.45	0.09
Rate SJT	.47	TF SS	.28	1.07	0.19
Conscientiousness					
Rank SJT	.16	Rate SS	.56	-2.25	0.47
Rank SJT	.16	TF SS	.41	-1.30	0.27
Rate SJT	.48	Rank SJT	.16	-1.75	0.36
Rate SJT	.48	Rate SS	.56	-0.51	0.11
Rate SJT	.48	TF SS	.41	0.45	0.09

Note. SS = single-statement; SJT = situational judgment test; TF = true-false. Cohen's *q* is an effect size which represents the absolute value of the difference between the Fisher-*z* transformed Pearson correlations. **p* < .05 (one-tailed).

With respect to OCBs, the results presented in Table 6 indicate that the criterion-related validity for the rank SJT for agreeableness (*r* = .36) was lower but not significantly different from the Likert measure (*r* = .40; *z* = -0.24, *p* > .05). For CWBs, as presented in Table 7, the criterion-related validity for the rank SJT for agreeableness (*r* = -.29) was not significantly different from that for the Likert measure (*r* = -.36; *z* = 0.39, *p* > .05). Thus, Hypothesis 5(a)(i) was not supported. Similar results were obtained in comparisons of the association between scores on the rate SJT of agreeableness and OCBs (*r* = .47) and the association between the Likert measure of agreeableness and OCBs (*r* = .40, *z* = 0.45, *p* > .05). For CWBs, the criterion-related validity for the rate SJT of agreeableness (*r* = -.37) was not significantly different from that for the Likert measure (*r* = -.36; *z* = -0.06; *p* > .05). Thus, Hypotheses 5(a)(ii) also was not supported.

Table 7
Comparisons of Correlations Between Personality Measures and Counterproductive Work Behaviors

				Statistical Comparison	
Measure	*r*	Measure	*r*	*z*-test	*q*
Agreeableness					
Rank SJT	-.29	Rate SS	-.36	0.39	0.08
Rank SJT	-.29	TF SS	-.52	1.35	0.28
Rate SJT	-.37	Rank SJT	-.29	0.45	0.09
Rate SJT	-.37	Rate SS	-.36	-0.06	0.01
Rate SJT	-.37	TF SS	-.52	0.90	0.19
Conscientiousness					
Rank SJT	-.42	Rate SS	-.73	2.39	0.49
Rank SJT	-.42	TF SS	-.69	1.97	0.41
Rate SJT	-.62	Rank SJT	-.42	1.34	0.28
Rate SJT	-.62	Rate SS	-.73	1.04	0.21
Rate SJT	-.62	TF SS	-.69	0.62	0.13

Note. SS = single-statement; SJT = situational judgment test; TF = true-false. Cohen's *q* is an effect size which represents the absolute value of the difference between the Fisher-*z* transformed Pearson correlations. **p* < .05 (one-tailed).

Conscientiousness: Rank and rate SJT vs. Likert. Similar results were obtained for conscientiousness. As presented in Table 6, with respect to OCB, surprisingly, the criterion-related validity for the rank SJT of conscientiousness (*r* = .16) was not significantly higher than that for rate Likert measure of the same construct (*r* = .56; *z* = -2.25, *p* > .05). In addition, contrary to the hypothesis, for CWBs, as presented in Table 7, the criterion-related validity for the rank SJT of conscientiousness (*r* = -.42) was not significantly stronger than the Likert measure (*r* = -.73; *z* = 2.39, *p* > .05). Taking these results together, Hypothesis 5(b)(i) was not supported. The association between the rate SJT of conscientiousness and OCBs (*r* = .48) was also lower than the association between the Likert measure and OCBs (*r* = .56; *z* = -0.51, *p* > .05). For CWBs, the criterion-related validity for the rate SJT of conscientiousness (*r* = -.62) was not significantly stronger than that for the Likert measure (*r* = -.73; *z* = 1.04, *p* > .05) and thus Hypothesis 5(b)(ii) was not supported.

Agreeableness: Rank and rate SJT vs. true-false. Hypothesis 6 posited that when participants are explicitly instructed to fake, the associations between scores on the two personality constructs and the criteria will be higher for those completing SJT-based assessments using rank and rate formats compared to those completing true-false measures. With regard to OCB, as presented in Table 6, the criterion-related validity for the rank SJT of agreeableness (*r* = .36) was higher but not significantly different from that for the true-false measure (*r* = .28; *z* = 0.39, *p* > .05). For CWBs, as presented in Table 7, the association with the rank SJT of agreeableness (*r* = -.29) was weaker than the association between CWBs and the true-false measure (*r* = -.52; *z* = 1.35, *p* > .05). Thus, Hypothesis 6(a)(i) was not supported.

For OCBs, the criterion-related validity for the rate SJT of agreeableness (*r* = .47) was higher but not significantly different from that for the true-false measure (*r* = .28; *z* = 1.07, *p* > .05). For CWBs, the criterion-related validity for the rate SJT of agreeableness (*r* = -.37) was not significantly different from that for the true-false measure (*r* = -.52; *z* = 0.90, *p* > .05). Thus, Hypotheses 6(a)(ii) was not supported.

Conscientiousness: Rank and rate SJT vs. true-false. With regard to OCB, the criterion-related validity for the rank SJT of conscientiousness (*r* = .16) was lower than the criterion-related validity for the true-false measure (*r* = .41; *z* = -1.30, *p* > .05; see Table 6). Also contrary to the hypothesis, for CWBs, the criterion-related validity for the rank SJT of conscientiousness (*r* = -.42) was lower than that for the true-false measure (*r* = -.69; *z* = 1.97, *p* < .05; see Table 7). Thus, Hypothesis 6(b)(i) was not supported.

The association between the rate SJT of conscientiousness and OCBs (*r* = .48) was also found to be higher but not significantly different from that for the true-false measure (*r* = .41; *z* = 0.45, *p* > .05). For CWBs, the criterion-related validity for the rate SJT of conscientiousness (*r* = -.62) was weaker than that for the true-false measure (*r* = -.69; *z* = 0.62, *p* > .05) and thus Hypothesis 6(b)(ii) was not supported.

Agreeableness: Rate SJT vs. rank SJT. Hypothesis 7(a) posited that the association between agreeableness and OCBs and agreeableness and CWBs would be stronger for the rank SJT compared to the rate SJT. Contrary to the hypothesis, the criterion-related validities for the agreeableness SJT using the rank format (*r* = .47 for OCBs, presented in Table 6, and *r* = -.37 for CWBs, presented in Table 7) were not statistically different from those for the rate format (*r* = .36 for OCBs and *r* = -.29 for CWBs). With respect to OCBs, the *z*-test was not statistically significant (*z* = -0.69, *p* > .05) and the same pattern held for CWBs (*z* = 0.45, *p* > .05; see Table 7). Thus, Hypothesis 7(a) was not supported.

Conscientiousness: Rate SJT vs. rank SJT. Hypothesis 7(b) posited that the association between conscientiousness and OCBs and conscientiousness and CWBs would be stronger for the rank SJT compared to the rate SJT. However, the criterion-related validities for the rate SJT of conscientiousness (*r* = .48 for OCBs, presented in Table 6, and *r* = -.62 for CWBs, presented in Table 7) were not significantly different from those for the rank SJT (*r* = .16 for OCBs and *r* = -.42 for CWBs). The corresponding *z*-tests comparing the OCB correlations (*z* = -1.75, *p* >.05) and CWB correlations (*z* =1.34, *p* > .05) were not significant, and thus Hypothesis 7(b) was not supported.

Table 8

Descriptive Statistics and Correlations, Collapsed Across Test Type, Response Format, and Condition

Variable	*M*	*SD*	1	2	3	4	5
1. AGREE	56.96	31.17					
2. CONSC	60.65	30.53	.84*				
3. CWB	30.29	16.83	-.26*	-.38*			
4. GMA_{60}	37.78	11.64	.21*	.32*	-.63*		
5. OCB	44.76	7.18	.27*	.30*	-.30*	.32*	
6. BIDR-IM	6.24	4.74	.25*	.24*	-.33*	.24*	.39*

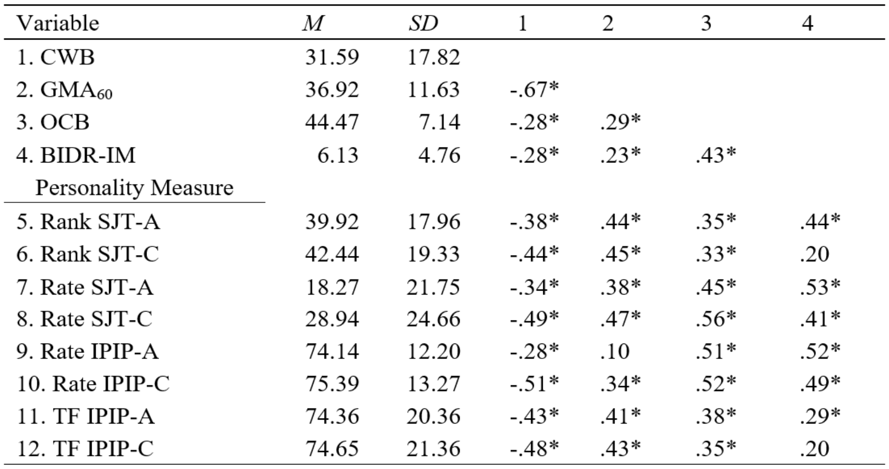
Note. AGREE and CONSC = agreeableness and conscientiousness measures across test type and response format, respectively; CWB and OCB = organizational citizenship behaviors and counterproductive work behaviors, respectively, measured across honest and faking condition; GMA = general mental ability test; BIDR-IM = the Balanced Inventory of Desirable Responding Impression Management scale and measured across both conditions; **p* < .05 (two-tailed).

Personality-GMA Associations

Correlations for all variables (across both honest and faking conditions) are presented in Table 8. Condition-specific correlations are presented in Table 9 (honest condition) and Table 10 (faking condition). In the absence of formal hypotheses, personality-GMA correlations were examined on an exploratory basis. Specifically, the associations between general mental ability and personality scores (a) across conditions, (b) in the honest condition, and (c) in the faking condition were examined. As presented in Table 8, GMA was significantly correlated with personality scores collapsed across testing method, response format, and condition (*r* = .21 for agreeableness; *r* = .32 for conscientiousness). When taking the conditions into account, the association between personality test scores and GMA appears to be stronger in the faking condition.

Table 9

Descriptive Statistics and Correlations for Participants in the Honest Condition

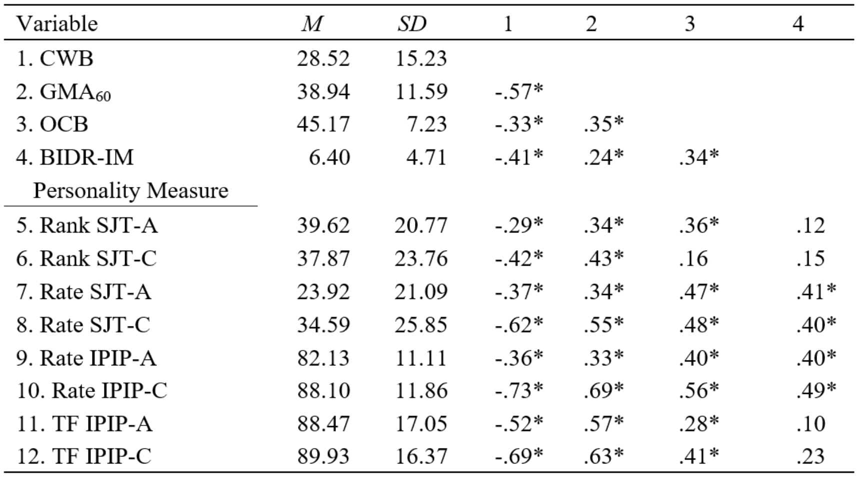


Variable	*M*	*SD*	1	2	3	4
1. CWB	31.59	17.82				
2. GMA_{60}	36.92	11.63	-.67*			
3. OCB	44.47	7.14	-.28*	.29*		
4. BIDR-IM	6.13	4.76	-.28*	.23*	.43*	
Personality Measure						
5. Rank SJT-A	39.92	17.96	-.38*	.44*	.35*	.44*
6. Rank SJT-C	42.44	19.33	-.44*	.45*	.33*	.20
7. Rate SJT-A	18.27	21.75	-.34*	.38*	.45*	.53*
8. Rate SJT-C	28.94	24.66	-.49*	.47*	.56*	.41*
9. Rate IPIP-A	74.14	12.20	-.28*	.10	.51*	.52*
10. Rate IPIP-C	75.39	13.27	-.51*	.34*	.52*	.49*
11. TF IPIP-A	74.36	20.36	-.43*	.41*	.38*	.29*
12. TF IPIP-C	74.65	21.36	-.48*	.43*	.35*	.20

Note. OCB and CWB = organizational citizenship behaviors and counterproductive work behaviors, respectively; GMA = general mental ability test; BIDR-IM = Balanced Inventory of Desirable Responding Impression Management scale; IPIP = International Personality Item Pool; SJT = situational judgment test; A = agreeableness; C = conscientiousness; TF = true-false. **p* < .05 (two-tailed).

Table 10

Descriptive Statistics and Correlations for Participants in the Faking Condition



Variable	*M*	*SD*	1	2	3	4
1. CWB	28.52	15.23				
2. GMA_{60}	38.94	11.59	-.57*			
3. OCB	45.17	7.23	-.33*	.35*		
4. BIDR-IM	6.40	4.71	-.41*	.24*	.34*	
Personality Measure						
5. Rank SJT-A	39.62	20.77	-.29*	.34*	.36*	.12
6. Rank SJT-C	37.87	23.76	-.42*	.43*	.16	.15
7. Rate SJT-A	23.92	21.09	-.37*	.34*	.47*	.41*
8. Rate SJT-C	34.59	25.85	-.62*	.55*	.48*	.40*
9. Rate IPIP-A	82.13	11.11	-.36*	.33*	.40*	.40*
10. Rate IPIP-C	88.10	11.86	-.73*	.69*	.56*	.49*
11. TF IPIP-A	88.47	17.05	-.52*	.57*	.28*	.10
12. TF IPIP-C	89.93	16.37	-.69*	.63*	.41*	.23

Note. CWB and OCB = organizational citizenship behaviors and counterproductive work behaviors, respectively; GMA = general mental ability test; BIDR-IM = Balanced Inventory of Desirable Responding Impression Management scale; IPIP = International Personality Item Pool; A = agreeableness; C = conscientiousness; TF = true-false; SJT = situational judgment test. **p* < .05 (two-tailed).

Personality-GMA associations by condition. The results in Table 9 indicate that across testing methods and response formats, in the honest condition, the association between agreeableness and GMA ranged from .10 to .44, and the association between conscientious and GMA ranged from .34 to .47. In contrast, in the faking condition (Table 10), the correlations were noticeably larger—ranging from .33 to .57 for the agreeableness-GMA association and .43 to .69 for the conscientiousness-GMA association.

Testing method and response format differences. The extent to which the personality-GMA associations were influenced by testing method and response format was also examined. In the honest condition, as presented in Table 9, the correlations were smaller but not substantially so for the agreeableness single-statement measures compared to the SJTs—correlations were .38 and .44 for the rate and rank SJTs, respectively, whereas the correlations were .10 and .41 for the Likert and true-false measures, respectively. The correlations for conscientiousness were largely similar across testing method—the correlations were .47 and .45 for the rate and rank SJTs, respectively, and .34 and .43 for the Likert and true-false measures, respectively. Overall, the results indicate that in the honest condition, the personality-GMA correlations did not show substantial differences between testing methods.

Surprisingly, this pattern holds in the faking condition, as the SJT and single-statement measures displayed similar associations with GMA across personality constructs (see Table 10)—correlations ranged from .34 to .55 for the SJT measures, and from .33 to .69 for the single-statement measures. This result was unexpected, as historically, SJTs have been thought to have higher correlations with GMA than single-statement measures regardless of the construct tested. The results obtained here are in direct contrast to this.

There also appeared to be no substantial differences between the different response formats nested within each testing method. That is, across personality constructs and honest and faking conditions, when directly comparing the personality-GMA correlations for the rank SJT (*r* = .34 to *r* = .45) to the rate SJT (*r* = .34 to *r* = .55) and Likert (*r* = .10 to *r* = .69) to the true-false measures (*r* = .41 to *r* = .63), the personality-GMA correlations appear to be largely similar with no substantial differences emerging for the different response formats, as reflected in both Table 9 and 10.

Personality Construct-Social Desirability Responding Associations

One important metric which can be used to determine the resilience of a personality measure to response distortion is a measure's correlations with an external social desirability measure (Arthur, Hagen, & George, 2021). In the present study, one such measure of social desirability responding, the BIDR measure, was administered.

As presented in Table 9, for the honest condition, the SJT-BIDR associations ranged from .20 to .53 for the SJT, whereas they ranged from .20 to .52 for the single-statement measures. Thus, overall, the SJT and single-statement measures had largely similar associations with the BIDR when participants were given an honest instruction set. As presented in Table 10, similar results were obtained in the faking condition; specifically, SJT-BIDR associations ranged from .12 to .41 for the SJT and from .10 to

.49 for the single-statement measures. In the faking condition, more substantial differences between testing methods are revealed when examined in the context of the various response formats used. Of the two SJT response formats (rank and rate) and the two single-statement formats (Likert and true-false), the rank SJT (*r* = .12 for agreeableness and *r* = .15 for conscientiousness) and true-false formats (*r* = .10 for agreeableness and *r* = .23 for conscientiousness) appeared to have weaker correlations with the BIDR. This provides further evidence that the rank SJT appears to be superior to the rate SJT and Likert measures in terms of faking resistance. It also provides some evidence that the true-false format compares favorably to the Likert format under specific conditions. A summary of the results for the hypotheses pertaining to the mean differences between honest and faking conditions and differences in criterion-related validity are presented in Table 11 and Table 12, respectively.

Table 11

Summary of Results for Hypotheses: Differences Between Honest and Faking Conditions

	Hypothesis	Results
	There will be smaller mean differences between the "fake good" and honest conditions of the construct-laden SJT measure of:	
H1(a)	agreeableness compared to the single-statement measure of the same constructs.	**Supported -** 0.60 difference between *d*s.
H1(b)	conscientiousness compared to the single-statement measure of the same constructs.	**Supported -** 0.85 difference between *d*s.
H2(a)(i)	agreeableness using a rank format compared to the single-statement measure of the same constructs using a Likert response format.	**Supported -** 0.70 difference between *d*s.
H2(a)(ii)	agreeableness using a rate format compared to the single-statement measure of the same constructs using a Likert response format.	**Supported –** effect in hypothesized direction and significant.
H2(b)(i)	conscientiousness using a rank format compared to the single-statement measure of the same constructs using a Likert response format.	**Supported -** large difference between *d*s (1.22).
H2(b)(ii)	conscientiousness using a rate format compared to the single-statement measure of the same constructs using a Likert response format.	**Supported -** large difference between *d*s (0.78).
H3(a)(i)	agreeableness using a rank format compared to the single-statement measure of the same constructs using a true-false response format.	**Supported -** 0.77 difference between *d*s.
H3(a)(ii)	agreeableness using a rate format compared to the single-statement measure of the same constructs using a true-false response format.	**Supported -** 0.49 difference between *d*s.
H3(b)(i)	conscientiousness using a rank format compared to the single-statement measure of the same constructs using a true-false response format.	**Supported –** 1.01 difference between *d*s.
H3(b)(ii)	conscientiousness using a rate format compared to the single-statement measure of the same constructs using a true-false response format.	**Supported -** 0.57 difference between *d*s.
H4(a)	agreeableness using a rank format compared to the construct-laden SJT measure of agreeableness using a rate format.	**Not supported -** difference between *d*s in hypothesized direction, but not significant.
H4(b)	conscientiousness using a rank format compared to the construct-laden SJT measure of conscientiousness using a rate format.	**Supported –** significant difference between *d*s.

Note. SJT = situational judgment test.

Table 12

Summary of Results for Hypotheses: Differences in Criterion-Related Validity

	Hypothesis	Results
	When respondents are explicitly instructed to fake, the criterion-related validity for the SJT of:	
H5(a)(i)	agreeableness using a rank format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.	**Not supported -** difference in *r*s not in hypothesized direction.
H5(a)(ii)	agreeableness using a rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.	**Not supported -** difference in hypothesized direction.
H5(b)(i)	conscientiousness using a rank format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.	**Not supported –** *r*s higher for Likert measure.
H5(b)(ii)	conscientiousness using a rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a rate response format.	**Not supported -** small difference in *r*s.
H6(a)(i)	agreeableness using a rank format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a true-false response format.	**Not supported** - difference in hypothesized direction (OCB).
H6(a)(ii)	agreeableness using a rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a true-false response format.	**Not supported -** difference in hypothesized direction (OCB).
H6(b)(i)	conscientiousness using a rank format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a true-false response format.	**Not supported -** TF higher than rank SJT for CWBs but not OCBs.
H6(b)(ii)	conscientiousness using a rate format will be higher than the criterion-related validity for the single-statement measure of the same constructs using a true-false response format.	**Not supported -** difference in hypothesized direction (OCB).
H7(a)	agreeableness using a rank format will be higher than the criterion-related validity for the SJT using a rate response format.	**Not supported -** rate correlations stronger than rank.
H7(b)	conscientiousness using a rank format will be higher than the criterion-related validity for the SJT using a rate response format.	**Not supported -** rate correlations stronger than rank.

Note. SJT = situational judgment test; TF = true-false; OCB = organizational citizenship behaviors; CWB = counterproductive work behaviors.

8. DISCUSSION AND CONCLUSION

Summary of Findings

The present study examined whether the effects of response distortion could be mitigated by the use of a construct-laden SJT. The expected results were grounded in the premise that measures with predetermined scoring keys—those for which each item has a right answer—should have a higher resistance to faking compared to those that do not. Because empirical evidence by Arthur (2017a), Arthur et al. (2014), and Kasten et al. (2020) supports this assertion, the present study attempted to replicate and extend these findings. Specifically, the present study examined whether the resistance of SJTs to faking is also a function of the response format.

It was hypothesized that there would be larger differences between honest and faking scores on single-statement measures of agreeableness and conscientiousness than on SJT measures using rank and rate response formats. In addition, an SJT using a rank response format was expected to be more resistant to faking than an SJT using a rate response format. It was also hypothesized that under faking conditions, the criterion-related validity of the personality measures would be higher for (a) those completing the SJT compared to the single-statement measures, and (b) those completing the rank SJT compared to the rate SJT. As discussed below, the obtained results supported the hypotheses positing that SJTs would have smaller mean differences between honest and faking conditions than single-statement measures. In contrast, all hypotheses pertaining to the criterion-related validity of the SJTs were not supported.

Difference between *d*s. A primary goal of the present study was to replicate Kasten et al.'s (2020) finding that SJTs were able to prevent mean shifts when participants were given explicit instructions to fake. This goal was achieved, as the mean differences between honest and faking conditions for those completing the SJT were similar to those obtained by Kasten et al. Specifically, the difference between the SJT and Likert faking and honest conditions for the present study was 0.60 for agreeableness and 0.85 for conscientiousness. Kasten et al. (2020), who assessed only conscientious, obtained a difference between the SJT and Likert faking and honest conditions of 0.93 in Study 1 (within-subjects) and 0.48 in Study 2 (between-subjects). Furthermore, in the present study, across constructs, the SJT measures had demonstrably smaller mean differences (*d* = 0.05) than the single-statement measures (Likert *d* = 0.84 and true-false *d* = 0.77).

The findings of the present study extend Kasten et al.'s findings by demonstrating that the extent to which an SJT can prevent faking is dependent on the response format that it uses; the rank SJT was particularly effective at preventing faking compared to single-statement measures. Accordingly, all hypotheses pertaining to the rank SJT's resilience to faking compared to single-statement measures (Likert and true-false) were supported. Similarly, the hypotheses for the rate SJT were supported. In addition, the rank SJT of conscientiousness had significantly smaller mean differences across conditions (*d* = -0.22) than the rate SJT (*d* = 0.22). This effect was not obtained for agreeableness (rank SJT *d* = -0.02, rate SJT *d* = 0.26). So, given the significant difference between *d*s when assessing conscientiousness (-0.22 – 0.22 = -0.44), and the moderate difference between *d*s when assessing agreeableness (-0.02 – 0.26 = -0.28), the present study contributes to the extant literature by demonstrating that the faking resiliency of SJTs is also a function of the response format. Regarding the nonsignificant finding when comparing the rank and rate SJTs of agreeableness, it should be contextualized by the analytical method used. That is, comparing standardized mean differences using *z*-tests, as was done in the present study, was an especially conservative method to assess differences between *d*s; thus, although there was a moderate difference between *d*s (0.28), the comparison between the rank SJT and rate SJT of agreeableness did not reach statistical significance.

Criterion-related validity and the true vs. error variance debate. It is reasonable to posit that if a testing method or response format is able to limit test-takers' ability to distort their responses, then it should also lead to higher criterion-related validities relative to testing methods and formats that are more susceptible to faking. However, the literature does not strongly support this line of reasoning (Hough & Oswald, 2000) and recognizing that this is but one single study, neither do the findings presented herein. Despite the comparatively small mean differences between the honest and faking conditions for both SJT formats compared to the single-statement measures, the SJTs generally had smaller criterion-related validities, although these differences were not statistically significant in most instances. So, although on the one hand, the SJTs did not have higher criterion-related validity than the single-statement measures, on the other hand, they also did not have substantially smaller criterion-related validities. Thus, SJTs should provide those involved with testing programs an option that both reduces faking but does not significantly impact the associations between personality constructs and performance. Given the use of supervisor-perspective ratings rather than supervisor ratings, further research should be conducted to examine this conclusion.

Pertaining to the ongoing scholarly debate regarding whether faking can be considered true variance or error variance (Ziegler & Buehner, 2009) in personality score-criterion associations, the results of the present study are more suggestive of the former. As previously noted, if faking is a source of error in the assessment of noncognitive constructs, then any mechanism that prevents faking should ostensibly lead to stronger associations between measures of said constructs and criteria. The failed support for the criterion-related validity hypotheses challenges the error variance notion, as preventing faking did not improve criterion-related validity, and in fact, lowered it in some cases. This pattern of results is more suggestive of true rather than error variance (c.f., Ziegler & Buehner, 2009). Whereas construct-irrelevant variance should typically be avoided, the results—failed support for all ten hypotheses concerning criterion-related validity—suggest that faking may be a substantive individual difference that is in itself valuable to assess. Furthermore, the findings suggest that preventing intentional response distortion via the use of SJTs might not have the intended effect of increasing the criterion-related of validity noncognitive measures when the outcome is contextual performance, which is in line with assertions made by various scholars (e.g., Hough & Oswald, 2000; Viswesvaran & Ones, 1999) who remain unconvinced that faking diminishes the predictive power of noncognitive measures. Accordingly, the most important question might shift from “can we prevent faking” to “should we prevent faking?" That said, from an ethical standpoint, particularly since concerns regarding rank-order changes remain (Griffith et al., 2007), the perspective adopted in this paper is that faking should still be controlled and mitigated as long as the prevention of faking is not accompanied by a decrease in criterion-related validity.

Personality-score SDR correlations. A primary aim of the present study was to replicate and extend Arthur's (2017a) findings that SJTs have weaker correlations with SDR measures than Likert measures. Specifically, Arthur (2017a) obtained SJT-SDR correlations of .07 for agreeableness and .01 for conscientiousness, and correlations of .20 and .28, respectively, for single-statement measures of the same constructs. Although Kasten et al. (2020) did not directly examine the relationship between SDR and scores on noncognitive measures, they obtained correlations similar to Arthur (2017a) between SJT personality measures and a proxy for SDR, self-monitoring (*r* = -.10 - .04). The magnitude of the correlations between the SDR and SJT personality measures obtained by Kasten et al. (2020) and Arthur (2017a) were quite different from those obtained in the faking condition of the present study (*r* = .12 - .15 for the rank SJT and *r* = .40 - .41 for the rate SJT). Furthermore, the correlations were quite large in the honest condition with the SJT-SDR correlations ranging from .41 - .53 for the rate SJT and .20 - .44 for the rank SJT. So, although in the faking condition the personality score-SDR correlations were comparable to Arthur's (2017a) findings, in general, the results of the present study did not fully replicate Arthur's (2017a) or Kasten et al.'s (2020) findings pertaining to the magnitude of the correlations between SJT and SDR measures.

However, it is important to note that across all three studies, different measures of SDR were used, and so the choice of SDR measures warrants further discussion.

The present study used the BIDR whereas Arthur (2017a) used the SDR_{10} (Arthur, 2014) and Kasten et al. (2020) used Graf's (2004) self-monitoring scale. To the extent that these social desirability measures might have different psychometric properties or tend to display different correlations with personality measures, comparisons across studies are more appropriate if the SDR measure is held constant. As such, comparing Arthur et al.'s (2014) findings from Study 2 to the present study are more appropriate, because both studies used the BIDR. This comparison indicates that the personality score-SDR correlations obtained by Arthur et al. (2014; *r* = .13 - .30) were similar to those obtained in the present study (*r* = .12 - .41 in the faking condition, and *r* = .20 - .53 in the honest condition). In summary, for a measure to be considered resistant to faking, it should ideally have no increase in its correlation with an SDR measure across honest and faking conditions. This is what was found for the rank SJT in the present study, providing further evidence of the rank SJT's ability to prevent faking.

Because test-takers' motivation to fake is an important consideration when examining response distortion (Ellingson, 2012), the present study's findings regarding differences in SJT-SDR correlations across the honest and faking condition are not surprising. Those in the faking condition were given explicit instructions to fake, which conceptually has the effect of simulating a high-stakes testing situation. That is, participants should be motivated to achieve higher scores in the faking condition—not because their performance is being assessed for hiring or promotional reasons, but rather because they were *instructed* to do so. In this simulation of the effects of a high-stakes situation, the correlations between the more resilient measure—the SJT in this case—should be substantially lower than measures that are less resistant to faking (e.g., Likert and true-false measures), which is what was observed. This effect, however, is unlikely to hold in honest conditions because test-takers' motivation to distort their scores is already low, and thus the SJT's ability to prevent faking is less important.

Implications for Science and Practice

SJTs vs. single-statement measures. A particularly meaningful finding from the present study is that the SJTs were more resilient to faking than single-statement measures (Likert and true-false). The substantial mean differences between honest and faking conditions for the true-false measures challenges early research showing true-false measures to be quite impervious to faking (Braun & Costantini, 1970; Hoffmann & Nelson, 1971). Notwithstanding those early results, this is not surprising given the commonality between the true-false and Likert formats; instead of using a predetermined scoring key, ratings are summed across items to compute score totals. SJTs should also be less susceptible to faking than other single-statement measures that do not use predetermined scoring keys to score responses, for example, frequency-based personality measures (Edwards & Woehr, 2007). Future research should compare SJTs to other single-statement formats to examine this claim.

In light of the findings of the present study, practitioners should consider whether SJTs have greater utility than single-statement measures, given that SJTs do not appear to have higher criterion-related validity. There are no easy answers to this question, and it will be necessary to consider a variety of factors pertinent to the given context when comparing SJTs and single-statement measures. On the one hand, those administering noncognitive measures in operational employment contexts might opt for SJTs if preventing artificially elevated scores in high-stakes situations is critical to the goals of the assessment program. On the other hand, those administering noncognitive measures for research purposes in low-stakes settings might not consider SJTs to be worthy of the additional development time and effort. Finally, if maximizing criterion-related validity is the prime consideration, then single-statement measures continue to be a strong option.

Smart faker hypothesis. The smart faker hypothesis specifies that the extent to which individuals can distort their responses on noncognitive measures is associated with their cognitive ability, particularly their comprehension-knowledge (MacCann, 2013). The findings of the present study support findings from Kasten et al. (2020) regarding the relationship between response distortion and cognitive ability. That is, response distortion on noncognitive measures appears to be significantly associated with cognitive ability when comparing participants' performance across honest and faking conditions. In the present study, whereas the personality-GMA correlations ranged between .10 and .45 in the honest condition, they ranged from .33 to .69 in the faking condition. Similarly, Kasten et al. (2020) obtained personality-GMA correlations between -.08 and -.04 in the honest condition and between .24 and .25 in the faking condition. Although the SJT measures were effective at preventing faking, the reduction in faking did not neutralize the personality score-GMA associations, as the SJT-GMA and single-statement-GMA correlations did not differ substantially. In addition, the present study did not find substantially larger personality score-GMA correlations for the rank SJT compared to the rate SJT (c.f., Arthur et al., 2014) despite the purported higher cognitive load of SJTs using the rank format (Arthur et al., 2014).

Limitations

Most research scholars would agree that the effectiveness of an experimental study is bound by its ability to make a meaningful distinction between the experimental condition—in this case, the fake "good" condition—and some other comparison condition (e.g., a control condition). Put differently, a failed manipulation can threaten the ability to make strong inferences from the results obtained, and thus, a more detailed analysis of the results of the manipulation check in the present study is informative. To further assess the impact of the manipulation check in the present study, standardized mean differences (*d*s) on noncognitive measures were computed across the honest and faking conditions for participants who passed (*d* = 0.32) and those who failed (*d* = 0.17) the manipulation check. The larger *d* for those who passed suggests that the manipulation check did indeed work as expected for those who were giving adequate attention to completing the measures.

With respect to the large percentage of participants who typically fail attention checks on MTurk, recent empirical evidence suggests that subject pool participants (particularly university students) fail online attention checks at an even higher rate than those on MTurk (Hauser & Schwarz, 2016). This suggests that MTurk is not necessarily a source of low-quality data but the percentage of failures may be a consequence of the online format of the assessments. Overall, the exclusion of those participants who failed attention checks resulted in higher data quality in the present study.

It is also important to consider any impact that the attention check failure rate had on the present study's statistical power. That is, due to a significant reduction in sample size after removing those failing attention checks, the final sample size of 583 participants was much smaller than the 800 needed to achieve a power level of .80 for the postulated a priori effect size based on Kasten et al. (2020). Consequently, a post-hoc power analysis was conducted, revealing that the achieved power of the present study was .76, which was lower than the .80 that had been originally sought.

Furthermore, the exclusion of those failing the attention check had the additional effect of causing the research design to be unbalanced. Instead of 100 participants in each cell of the research design (as illustrated in Figure 2), the final sample was not balanced with respect to testing method and response format. For example, there were 86 participants in the honest condition who completed the true-false measures and 72 in the faking condition who completed the true-false measures (see Table 4 for the number of participants who completed each measure). Despite the reduction in sample size and unbalanced design, Kasten et al.'s (2020) effect size estimate (q = 0.21) was lower than the effect sizes obtained in the present study for the primary hypothesis comparing the mean differences of SJTs and single-statement measures (q = 0.29 for agreeableness, q = 0.41 for conscientiousness). This provides evidence that the reduction in sample size did not significantly impact the ability to answer the specified research questions.

Finally, the present study used a manipulation check, administered after all noncognitive measures were completed, to determine if participants were adequately attending while completing the measures. The manipulation check procedure could have been strengthened by what some have labeled an instructional manipulation check (IMC). An IMC consists of a question that resembles other items embedded within experimental materials. However, instead of participants responding using the standard response format, the IMC instructs them to perform an action to confirm that they understand the instructions (e.g., typing a specified phrase; Oppenheimer, Meyvis, & Davidenko, 2009). Oppenheimer et al. describe IMCs as an effective method to increase statistical power in experiments.

Suggestions for Future Research

While the present study focused on faking without taking the job context into account, there is empirical evidence that individuals employ job-specific patterns of faking on noncognitive measures (Mahar et al., 1995). Future research should replicate the present study’s findings while explicitly including job type as a factor in the study design. Only then can it be determined whether the resilience of SJTs to faking holds when individuals alter their response styles to fit the job for which they are applying. Furthermore, more robust indicators of job performance (e.g., supervisor ratings of performance) should be used in future studies on this topic. After all, there is nothing precluding individuals from also faking self-reported OCBs and CWBs. Indeed, the correlation between SDR-performance ratings was quite strong in the present study; the correlation between CWB-SDR correlation was -.33 and the OCB-SDR correlation was .39.

Researchers should also consider using enhanced technological methods, for example, computer adaptive testing, since scholars have had some success combining the forced-choice response format with the computer adaptive testing format (e.g., TAPAS; Stark et al., 2014). Specifically, researchers should determine whether SJTs that adapt to the test-takers' responses are a viable option to mitigate concerns regarding faking. Scholars have already begun to examine the design of branching SJTs and how branching SJTs impact test-taker reactions (e.g., Reddock, Auer, & Landers, 2020). Future research should also examine the extent to which branching SJTs are susceptible to response distortion. The pairing of SJTs with IRT methods is also worthy of increased scholarly attention.

In addition, researchers may look to a response surface analytical (RSA) approach to determine whether the divergence between faking and honest scores (within-person) is associated with cognitive ability. RSA allows researchers to jointly model the association between two predictors and an outcome or antecedent variable. This could be used to determine if test-takers who improve their standing among test-takers in the faking condition are qualitatively different from those who do not. In terms of explanatory power, RSA has advantages over the difference score or pre-post scoring techniques (Shanock, Baran, Gentry, Pattison, & Heggestad., 2010) that are typically employed in within-person faking studies (e.g., Kasten et al., 2020).

Another potential future research direction is to examine whether the most/least or best/worst response format performs similar to the rank response format in terms of resilience to faking. There is empirical evidence that the best/worst format would also outperform Likert measures in terms of faking resiliency (Arthur et al., 2014). For instance, Rasmussen (2009) found that scores on an SJT using the best/worst response format had stronger associations with performance than a modified SJT measure requiring test-takers to rate response options using a Likert scale. This could be due to differences in resistance to faking for the best/worst response format relative to other formats. Overall, the results obtained in the present study suggests that increased scholarly attention should be given to other response formats to determine which formats perform as well as the rank SJT in mitigating faking concerns.

Conclusion

That the main hypothesis was supported, that is, across response formats, a construct-laden SJT of agreeableness and conscientiousness had smaller mean differences between honest and faking conditions than single-statement measures, contributes to organization science and practice in several ways. First, organizations concerned with faking or socially desirable responding in personality assessment could opt to administer construct-laden SJTs to mitigate these concerns. Furthermore, that the sample consisted of employed individuals (85.76% full-time, 14.24% part-time), strengthens the generalizability of the finding that SJTs are more resilient to faking than single-statement measures. Second, because faking is associated with compromised psychometric properties of personality measures—particularly, changes in the rank ordering of test-takers—the use of measures that are more resistant to faking such as rank SJTs may improve such concerns and lead to more equitable selection programs compared to those that rely on Likert scales for the assessment of personality.

However, the findings of the present study provide critical evidence that, perhaps, faking is not as significant a threat to the criterion-related validity of noncognitive measures as previously thought. Indeed, faking can and often will occur when assessing noncognitive constructs, resulting in mean differences when comparing scores across low-stakes and high-stakes situations. However, to the extent that faking has no effect—or perhaps, even a detrimental one—on personality score-criterion relationships, it is premature to call for a moratorium on single-statement measures and replace them with SJTs when assessing job applicants. Although the SJT was indeed more resilient to attempts to fake, the psychometrics benefits of preventing faking have yet to be established empirically. Considered by some to be a source of measurement error (Ziegler & Buehner, 2009), the findings of the present study are more aligned with the notion that faking may be substantive variance (Hogan et al., 2007; Morgeson et al., 2007); that is, faking is simply a response style used by well-adjusted individuals when they are motivated to do so. Perhaps, then, scholars of organization science and practitioners should shift from studying the prevention, detection, and correction of faking to determining whether those who fake are qualitatively different from those who do not; if they are, it is important that scholars work to construct a coherent theory to explain these differences rather than treating faking as a problem. Such a theory might use Snell’s theory of applicant faking as a foundation (Snell et al., 1999) and should explain the individual differences associated with faking and honest responding styles. Until then, the debate about the importance of faking will likely continue.

REFERENCES

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7,* 32-39.

Arnold, H. J., Feldman, D. C., & Purbhoo, M. (1985). The role of social-desirability response bias in turnover research. *Academy of Management Journal*, *28*, 955-966.

Arthur, W., Jr. (2014). *Development of a social desirability responding measure*. Unpublished manuscript, Texas A&M University, College Station, TX.

Arthur, W., Jr. (2017a). Construct-laden situational judgment tests of personality: Ingenuity or folly? In J. Golubovich, & C. Angiano-Carrasco (Chairs), *Development of and scoring of construct-focused situational judgment tests.* Session presented at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Arthur, W., Jr. (2017b). *An unproctored internet-based test of general mental ability. A validation report.* College Station, TX: Author.

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99,* 535-545.

Arthur, W., Jr., Hagen, E., & George, F., Jr. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior, 8*. Review in advance publication. doi:10.1146/annurev-orgpsych-012420-05532

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *9,* 435-442.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology review*, *11*, 150-166.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, *81*, 261-272.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment*, *13,* 225-232.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.

Bernal, D. (1998). Reducing the effects of individual faking on noncognitive measures: *Let's send those fakers to the bottom of the distribution.* Unpublished dissertation proposal, University of Akron, Akron, OH.

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.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*, 229-258.

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.

Braun, J. R., & Costantini, A. (1970). Faking and faking detection on the Personality Research Form, AA. *Journal of Clinical Psychology, 26,* 516-518.

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

Carpenter, N. C., Newman, D. A., & Arthur, W., Jr., (2020). *What are we measuring? Evaluations of items measuring task performance, organizational citizenship, counterproductive, and withdrawal behaviors.* Manuscript submitted for publication.

Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81,* 694-704.

Cho, I., Payne, S. C., Berry, C. M., & Lee, P., (2020). *Too good to be true: Self-ratings from a supervisor's perspective are not a valid substitute for actual supervisor ratings.* Manuscript submitted for publication.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83-117.

Christiansen, N. D., Edelstein, S., & Fleming, B. (1998). *Reconsidering forced-choice formats for applicant personality assessment.* Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Cronbach, L. J. (1990). *Essentials of psychological Testing*. New York: Harper Collins Publisher, Inc.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

Cucina, J. M., Vasilopoulos, N. L., Su, C., Busciglio, H. H., Cozma, I., DeCostanza, A. H., ... & Shaw, M. N. (2018). The effects of empirical keying of personality measures on faking and criterion-related validity. *Journal of Business and Psychology, 34,* 337-356.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, *34,* 267-285.

de Meijer, L. A., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist, 15*, 229-236.

DeYoung, C. G., Hirsh, J. B., & Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological Science*, *21*, 820-828.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*, 192-203.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology*, *29*, 479-493.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996, August). *The validity of non-cognitive measures decays when applicants fake.* Proceedings of the 56th Annual Meeting of the Academy of Management (Vol. 1996, No. 1, pp. 127-131). Cincinnati, OH.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone *was right! Industrial and Organizational Psychology, 3,* 465-476.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16,* 1-23.

Edwards, B. D., & Woehr, D. J. (2007). An examination and evaluation of frequency-based personality measurement. *Personality and Individual Differences, 43,* 803-814.

Ellingson, J. E. (2012). *People fake only when they need to fake*. In Annual Meeting of the Society for Industrial and Organizational Psychology, April, 2009, New Orleans, LA, US; Oxford University Press.

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

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97,* 866-880.

Fluckinger, C. D., McDaniel, M. A., & Whetzel, D. L. (2008). Review of faking in personnel selection. In M. Mandel (Ed.), *In search of the right personnel* (pp. 90-109). New Delhi, India: McMillian

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.). *Personality psychology in Europe* (pp. 7–28). The Netherlands: Tilburg University Press.

Graf, A. (2004). Eine deutschsprachige Version der Self-Monitoring-Skala. *Zeitschrift für Arbeits-und Organisationspsychologie A&O, 48*, 109-121.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17-41.

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

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341-355.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48,* 400-407.

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.

Hoffmann, H., & Nelson, J. I. (1971). Desirability responses in the Personality Research Form by a sample of alcoholics. *Psychological Reports*, *29*, 559-562.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology*, *88*, 100-112.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future--Remembering the past. *Annual Review of Psychology*, *51*, 631-664.

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

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS) (NPRST-TR-06-2).* Millington, TN: Navy Personnel Research, Studies, and Technology.

Hurd, J. M., Barrett, G. V., Miguel, R. F., Tan, J. A., & Lueke, S. B. (2001, April). *When do response distortion scales reflect faking? A meta-analysis*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Jackson, D. N., Neill, J. A., & Bevan, A. R. (1973). An evaluation of forced-choice and true-false item formats in personality assessment. *Journal of Research in Personality*, *7,* 21-30.

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.

Kasten, N., Freund, P. A., & Staufenbiel, T. (2020). "Sweet little lies": An in-depth analysis of faking behavior on Situational Judgment Tests compared to personality questionnaires. *European Journal of Psychological Assessment, 36,* 136-148.

Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76,* 889-896.

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*, 202-210.

Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using blatant extreme responding for detecting faking in high-stakes selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment, 22,* 371-383.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37,* 426-441.

Lim, B. C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-multimethod examination. *Organizational Research Methods, 9*, 29-54.

MacCann, C. (2013). Instructed faking of the HEXACO reduces facet reliability and involves more Gc than Gf. *Personality and Individual Differences, 55,* 828-833.

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.

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.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60,* 63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683-729.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*, 331-346.

Naber, A. M., Arthur, W., Jr., Edwards, B. D., & Franco-Watkins, A. (2020). *Increased retest scores on cognitive tests: Retrieval or acquisition effects.* Manuscript submitted for publication.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250-260.

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239-250.

Ones, D. S., Viswesvaran, C., & Korbin, W. P. (1995, May). *Meta-analyses of fakability estimates: Between-subjects versus within-subjects designs.* Paper presented at 10th annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*, 1-29.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867-872.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*, 562–571.

Paulhus, D. L. 1988. *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding.* Unpublished manuscript, University of British Columbia, Vancouver, British Columbia.

Paulhus, D. L. (1991). Balanced Inventory of Desirable Responding (BIDR) reference manual for version 6. (Manual available from author at Department of Psychology. University of British Colombia, Vancouver, B. C., Canada V6TIY7.).

Paulhus, D. L., & Bruce, M. N. (1991, August). *Faking job profiles.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin, 21,* 100-108.

Pauls, C. A., & Crost, N. W. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences, 26,* 194-206.

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

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), *SIOP* organizational series. *Situational judgment tests: Theory, measurement, and application* (pp. 83-105). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Rasmussen, J. L. (2009). *Situational judgment test responding: Best and worst or rate each response.* Master's thesis, Texas A&M University.

Reddock, C. M., Auer, E. M., & Landers, R. N. (2020). A theory of branched situational judgment tests and their applicant reactions. *Journal of Managerial Psychology, 35*, 225-270.

Reeve, C. L., Heggestad, E. D., & George, E. (2005). Estimation of transient error in cognitive ability scales. *International Journal of Selection and Assessment, 13,* 316-320.

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance,* 21, 89-106.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634-644.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management, 41,* 149-174.

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.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 91,* 613-621.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T. Y. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88,* 979-988.

Schoorman, F. D., & Mayer, R. C. (2008). The value of common perspectives in self-reported appraisals: You get what you ask for. *Organizational Research Methods, 11*, 148-159.

Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology, 25,* 543-554.

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.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

Stanush, P. L. (1997). *Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation* (Doctoral dissertation, Texas A&M University, 1997). Dissertations Abstracts International, Section B: The Sciences and Engineering, 58, 2167.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multi-unidimensional paired comparison responses* (Doctoral dissertation, ProQuest Information & Learning, 2002). Dissertation Abstracts International: Section B: The Sciences and Engineering, 63, 1084.

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (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.

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

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist*, *50,* 912-927.

Stewart, G. L., Darnold, T. C., Zimmerman, R. D., Parks, L., & Dustin, S. L. (2010). Exploring how response distortion of personality measures affects individuals. *Personality and Individual Differences, 49,* 622-628.

Teng, Y., Brannick, M. T., & Borman, W. C. (2020). Capturing resilience in context: Development and validation of a situational judgment test of resilience. *Human Performance*. Advance online publication. doi:10.1080/08959285.2019.1709069

Tsaousis, I., & Nikolaou, I. E. (2001). The stability of the Five-Factor Model of personality in personnel selection and assessment in Greece. *International Journal of Selection and Assessment, 9*, 290-301.

Underhill, C. M., Bearden, R. M., & Chen, H. T. (2008). *Evaluation of the fake resistance of a forced-choice paired-comparison of computer adaptive personality measures.* Millington, TN: Navy Personnel Research, Studies, and Technology (NPRST-TR-08-2)

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*, 243-262.

Vecchione, M., Dentale, F., Alessandri, G., & Barbaranelli, C. (2014). Fakability of implicit and explicit measures of the Big Five: Research findings from organizational settings. *International Journal of Selection and Assessment, 22,* 211-218.

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

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2003, April). *Personality and situational judgment tests across applicant and incumbent settings.* Paper presented at 18th annual conference of the Society for Industrial Organizational Psychology, Orlando, US.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291-309.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

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

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

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69,* 548-565.

APPENDIX

AGREEABLENESS AND CONSCIENTIOUSNESS: LIKERT (RATE RESPONSE FORMAT)

HONEST DIRECTIONS

Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself in relation to other people you know of the same sex and roughly the same age as you. Please read each statement carefully, and then rate the extent to which it accurately describes you.

FAKING DIRECTIONS

Listed below are phrases describing people's behaviors. **Imagine that you are applying for a job that you really want. When completing the following items, present yourself in the most favorable light. Please read each statement carefully, and then respond in a manner that will maximize your chances of being hired.**

①	②	③	④	⑤
Very inaccurate	Inaccurate	Neither inaccurate nor accurate	Accurate	Very accurate

1.	Have a soft heart.	① ② ③ ④ ⑤
2.	Am always prepared.	① ② ③ ④ ⑤
3.	Sympathize with others' feelings.	① ② ③ ④ ⑤
4.	Get chores done right away.	① ② ③ ④ ⑤
5.	Feel others' emotions.	① ② ③ ④ ⑤
6.	Make a mess of things.	① ② ③ ④ ⑤
7.	Am not really interested in others.	① ② ③ ④ ⑤
8.	Am exacting in my work.	① ② ③ ④ ⑤
9.	Feel little concern for others.	① ② ③ ④ ⑤
10.	Like order.	① ② ③ ④ ⑤
11.	Make people feel at ease.	① ② ③ ④ ⑤

12.	Leave my belongings around.	① ② ③ ④ ⑤
13.	Am not interested in other people's problems.	① ② ③ ④ ⑤
14.	Pay attention to details.	① ② ③ ④ ⑤
15.	Take time out for others.	① ② ③ ④ ⑤
16.	Shirk my duties.	① ② ③ ④ ⑤
17.	Insult people.	① ② ③ ④ ⑤
18.	Follow a schedule.	① ② ③ ④ ⑤
19.	Am interested in people.	① ② ③ ④ ⑤
20.	Often forget to put things back in their proper place.	① ② ③ ④ ⑤
21.	Inquire about others' well-being.	① ② ③ ④ ⑤
22.	Do things according to a plan.	① ② ③ ④ ⑤
23.	Am hard to get to know.	① ② ③ ④ ⑤
24.	Continue until everything is perfect.	① ② ③ ④ ⑤
25.	Know how to comfort others.	① ② ③ ④ ⑤
26.	Neglect my duties.	① ② ③ ④ ⑤
27.	Love children.	① ② ③ ④ ⑤
28.	Make plans and stick to them.	① ② ③ ④ ⑤
29.	Am indifferent to the feelings of others.	① ② ③ ④ ⑤
30.	Love order and regularity.	① ② ③ ④ ⑤
31.	Am on good terms with nearly everyone.	① ② ③ ④ ⑤
32.	Waste my time.	① ② ③ ④ ⑤
33.	Have a good word for everyone.	① ② ③ ④ ⑤
34.	Like to tidy up.	① ② ③ ④ ⑤
35.	Show my gratitude.	① ② ③ ④ ⑤
36.	Do things in a half-way manner.	① ② ③ ④ ⑤
37.	Think of others first.	① ② ③ ④ ⑤
38.	Find it difficult to get down to work.	① ② ③ ④ ⑤
39.	Love to help others.	① ② ③ ④ ⑤
40.	Leave a mess in my room.	① ② ③ ④ ⑤

Note. Items 6, 7, 9, 12, 13, 16, 17, 20, 23, 26, 29, 32, 36, 38, and 40 are reverse-coded.

AGREEABLENESS AND CONSCIENTIOUSNESS: SINGLE-STATEMENT

(TRUE-FALSE RESPONSE FORMAT)

HONEST DIRECTIONS

Listed below are phrases describing people's behaviors. Please use the scale provided below to determine whether each statement describes you. Compare yourself with the statement in terms of how you generally are now, not as you wish to be in the future. Respond as you honestly see yourself in relation to other people you know of the same sex and roughly the same age as you. Please read each statement carefully, and then respond true if it accurately describes you or false if it does not.

FAKING DIRECTIONS

Listed below are phrases describing people's behaviors. **Imagine that you are applying for a job that you really want. When completing the following items, present yourself in the most favorable light. Please read each statement carefully, and then respond true or false in a manner that will maximize your chances of being hired.**

Ⓣ	Ⓕ
True	False

1.	Have a soft heart.	Ⓣ	Ⓕ
2.	Am always prepared.	Ⓣ	Ⓕ
3.	Sympathize with others' feelings.	Ⓣ	Ⓕ
4.	Get chores done right away.	Ⓣ	Ⓕ
5.	Feel others' emotions.	Ⓣ	Ⓕ
6.	Make a mess of things.	Ⓣ	Ⓕ
7.	Am not really interested in others.	Ⓣ	Ⓕ
8.	Am exacting in my work.	Ⓣ	Ⓕ
9.	Feel little concern for others.	Ⓣ	Ⓕ
10.	Like order.	Ⓣ	Ⓕ
11.	Make people feel at ease.	Ⓣ	Ⓕ
12.	Leave my belongings around.	Ⓣ	Ⓕ
13.	Am not interested in other people's problems.	Ⓣ	Ⓕ

14.	Pay attention to details.	Ⓣ Ⓕ
15.	Take time out for others.	Ⓣ Ⓕ
16.	Shirk my duties.	Ⓣ Ⓕ
17.	Insult people.	Ⓣ Ⓕ
18.	Follow a schedule.	Ⓣ Ⓕ
19.	Am interested in people.	Ⓣ Ⓕ
20.	Often forget to put things back in their proper place.	Ⓣ Ⓕ
21.	Inquire about others' well-being.	Ⓣ Ⓕ
22.	Do things according to a plan.	Ⓣ Ⓕ
23.	Am hard to get to know.	Ⓣ Ⓕ
24.	Continue until everything is perfect.	Ⓣ Ⓕ
25.	Know how to comfort others.	Ⓣ Ⓕ
26.	Neglect my duties.	Ⓣ Ⓕ
27.	Love children.	Ⓣ Ⓕ
28.	Make plans and stick to them.	Ⓣ Ⓕ
29.	Am indifferent to the feelings of others.	Ⓣ Ⓕ
30.	Love order and regularity.	Ⓣ Ⓕ
31.	Am on good terms with nearly everyone.	Ⓣ Ⓕ
32.	Waste my time.	Ⓣ Ⓕ
33.	Have a good word for everyone.	Ⓣ Ⓕ
34.	Like to tidy up.	Ⓣ Ⓕ
35.	Show my gratitude.	Ⓣ Ⓕ
36.	Do things in a half-way manner.	Ⓣ Ⓕ
37.	Think of others first.	Ⓣ Ⓕ
38.	Find it difficult to get down to work.	Ⓣ Ⓕ
39.	Love to help others.	Ⓣ Ⓕ
40.	Leave a mess in my room.	Ⓣ Ⓕ

Note. Items 6, 7, 9, 12, 13, 16, 17, 20, 23, 26, 29, 32, 36, 38, and 40 are reverse-coded.

ORGANIZATIONAL CITIZENSHIP BEHAVIORS

DIRECTIONS: How would your supervisor rate you on the following job duties? Even if you disagree with how your supervisor would rate you, please use the scales presented below to rate the degree to which your supervisor would agree with the following statements about you.

①	②	③	④	⑤
Strongly Disagree	**Moderately Disagree**	**Neither Disagree nor Agree**	**Moderately Agree**	**Strongly Agree**

1.	Assists supervisor with his or her work (when not asked)	① ② ③ ④ ⑤
2.	Takes time to listen to coworkers' problems and worries	① ② ③ ④ ⑤
3.	Goes out of way to help new employees	① ② ③ ④ ⑤
4.	Takes a personal interest in other employees	① ② ③ ④ ⑤
5.	Conserves and protects organizational property	① ② ③ ④ ⑤
6.	Attendance at work is above the norm	① ② ③ ④ ⑤
7.	Helps others who have heavy work loads	① ② ③ ④ ⑤
8.	Gives advance notice if unable to come to work	① ② ③ ④ ⑤
9.	Helps others who have been absent	① ② ③ ④ ⑤
10.	Adheres to informal rules devised to maintain order	① ② ③ ④ ⑤
11.	Passes along information to co-workers	① ② ③ ④ ⑤

COUNTERPRODUCTIVE WORK BEHAVIORS

DIRECTIONS: How would your supervisor rate you on the following behaviors? Even if you disagree with how your supervisor would rate you, please use the scales presented below to rate how often your supervisor would say that you engage in the following behaviors at work.

①	②	③	④	⑤
Never	**Once or Twice**	**Once or Twice Per Month**	**Once or Twice Per Week**	**Every Day**

1.	Complains about insignificant things at work	① ② ③ ④ ⑤
2.	Makes fun of someone at work	① ② ③ ④ ⑤
3.	Says something hurtful to someone at work	① ② ③ ④ ⑤
4.	Makes an ethnic, religious, or racial remark at work	① ② ③ ④ ⑤
5.	Curses at someone at work	① ② ③ ④ ⑤
6.	Plays a mean prank on someone at work	① ② ③ ④ ⑤
7.	Acts rudely toward someone at work	① ② ③ ④ ⑤
8.	Takes property from work without permission	① ② ③ ④ ⑤
9.	Takes an additional or longer break than is acceptable at your workplace	① ② ③ ④ ⑤
10.	Comes in late to work without permission	① ② ③ ④ ⑤
11.	Neglects to follow his/her supervisor's instructions	① ② ③ ④ ⑤
12.	Intentionally works slower than he/she could have worked	① ② ③ ④ ⑤
13.	Uses an illegal drug or consumes alcohol on the job	① ② ③ ④ ⑤
14.	Puts little effort into his/her work	① ② ③ ④ ⑤
15.	Drags out work in order to get overtime	① ② ③ ④ ⑤
16.	Does poor quality work	① ② ③ ④ ⑤
17.	Uses equipment for personal purposes without permission.	① ② ③ ④ ⑤

FIVE-FACTOR MODEL OF PERSONALITY: SITUATIONAL JUDGMENT TEST SAMPLE ITEMS

HONEST INSTRUCTIONS

This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. Your task is to rate each alternative in terms of its effectiveness as a response to the given scenario.

FAKING INSTRUCTIONS

This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. **<u>Imagine that you are applying for a job that you really want. Your task is to rate each alternative in terms of its effectiveness to the given scenario in a way that presents yourself in the most favorable light in order to maximize your chances of being hired.</u>**

①	②	③	④	⑤
Very Ineffective	**Ineffective**	**Neither Effective nor Ineffective**	**Effective**	**Very Effective**

1. You and your coworker, Alex, have a meeting scheduled to take care of the next steps on a project. However, Alex has rescheduled the meeting twice in the last week. You have both agreed to meet today to discuss the project but Alex has just rescheduled the meeting again. What would you do?

A.	Send Alex an email letting her know that the project needs to be completed in a timely manner.	① ② ③ ④ ⑤
B.	Ask Alex for a new meeting date that works for her.	① ② ③ ④ ⑤
C.	Send Alex an email letting her know how unprofessional this behavior is.	① ② ③ ④ ⑤
D.	Ask to speak to Alex for a few moments with the intent of professionally reminding her about the importance of keeping commitments and respecting each other’s time.	① ② ③ ④ ⑤

2. You have to be at work by 7:30 a.m. It is typically a 20 minute drive to work. You leave your apartment at 7:00 a.m., go down to start your car, and it does not start. What would you do?

A.	Call your supervisor and tell her that you are having car trouble and will not be able to make it to work today.	① ② ③ ④ ⑤

B.	Try and track down a friend and ask him to give you a ride to work.	① ② ③ ④ ⑤
C.	Contact your supervisor and inform her you may be late, and will be there as soon as possible.	① ② ③ ④ ⑤
D.	Try and fix the car yourself and if unsuccessful, then call your supervisor and inform her you will be late to work.	① ② ③ ④ ⑤

MANIPULATION CHECK (RATE LIKERT)

When you took the personality test, which of the two instructions were you given?

☐ Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself in relation to other people you know of the same sex and roughly the same age as you. Please read each statement carefully, and then rate the extent to which it accurately describes you.

☐ Listed below are phrases describing people's behaviors. **Imagine that you are applying for a job that you really want. When completing the following items, present yourself in the most favorable light. Please read each statement carefully, and then respond in a manner that will maximize your chances of being hired.**

MANIPULATION CHECK (TRUE-FALSE)

When you took the personality test, which of the two instructions were you given?

☐ Listed below are phrases describing people's behaviors. Please use the scale provided below to determine whether each statement describes you. Compare yourself with the statement in terms of how you generally are now, not as you wish to be in the future. Respond as you honestly see yourself in relation to other people you know of the same sex and roughly the same age as you. Please read each statement carefully, and then respond true if it accurately describes you or false if it does not.

☐ Listed below are phrases describing people's behaviors. **Imagine that you are applying for a job that you really want. When completing the following items, present yourself in the most favorable light. Please read each statement carefully, and then respond true or false in a manner that will maximize your chances of being hired.**

MANIPULATION CHECK (RATE SJT)

When you took the personality test, which of the two instructions were you given?

☐ This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. Your task is to rate each alternative in terms of its effectiveness as a response to the given scenario.

☐ This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. **Imagine that you are applying for a job that you really want.** **Your task is to rate each alternative in terms of its effectiveness to the given scenario in a way that presents yourself in the most favorable light in order to maximize your chances of being hired.**

MANIPULATION CHECK (RANK SJT)

When you took the personality test, which of the two instructions were you given?

☐ This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. Your task is to rank each alternative in terms of its effectiveness as a response to the given scenario.

☐ This measure consists of 14 scenarios. For each scenario, you will be presented with 4 alternatives. **Imagine that you are applying for a job that you really want.** **Your task is to rank each alternative in terms of its effectiveness to the given scenario in a way that presents yourself in the most favorable light in order to maximize your chances of being hired.**

BALANCED INVENTORY OF DESIRABLE RESPONDING

Using the scale below as a guide, write a number beside each statement to indicate how true it is.

1 ----------- 2 ----------- 3 ----------- 4 ----------- 5 ----------- 6 -----------7
Not True Somewhat True Very True

____ 1. My first impressions of people usually turn out to be right.
____ 2. It would be hard for me to break any of my bad habits.
____ 3. I don't care to know what other people really think of me.
____ 4. I have not always been honest with myself.
____ 5. I always know why I like things.
____ 6. When my emotions are aroused, it biases my thinking.
____ 7. Once I've made up my mind, other people can seldom change my opinion.
____ 8. I am not a safe driver when I exceed the speed limit.
____ 9. I am fully in control of my own fate.
____ 10. It's hard for me to shut off a disturbing thought.
____ 11. I never regret my decisions.
____ 12. I sometimes lose out on things because I can't make up my mind soon enough.
____ 13. The reason I vote is because my vote can make a difference.
____ 14. My parents were not always fair when they punished me.
____ 15. I am a completely rational person.
____ 16. I rarely appreciate criticism.
____ 17. I am very confident of my judgments
____ 18. I have sometimes doubted my ability as a lover.
____ 19. It's all right with me if some people happen to dislike me.
____ 20. I don't always know the reasons why I do the things I do.

____ 21. I sometimes tell lies if I have to.
____ 22. I never cover up my mistakes.
____ 23. There have been occasions when I have taken advantage of someone.
____ 24. I never swear.
____ 25. I sometimes try to get even rather than forgive and forget.
____ 26. I always obey laws, even if I'm unlikely to get caught.
____ 27. I have said something bad about a friend behind his/her back.
____ 28. When I hear people talking privately, I avoid listening.
____ 29. I have received too much change from a salesperson without telling him or her.
____ 30. I always declare everything at customs.
____ 31. When I was young I sometimes stole things.
____ 32. I have never dropped litter on the street.
____ 33. I sometimes drive faster than the speed limit.
____ 34. I never read sexy books or magazines.
____ 35. I have done things that I don't tell other people about.
____ 36. I never take things that don't belong to me.
____ 37. I have taken sick-leave from work or school even though I wasn't really sick.
____ 38. I have never damaged a library book or store merchandise without reporting it.
____ 39. I have some pretty awful habits.
____ 40. I don't gossip about other people's business.
Self-deception = 1-20. Impression management = 21-40.

GENERAL MENTAL ABILITY TEST
SAMPLE ITEMS

This is a 10-minute timed test and because it is timed, it requires uninterrupted time to complete it. Once you start this test, you will be unable to pause it.

There are a total of 60 items, but the test will probably be too long for you to finish. However, complete as many items as you can in the allotted time. Work quickly and accurately. Do not spend too much time on any one item. Your score will be the number of items that you answer correctly.

You may also want to have scratch paper and a pen or pencil ready before you start since some of the problems you will encounter may require some "figuring out". Please do **not** use a calculator, a dictionary, or any other aid.

1. What is 15% of 200?
   - 20
   - 30
   - 45
   - 50

2. BOOK is to CHAPTER as ORGANIZATION is to
   - Corporation
   - Department
   - Bureaucracy
   - Regulation

3. Which of the following words is different from the others?
   - Minute
   - Small
   - Moderate
   - Diminutive

DEMOGRAPHICS

Is English your primary language?
☐ Yes
☐ No

What is your age?
☐ Under 18
☐ 18-40
☐ 41-65
☐ Over 65

Are you currently employed?
☐ Yes - full time
☐ Yes - part time
☐ No

What is your sex:
☐ Male
☐ Female
☐ Other (please specify): ___________________________

What is your age in years? ___________________

Race and Ethnicity:
☐ Hispanic or Latino
☐ White (Not Hispanic or Latino)
☐ Black or African American (Not Hispanic or Latino)
☐ Native Hawaiian or Other Pacific Islander (Not Hispanic or Latino)
☐ Asian (Not Hispanic or Latino)
☐ American Indian or Alaska Native (Not Hispanic or Latino)
☐ Two or More Races (Not Hispanic of Latino)
☐ Other (please specify): ___________________________

Highest Education Earned:
Degree: ▾
Status: ▾

Degree (1st drop down)	Status (2nd drop down)
High School	
	GED
	9th grade
	10th grade
	11th grade
	12th grade
	Completed
Technical/Vocational School	
	Completed
	In-progress
Associate’s Degree [2-year]	
	Completed
	In-progress
Bachelor’s Degree [4-year]	
	Freshman
	Sophomore
	Junior
	Senior
	Completed
Master’s Degree	
	Completed
	In-progress
Ph.D.	
	Completed
	In-progress

If employed, how long have you been in your current organization in years__________and months ____________?