Masthead Logo

Via Sapientiae:
The Institutional Repository at DePaul University

College of Science and Health Theses and Dissertations

College of Science and Health

6-13-2014

Interventions for Addressing Faking on Personality Assessments for Employee Selection: A Meta-Analysis

Christopher Adair
DePaul University, CKADAIR@GMAIL.COM

Recommended Citation
Adair, Christopher, "Interventions for Addressing Faking on Personality Assessments for Employee Selection: A Meta-Analysis" (2014). *College of Science and Health Theses and Dissertations*. 93.
https://via.library.depaul.edu/csh_etd/93

This Dissertation is brought to you for free and open access by the College of Science and Health at Via Sapientiae. It has been accepted for inclusion in College of Science and Health Theses and Dissertations by an authorized administrator of Via Sapientiae. For more information, please contact wsulliv6@depaul.edu, c.mcclure@depaul.edu.

Interventions for Addressing Faking on Personality Assessments for Employee Selection: A Meta-Analysis

A Dissertation

Presented in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

By

Christopher Kenny Adair

June, 2014

Department of Psychology

College of Science and Health

DePaul University

Chicago, Illinois

Dissertation Committee

Suzanne T. Bell, Ph.D., Chairperson

Jane Halpert, Ph.D.

Alice Stuhlmacher, Ph.D.

Charles Naquin, Ph.D.

Alexandra Murphy, Ph.D.

Acknowledgments

I would like to express my deepest appreciation to my dissertation chair Dr. Suzanne T. Bell for her continued guidance and expertise throughout this process. I would also like to thank my committee members. Thank you to Dr. Alice Stuhlmacher and Dr. Jane Halpert for their assistance not only in this paper but throughout my graduate career. Thank you to Dr. Charles Naquin and Dr. Alexandra Murphy for their contributions to the quality of this project. Thank you to Daniel Abben for his assistance in coding the studies included in this meta-analysis. Thank you to my parents, Bill and Mary, and my sister, Katie, for being a constant source of support. I would also like to thank my wonderful fiancé, Sara, for keeping me sane and calm over the last few months. I could not have done this without you. And finally to my dog, Enzi, who reminds me that an afternoon lying in the grass in the park is often the best place to think.

Biography

The author was born in Kansas City, Missouri on May 23, 1985. He graduated from Rockhurst High School in 2003, received his Bachelor of Arts degree from Saint Louis University in 2007, and a Master of Arts degree from DePaul University in 2011.

Table of Contents

List of Tables

Table 11

List of Figures

Abstract

It is common practice to administer personality assessments in personnel selection due to their ability to cost-effectively predict organizationally relevant criteria with relatively small subgroup differences. However, concerns are often raised about test-taker response bias. The proposed research focuses on one issue related to personality test accuracy, namely faking. Also called response distortion or inflation, faking represents a multidimensional behavior that is both intentional and deceptive and seeks to benefit one's own interests. The current study uses the Theory of Planned Behavior (Ajzen, 1985) and expectancy theories of motivation (e.g., Vroom, 1964) as a theoretical basis for understanding faking. Prevalence estimates vary, but common estimates state that around 30% of applicants can be classified as fakers. Faking on personality assessments can influence response scores, the ability to make valid inferences from these scores, and even selection decisions. The utility of the selection system is critically undermined to the extent that any or all of these factors are altered.

Given the prevalence of faking, research is focused on preventing faking on personality assessments and/or reducing its negative impact on organizational decision-making. This dissertation uses meta-analysis to evaluate the efficacy of different faking interventions. There are two main categories of interventions: preventive and remedial. Remedial interventions are focused on altering the interpretations or decisions made from personality scores after test-taker data has been collected. Preventive strategies, on the other hand, seek to limit faking before the behavior occurs. For instance, warnings seek to limit faking intention while time limits, forced choice formats, and decreased item transparency seek to

limit faking ability. Meta-analysis will be used to quantitatively aggregate the results of multiple primary studies. Meta-analysis can be used to test relationships not addressed in the primary study, and can provide summary statements about effects observed in the largely disjointed faking literature. Two meta-analyses were conducted in order to better understand the effectiveness of faking interventions. First, effect size estimates of the difference between personality scale mean scores (i.e., sample-weighted *d*'s) across intervention conditions was conducted. Second, meta-analysis of the relationships between personality traits and performance outcomes (i.e., correlation) was conducted in order to estimate the criterion-related validities of personality traits across different intervention conditions.

Results suggest that interventions are generally effective at reducing faking behavior, evidenced by smaller sample-weighted mean *d*'s for studies with a faking intervention compared to those without any intervention. Warnings are generally more effective than forced-choice or item transparency interventions at reducing faking behavior. Randomizing items, on the other hand, does little to influence faking. Although based on a limited primary studies, the criterion-related validity of personality scores on performance outcomes were not enhanced due to the presence of a faking intervention. Taken together, these results suggest that faking interventions may influence observed personality scores but did not seem to influence the ability to make valid inferences based on the scores.

Introduction

Personality assessments have regained popularity in Industrial-Organizational (I-O) psychology within the last 25 years (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) and are now utilized for selection purposes by over 40% of Fortune 100 companies (Rothstein & Goffin, 2006). Beyond their usefulness for predicting organizationally-relevant criteria (e.g., Hurtz & Donovan, 2000), personality assessments are attractive selection tools because they typically have smaller subgroup differences than other personnel selection tools (Hough & Oswald, 2008; Hough, Oswald, & Ployhart, 2000). However, the ability for these assessments to be faked raises concerns about the measures' usefulness (Morgeson et al., 2007a). The purpose of this dissertation is to examine the effectiveness of various approaches designed to reduce faking on personality assessments when used for employee selection.

Personality Assessments and Personnel Selection

The general consensus on a taxonomy of personality traits, namely the Five Factor Model (FFM; McCrae & John, 1992), was a main catalyst for a renewed interest in personality testing that started in the 1980's. Although there are several theories with regards to how to personality dimensionality, most current personality research and practice is based on the five dimensions of openness to experience (openness), conscientiousness, extraversion, agreeableness, and emotional stability. The current research focuses on these five dimensions because they are the most widely researched, and other personality frameworks can be mapped onto the FFM. Openness is characterized by a healthy curiosity about the world and an interest in new intellectual or imaginative

experiences. People with high levels of openness are often described as curious, artistic, or insightful. Conscientious individuals are characterized by self-control, persistence, and focus on details. Individuals with high levels of conscientiousness are often described as efficient, organized, and achievement striving. Extraverted individuals are driven by the need to seek social stimulation. Individuals with high levels of extraversion are often described as assertive, gregarious, or outgoing. Agreeable individuals are characterized by a tendency to be friendly and cooperative. People with high levels of agreeableness are often described as appreciative, kind, or accommodating. Finally, emotionally stable individuals tend to have solid control over their emotions and are relatively resilient to negative cues in their environment. Individuals with low levels of emotional stability are often described as anxious, irritable, or stressed.

Questioning assumptions of personality assessments. The practice of administering personality assessments for personnel selection has two strong assumptions. First, by requiring applicants to complete a personality test, administrators (e.g., employers) assume that the measure accurately assesses the true, underlying disposition. Thus, a test-taker that scores high on a measure of conscientiousness truly possesses that trait. Research supports this assumption. Widely used personality measures tend to demonstrate strong psychometric properties across a variety of settings (*cf*. Hough & Ones, 2002), and there is significant cross-observer agreement across the five factors (Connelly & Ones, 2010; McCrae & Costa, 1987). The second assumption is that these scores serve as useful predictors of future behavior or performance. This assumption tends to

be supported for organizational outcomes such as effectiveness and performance (e.g., Barrick & Mount; 1991; Driskell, Hogan, & Salas, 1987; Morgeson et al. 2007a), organizational citizenship behaviors (Chiaburu, Oh, Berry, Li, & Gardner, 2011), and organizational commitment (Erdheim, Wang, & Zickar, 2011). However, the validity coefficients tend to be modest across these various criteria.

Despite some support for these assumptions across a variety of contexts, both researchers and practitioners have reason to doubt the viability of these two assertions. The properties of self-report measures are a major factor driving this doubt. Self-report assessments, the most common manner of collecting personality data, are limited in part due to their susceptibility to positive response distortion (Dunning, Heath, & Suls, 2004). In other words, test-takers can appear more externally desirable than they truly are. The result of response distortion (or, faking) may directly undermine the aforementioned assumptions. Although not all researchers believe that faking has these effects (e.g., Morgeson et al., 2007a), a considerable set of research suggests there may be cause for concern (e.g., Tett & Christiansen, 2007). Participants may be able to fake their responses in such a way that the test no longer accurately predicts the criterion of interest (*cf*. Rothstein & Goffin, 2006). Also, the mean scores of traits may be altered to such an extent that the tests are no longer able to differentiate test-takers (e.g., Ones, Viswesvaran, & Schmidt, 1996).

As such, there is a great amount of dissent in the literature regarding the utility of personality assessments for selection. Some researchers argue for the abandonment of these tests based on the lack of confidence in the scores (e.g.,

Morgeson et al., 2007a), while others argued that faking tends to be uniform across applicants and thus practical decisions such as hiring decisions are not adversely influenced (Morgeson et al., 2007b). Neither of these is desirable, as both fail to actually address the issue of faking.

The core assumptions of personality testing are questioned based on the expectation that test-takers fake as long as there is motivation or incentive to do so. Moreover, important hiring decisions can be influenced (e.g., Rosse, Stecher, Miller, & Levin, 1998), and these subpar hires can end up imposing substantial costs onto the organization. Depending on various factors such as the proportion of fakers in the sample and the selection ratio, faking can cost organizations over $2,000 per worker per year (Komar, Brown, Komar, & Robie, 2008). Beyond affecting valid inferences from test scores and the rank order of applicants, applicants may react negatively if they believe others have faked on an assessment (Converse et al., 2008). This can have a meaningful impact on the organization, as negative reactions during the selection process can result in lower test-taking motivation and an increase of applicants who self-select out of the process (Ployhart, McFarland, & Ryan, 2002). Due to these potential consequences, organizations should focus their efforts on how to mitigate the negative influence of faking. Research in this area has strong implications for practitioners, and will also drive future research in the field by shining a light on the more efficacious interventions for addressing faking.

Theoretical Perspectives on Faking

Faking is often regarded as a volitional behavior used by a test-taker to improve the likelihood of attaining desired outcomes. There are various intrapersonal and situational factors that influence the test-taker's decision to fake or not. A review of the theoretical perspectives on faking and its determining factors will offer a useful framework for understanding how interventions seek to influence faking behavior.

Social desirability as a theory of faking. Social desirability (SD), or the tendency for test-takers to respond in a way that they feel is externally or socially valued (Paulhus, 1991; Smith & Ellingson, 2002), is the most common theory applied to faking. Commonly used measures of SD such as the Balanced Inventory of Desirable Responding (BIDR) include two SD dimensions: self-deceptive positivity and impression management (Paulhus, 1984). Empirical tests of the scales demonstrate appropriate levels of discriminant validity to treat them as distinct constructs (e.g., Paulhus & Reid, 1991).

Self-deceptive enhancement is considered to be a more "natural" form of response distortion, as high scores do not represent a deliberate attempt to deceive test administrators. Instead, self-deceptive enhancement operates through the unconscious, positive biases individuals hold when evaluating themselves (Bing, Kluemper, Davison, Taylor, & Novicevic, 2011). While it still presents the test taker in an overly-positive light (i.e., not fully accurate), this form of response distortion is less of a focus for test administrators because it is done unconsciously. In other words, unconscious distortion will likely always be

present in any self-report test (Meehl & Hathaway, 1946; Quist, Merlini, & Griffith), and because it is done unconsciously is more difficult to correct. Moreover, it may be a partial extension of self-efficacy (Barrick & Mount, 1996).

Test administrators are more concerned with conscious distortion behaviors, as opposed to unconscious distortion. The impression management dimension of SD involves deliberate alteration of responses in order for test-takers to create an artificially positive image. A test taker engaging in impression management consciously responds to the situational demands and motives in a way that seeks to enhance the interpretation of the responses compared to others. Impression management is especially prevalent when the test taker is motivated to present one's self positively, such as in the case of applying for a job. In fact, Paulhus (1984) demonstrated that scores on impression management, and not self-deception, tended to increase from more to less verifiable settings. This implies that impression management scores can be consciously manipulated, because impression management, not self-deception, scores were influenced by the situational demands. High impression management scores also tend to be strongly associated with traditional lie scales used in personality test, such as the MMPI lie scale (Barrick & Mount, 1996), and other distortion measures (Quist, Arora, & Griffith, 2007).

However, there are two main reasons why social desirability and faking should not be considered isomorphic. First, SD has substantial overlap with substantive personality traits. Impression management in particular shares meaningful variance with traits such as conscientiousness and emotional stability

(Barrick & Mount, 1996; Li & Bagger, 2006; Ones, Viswesvaran, & Reiss, 1996). It is thus difficult to identify whether SD is indicative of a faking behavior or a disposition to fake. Knowing an individual's disposition to fake is certainly useful, but it is not equivalent to faking behavior. Second, SD cannot adequately address the complexity of faking behavior. Impression management is often used as a part of faking models (e.g., Mueller-Hanson, Heggestad, & Thornton, 2006), but it does not fully represent the construct. For instance, a test taker's efficacy of positive self-presentation was found to be more reflective of faking than impression management (Pauls & Crost, 2005). In other words, attitudes and beliefs regarding one's ability to fake may be as or more important for determining faking behavior than dispositional tendencies to fake. These two issues suggest that although SD provides a useful starting point, it does not offer a complete representation of faking.

Theory of Planned Behavior. The main limitation of SD as a theory for faking is that it represents an overly simplified description of a complex behavior. Ajzen's (1985; 1991) theory of planned behavior (TPB) is more complete as it includes crucial antecedents. TPB represents a meditational model from cognition to behavior. According to this theory, three behavior-specific cognitions (i.e., attitudes, social norms, and perceived behavioral control) influence behavior through individual intentions. Within the context of faking, test-takers looking to engage in faking behavior must first have cognitive dispositions that influence their *intent* to distort their responses (Yu, 2008). Indeed, TPB is used in many

studies to predict a range of behaviors from theft (Beck & Ajzen, 1991) to recycling (Boldero, 1995).

Each of the cognitions plays an important role in predicting intention and behavior. Attitudes represent an evaluative appraisal of the behavior in question. More positive attitudes suggest that an individual is more likely to engage in the specified behavior. Social norms operate on a more external level than attitudes. Instead of an internal evaluation, the presence of social pressure drives an individual to engage in a given behavior. If other people are engaging in the behavior or if the behavior is determined to be socially valued, an individual is more likely to engage in the focal behavior. The third cognition is perceived behavioral control. If a person believes he or she can perform the behavior, then they are likely to engage in said behavior. This is similar to efficacy of positive self-presentation discussed previously (Pauls & Crost, 2005). Finally, intentions serve a mediating role between the above cognitions and behavior (Ajzen, 1991).

TPB highlights the complexity of a given behavior, suggesting a variety of situationally dependent factors that play a part in an individual's choice to engage in a behavior. Many of these factors are largely intra-personal and thus may be dependent on the situation or various cognitions. This dynamic understanding of behavior calls into question more static explanations such as that offered by SD above. The theory also applies well to faking because it more accurately defines the construct as a behavior rather than a stable trait.

Two recent models of faking included TPB as a central component (i.e., McFarland & Ryan, 2006; Mueller-Hanson et al., 2006). Attitudes, social norms,

and perceived behavioral control all emerged as significant determinants of faking behavior in empirical tests of these models. Attitudes toward faking tended to be stronger predictors of faking behavior than dispositional factors such as Machiavellianism (Mueller-Hanson et al., 2006). This suggests faking is a behavior and not a trait. Social norms and perceived behavioral control also emerged as major factors in building an intention to fake. Multiple studies observed that expecting others to engage in faking, perceiving the situation as important for achieving goals, and perceiving control over the situation increase intention to fake (e.g., Mueller-Hanson et al., 2006; Pauls & Crost, 2005).

Additional research suggests that a model of faking that incorporates TPB offers a good fit to the data (McFarland & Ryan, 2006; see Figure 1). The authors tested several iterations of their model (McFarland & Ryan, 2000), and ultimately found that faking intentions were significantly related to faking behavior. Attitudes and perceived behavioral control both were significant predictors of faking intentions. In sum, TPB is a useful framework for discussing the mechanisms by which individuals choose to engage in behaviors in general, and faking behaviors in particular.

Motivation and faking. Factors outside of those offered in TPB can influence an individual's decision to fake. For instance, the motivation to achieve a desired outcome such as a job plays a meaningful role in guiding behavior. While the idea of motivation is partially addressed by "intentions" within the TPB

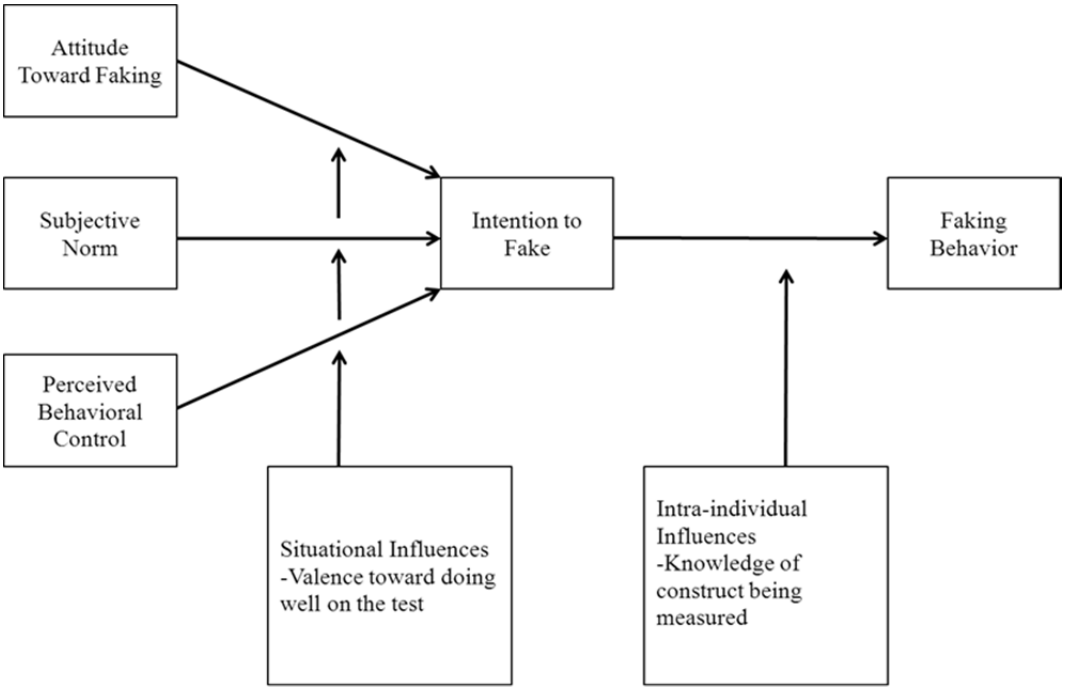

Figure 1. Model of Faking based on the Theory of Planned Behavior (based on McFarland & Ryan, 2006).

framework, we can gain a more complete understanding of faking through a more nuanced focus on motivational antecedents.

Expectancy theories of motivation are often used to explain and predict volitional behaviors (*cf.* Ellingson, 2011). In particular, Vroom's (1964) Valence-Instrumentality-Expectancy (VIE) theory is well positioned to contribute to the understanding of faking. VIE theory is used in many contexts to describe how an individual chooses from multiple courses of action in the pursuit of valued, extrinsic outcomes (Ellingson & McFarland, 2011). Although meta-analytic support is tenuous for the theory as a whole, the individual factors are useful for explaining motivation and behavior (van Eerde & Thierry, 1996). Valence is the affective component of the theory, referring to the preference for a particular outcome based on the anticipated satisfaction or desirability associated with it. Instrumentality refers to the belief that engaging in a particular behavior will actually result in achieving the outcome in question. This construct can be understood as having a clear "line of sight" between a behavior and an outcome. If instrumentality is high, individuals clearly see the connection between a behavior and an outcome. Finally, expectancy refers to the extent to which an individual believes he or she can accomplish the behavior. Test-takers will not be motivated to engage in faking behavior without the belief that they have the capability to fake and increase their scores on the assessment. Refinements to VIE theory also include an explicit ability component (e.g., Lawler & Suttle, 1973), meaning that people must truly possess the ability to achieve an outcome and not simply believe that they have the ability to do so.

Within the selection literature, VIE theory is used to understand motivation during applicant test-taking (Sanchez, Truxillo, & Bauer, 2000). Research on VIE and faking suggests that a multiplicative approach among factors is most appropriate (Ellingson, 2011). This means that positive, non-zero levels of each of the factors (i.e., valence, instrumentality, and expectancy) are needed in order to fake. If an individual already possesses the trait in question, then he or she would have no instrumentality because there is no reason to believe that faking will influence their chances of achieving a desired outcome (Ellingson & McFarland, 2011). Snell, Sydell, and Lueke (1999) went on to propose a model in which test-takers' increased desire for the job (i.e., valence) and confidence that they can successfully raise their test score (i.e., expectancy) lead to increased faking motivation. However, the authors and others note that certain assessment characteristics such as item type or item format can inhibit test-takers' ability to fake, regardless of their VIE levels. In other words, there are test-specific factors that can moderate faking behavior by influencing faking ability (Ellingson & McFarland, 2011; McFarland & Ryan, 2006; Tett, Freund, Chistiansen, Fox, & Coaster, 2012).

More systematic variance is often observed in personality scores from motivated contexts than from unmotivated contexts (e.g., Heggestead, 2011). In conjunction with the review of VIE above, these findings add to the understanding of faking in two ways. First, it complements the aforementioned discussion of intention within the TPB framework by furthering our understanding of when and why individuals *intend* to engage in a behavior. A

comparison of the model proposed by VIE theory in Figure 2 with that of TPB in Figure 3 shows how well the two theories complement each other. Second, it directly addresses one of the main conceptualizations of faking discussed in the literature. Many studies use laboratory groups that differ solely in their motivation to achieve a valued outcome, and used this as an experimental comparison of faking and honest responding. Integrating TPB and VIE theories provides a more comprehensive theoretical foundation that can be used to understand faking.

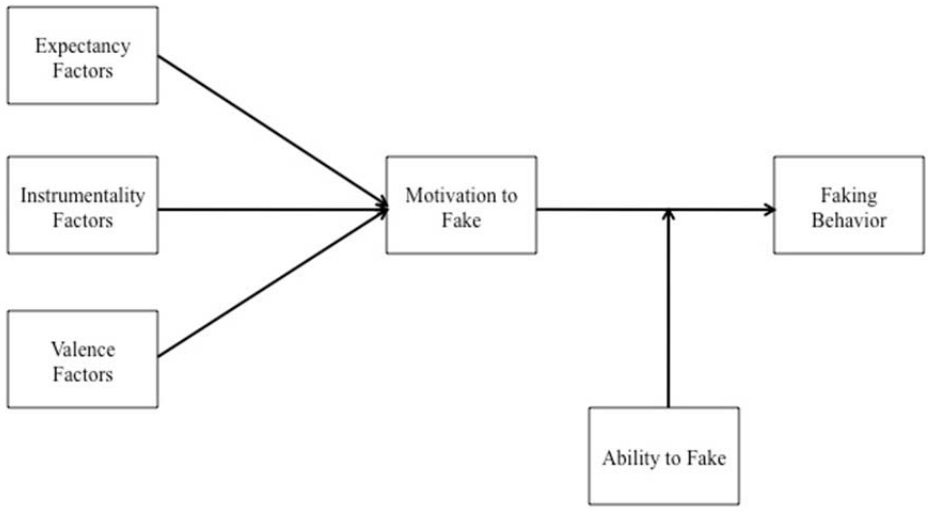

Figure 2. Model of Faking based on the Valence-Instrumentality-Expectancy-Theory

Faking within psychometric theory. A thorough understanding of faking must have some foundation within psychometric theory because of its core existence as a measurement issue (Heggestead, 2011). This additional lens will

apply parts of the theories discussed above in order to identify potential sources of faking behavior. The following discussion will round out our discussion of faking from a theoretical standpoint, and assist in defining the construct. The discussion uses classical test theory (CTT) as a framework. Although generalizability theory is more widely encouraged in the field, CTT offers a particularly accessible and clear mechanism for demonstrating the impact of faking and will assist in creating a general framework for the faking interventions discussed later.

According to classical test theory, an observed score is the function of an individual's true score and error, commonly represented by:

$$X = T + E. \quad (1.1)$$

The true score (*T*) represents the expected score for an individual if he or she were to complete a particular personality assessment across a large number of identical testing situations. Random error (*E*) can be understood as noise in the equation; inconsistent "variations in attention, mental efficiency, distractions, and so forth" (Schmidt, Le, & Ilies, 2003, p. 208). Because all testing operates within the social environment, there will always be a certain degree of error associated with the observed score (*X*).

As a measurement issue, faking must be incorporated into this equation. Because many studies find differences in mean trait scores between motivated and unmotivated contexts (e.g., Griffith, Chmielowski, & Yoshita, 2007), faking cannot be considered simple random error. One potential solution is to simply include it as one other variable in the equation. For example, if we are interested in faking on an extraversion scale (Heggestead, 2011), an equation may look like:

$$X = T_{extra} + T_{fake} + E. \quad (1.2)$$

Here, a unitary concept of faking is incorporated into the equation. This is problematic because faking has already been established as a multidimensional construct. Faking behavior has several antecedents and is conditional on a variety of factors, as discussed above. The complexity of behavior renders a single faking factor incomplete. Thus, several components of faking behavior should be incorporated into the equation. For example:

$$X = T_{extra} + (T_{F\text{-}instrumentality} + T_{F\text{-}ability} + T_{F\text{-}intent} \ldots + T_{Fn}) + E \quad (1.3)$$

where $T_{F\text{-}valence}$ .. T_{Fn} represent various sources of faking behavior. Instrumentality, ability, and intent are included as examples in this equation to address TPB and VIE discussed previously. For instance, observed scores can be influenced by the instrumentality of faking, the ability to fake on the assessment, and test-taker intent to fake on the assessment. There are a number of other sources of faking (e.g., valence, perceived behavioral control) excluded in the equation above for the sake of demonstration.

Many interventions designed to mitigate the effect of faking target one of these sources, most often faking intent or ability. For instance, warning statements target faking *intent* in an attempt to produce a more accurate portrayal of a true score, while using subtle items attempt to reduce faking *ability*. Other interventions disregard these sources and attempt to address faking after-the-fact. Indeed, many assessment publishers suggest correcting for social desirability when interpreting personality assessments. Understanding faking within the

context of CTT helps us to identify how such interventions differ in their approach to diminishing faking behavior.

This equation highlights the core message from the previous theoretical review: faking is a complicated behavior with several factors to consider. Framing faking within CTT more clearly identifies some of the "sources" of faking that can influence an observed score. The application of this final lens to faking does well to assist in understanding the role various faking interventions take in influencing faking behavior.

Summary and conceptual definition of faking. Based on the theoretical review above, faking is conceptually defined as a multidimensional behavior that is both intentional and deceptive, and seeks to benefit one's own interests. There are several important aspects of this definition. First, faking is multidimensional. Classical test theory and TPB suggest that faking, like other behaviors, is the result of a dynamic interaction of factors such as intentions and attitudes. It is important to understand faking as multidimensional from both a conceptual and practical perspective. For the latter, interventions looking to combat faking should be concerned with *how* the intervention is influencing behavior. As mentioned earlier, different interventions may be targeted at different sources of faking behavior such as intent and ability. It is additionally crucial to point out that faking is a behavior and not a trait. Differences in faking occur across people, but also across situations within individuals.

The conceptual definition also describes faking behavior as intentional. VIE theory suggests faking is a volitional behavior done to enhance the chances

of acquiring a desired outcome. Further, faking includes deception, which may be internally or externally focused. Regardless of the focus, the core intent of this deception is not malicious. It is primarily meant to benefit one's own interests or improve the possible outcomes (e.g., obtaining a job). While faking may have adverse outcomes for other parties such as the organization, these consequences are a by-product of the faking behavior and not the primary intent (Ellingson & McFarland, 2011).

Operationally Defining Faking

With a clear conceptual understanding of faking, the discussion can now move to how the construct is operationally defined in empirical research. The extant literature has found difficulty researching the topic because it requires deducing a complex array of human behavior into a single test or observation. However, a consistent understanding and measurement of faking is necessary. For this meta-analysis, faking will be operationalized by comparing mean scores across conditions that are theorized to only differ with respect to faking. Most of the literature regards "faked" scores as those that are outside of a 95% confidence interval away from the mean on that trait. There are some differences in techniques, as some researchers have used the standard error of measurement (SEM; e.g., Griffith et al., 2007), others have used the standard error of the difference (SED; et al., Peterson et al., 2009), and others have used the standard error of measurement for the difference score (e.g., Arthur, Glaze, Villado, & Taylor, 2010). There is evidence that these methods differ significantly in identifying fakers (Peterson, Griffith, Converse, & Gammon, 2011). However, the

basic premise in the methods is the same, as faked responses on a socially-desirable trait are expected to have a higher mean score than honest responses. Because of differences observed in classification methods, and to facilitate the aggregation of data, the current analysis focuses on the effect size (i.e., standardized mean difference) between conditions.

Comparison studies are typically either naturally occurring (e.g., comparing applicants to incumbents) or experimentally induced (e.g., comparing participants instructed to fake or not). Results from the two samples are compared, and the difference in personality mean scores is attributed to faking. This way of operationalizing faking rests largely on the assumption that people will present themselves in a positive light when motivated to do so, and respond more honestly (i.e., less socially desirable) when the motivation is absent (e.g., Ellingson & McFarland, 2011).

The rationale behind operationalizing faking as mean difference scores across motivated conditions is straightforward. Baseline or control data is collected in an unmotivated context (i.e., the incumbent or no faking instructions group). The scores are then either compared to another motivated group (between-subjects) or collected from the same test-takers again in a motivated context (within-subjects). For traits that should be perceived as desirable (e.g., conscientiousness), motivated contexts should display higher mean scores than in unmotivated contexts, while the inverse would be true for undesirable traits like neuroticism. Indeed, meta-analytic results suggest that participants are able to increase scores more than half a standard deviation under instructions to fake

good compared to no instructions (Viswesvaran & Ones, 1999). Table 1 demonstrates that test-takers are able to meaningfully change the mean dimension scores across each of the Five Factor traits. It should be noted that experimentally induced faking tends to produce a larger effect size than does naturally occurring faking (Holden & Book, 2011), although consideration of methodological moderators may limit these differences (Hooper, 2007).

The assumption of equality between the two samples is most frequently cited as a criticism of this understanding of faking (Mount & Barrick, 1995; Tristan, 2009). Incumbents may be substantially different from applicants, if for no other reason than because they already passed the selection assessment in question. Researchers and practitioners who believe faking is inconsequential or its research is unnecessary argue that incumbents will score high on the selection measure because they were already selected, and that the applicant sample will have more variability. While this likely true, it does not explain why applicants would then score *higher* on the personality assessment than incumbents. (e.g., Ellingson, Sackett, & Connelly, 2007; Griffith et al., 2007). In fact, applicants tend to score roughly one-third of a standard deviation higher (d = .35) across all of the FFM traits compared to incumbents (Tett et al., 2006).

At its core, the issue of nonequivalence represents a concern over between- versus within-group study designs. There are indeed numerous differences between samples, especially applicants versus incumbents, which are not accurately caught by a between-subjects design (Guion & Cranny, 1982;

Table 1.

Viswesvaran and Ones (1999) Mean Effect Sizes (d) for Five Factor Model Dimensions Under Fake Good Instructions.

	Fake Good		
Personality Trait	*Within*	*Between*	*Weighted Average***
Emotional Stability	0.93	0.64	0.76
	(921)*	(1357)	
Extraversion	0.54	0.63	0.61
	(391)	(1122)	
Openness	0.76	0.65	0.68
	(259)	(614)	
Agreeableness	0.47	0.48	0.48
	(408)	(1009)	
Conscientiousness	0.89	0.60	0.66
	(723)	(2650)	
Weighted Average	0.78	0.60	0.65

Note: Effect sizes represent the difference in mean personality trait scores between fake good and honest instructions.
* Numbers in parentheses represent the total *N* per mean effect size.
** Weighted averages are average mean effect sizes weighted by total *N* in each condition or trait.

Tristan, 2009). Although lab studies that use "instructionally-induced" faking can use random assignment to address non-equivalence, many faking researchers argue that between-subject designs are less methodologically sound than within-subject designs (*cf*. Griffith & Converse, 2011). Collecting personality data from the same individual in a motivated context and in an unmotivated context provides stronger evidence that differences are due to faking and not due to differences on extraneous variables. Between-subjects designs are frequently used in large part due to the practical ease of their implementation. The type of design is an important consideration when designing a faking study. Effect size calculations, as well as prevalence estimates, vary as a function of design type (e.g., Peterson, Griffith, O'Connell, & Isaacson, 2008).

Prevalence of Faking

Prevalence estimates of faking tend to vary depending on how faking is operationalized. When studies define faking as mean difference scores, the prevalence tends to be around 30 percent of test-takers. For instance, Griffith et al. (2007) classified applicants as fakers by examining those who elevated their score in a motivated context to a score that fell outside of a confidence interval around their honest score (i.e., the score obtained in an unmotivated condition). With a 95% confidence interval based on the standard error of measurement (SEM), roughly 31% were classified as fakers. However, the same study used the standard error of the difference SED to identify only 22% of the sample as fakers. Other studies on the prevalence of faking find similar estimates, ranging from 21%

(Donovan, Dwight, & Schneider, 2008) to 33% (Arthur et al., 2011) of samples identified as fakers (the latter using the SEM approach).

It is clear that prevalence estimates vary across different operationalizations of faking. Little research has directly compared prevalence estimates using different methods. Peterson et al. (2011) directly compared three methods on the same sample of simulated applicants (i.e., students instructed to fake as if they were an applicant for a job). The authors found that the choice of faking detection method significantly altered the percent of responses flagged as faked. Using the SEM consistently identified a larger proportion of fakers, while the SED approach identified the lowest proportion of "fakers" across personality traits.

The current study is not focused on identifying the prevalence of faking, but it is important to understand the extent of faking in order to understand its implications. Because the prevalence rates tend to vary, it is likely that roughly 30% of test takers can be classified as "fakers."

Implications of Faking

The nontrivial prevalence of faking may represent a threat to an organization's selection system. As discussed earlier, faking may produce a meaningful cost to organizations. It is thus quite important for organizations to be aware of the potential areas in which faking may have an effect. Most research on the implications of faking focuses on changes to the rank-order of the applicants (i.e., shifts in mean scores) and the validity of inferences from the assessment (i.e., criterion-related validity).

Faking and rank-order of applicants. Given the literature review to this point on the effect faking has on mean personality scores, it is unsurprising that faking can influence the rank-order of applicants in such a way that benefits fakers at the detriment of honest test-takers. This undermines one of the key assumptions of using personality tests discussed at the beginning of this chapter, as organizations assume that the test yields an accurate reading of a test-takers relative standing on a given trait.

Multiple simulations show that faking can change the rank-order of applicants, particularly at the upper-end of the distribution (Douglas et al., 1996; Zickar et al., 1996). This suggests that in a top-down selection system, hiring decisions may be altered by the prevalence of faking. Empirical studies find similar results, with substantially different rank-orders of applicants between adjusted and nonadjusted personality score conditions (Rosse, Stecher, Miler, & Levin, 1998).

Among other factors such as test reliability (Mueller-Hanson et al., 2003) and individual differences in willingness to fake (Griffith et al., 2006), researchers highlight the importance of selection-ratio in considering how faking influences rank-order or selection decisions (Christiansen, Goffin, Johnston, & Rothstein, 1994; Griffith et al., 2007; Rosse et al., 1998). As the selection ratio decreases, the number of fakers selected disproportionately increases (Mueller-Hanson et al., 2003). With a very high selection ratio (e.g., 60% and above), the percent of honest respondents selected is sufficiently similar to the percent of honest responders in the entire sample. However, when the selection ratios were

smaller (e.g., below 50%), the number of honest responders selected was significantly lower than the number in the entire sample. Meanwhile, Rosse and colleagues (1998) found that seven of the eight people hired were fakers when only the top 5% of applicants are hired. These results demonstrate that faking has a particular influence at the upper-end of the distribution of personality scores, which is exactly where it will have the most influence (Jenson & Sackett, 2012; Tett et al., 2006). This creates an increased opportunity for those applicants who are not naturally high on the job-related trait in question.

The results do not guarantee that fakers will always be hired above honest responders. For one thing, most organizations use a variety of selection tools when making hiring decisions. However, if the rank-order is influenced even slightly, selection decisions may be altered in turn.

Faking and criterion-related validity. A limited body of research within the faking literature focuses on criterion-related validity. It is also inexorably connected to one of the assumptions of the use of personality assessments in selection. Recall that the practice of administering a personality assessment to employees assumes that it is a useful predictor of future behavior or performance. Researchers arguing for the significance of faking suggest that faking on personality measures attenuates the criterion-related validity. This finding is widely debated within the field, with a fair amount of support for both sides of the argument.

Past research observes greater prediction error for incentivized participants compared to a control condition, particularly at the upper-end of the distribution

of scores (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Mueller-Hanson, Heggestad, & Thornton, 2003). In the incentive group, the relationship between personality and the criterion was significantly lower for test-takers that scored in the upper third of personality scores (r = .07) compared to test-takers at the bottom third of the distribution (r = .45). On the contrary, no differences were observed for the control group. Although this study did not demonstrate a statistically significant difference between validity coefficients for the two conditions (i.e., incentive and control), the difference was practically significant. Other research demonstrates significant differences in validity coefficients between conditions, such as fake good and honest responding conditions (e.g., Holden & Jackson, 1981). In a Monte Carlo simulation, Douglas, McDaniel, and Snell (1996) demonstrated how the inclusion of fakers into a sample lowers the criterion-related validity when predicting job performance. Specifically, the validity coefficients for conscientiousness and agreeableness dropped significantly as fakers were added to the simulated sample. Similarly, Komar et al. (2008) used a simulated dataset to demonstrate that faking can significantly decrease validity coefficients for predicting supervisory ratings of job performance under certain situations, such as the variability of faking and selection ratio.

Still, others maintain that faking has no noticeable effect on criterion-related validity across applicant and incumbent samples. Some note estimates are higher for applicants (r = .40) than for incumbents (r = .29) when examining the relation between integrity tests and overall job performance (Ones, Viswesvaran,

& Schmidt, 1993). In a similar vein, research suggests that the small differences in validity coefficients between applicant and incumbent samples (e.g., .07 difference; Hough, 1998) are not practically significant. Even in comparing "fake good" and "fake bad" conditions, validity coefficients remained stable. Finally, some researchers feel that faking is not a problem because of the already modest criterion-related validity coefficients for personality in most contexts (Morgeson et al., 2007a). Under ideal scenarios, personality tests can only account for about 15% of the variance in job performance. This does not leave much room for faking to meaningfully impact one's ability to make valid predictions based on the scores on a personality assessment.

The above review shows that research on faking and criterion-related validity is by no means consistent. One reason for discrepant findings regarding criterion-related validity may be in how the authors define a "significant change." That is, different estimates will result if operationalizing the change as significantly different from zero (i.e., single-group validity) versus significantly different from each other (i.e., the two groups). In any case, this disagreement simply highlights the continued need for research to refine our understanding for how response distortion affects criterion-related validity.

Interventions for Faking

Faking has a measureable impact on the efficacy of personality assessment for employee selection. As such, researchers and practitioners have developed various interventions that seek to address the problem. Conceptually, these interventions seek to eliminate the different "types" of error within the classical

test theory lens discussed above. A visual representation of how each intervention theoretically influences faking behavior is offered in Figure 3. The proposed model of interventions represents an application of the theoretical bases of faking behavior. Factors such as perceived control (Theory of Planned Behavior; TPB) and valence (Valence-Instrumentality-Expectancy; VIE) serve as antecedents. This list is far from exhaustive, and partially represents individual differences with regards to faking. As seen in the model, faking interventions primarily focus on situational factors rather than intraindividual factors. More specifically, faking interventions can be classified as either preventive or remedial. Preventive interventions seek to influence behavior before it occurs, while remedial approaches allow the assessment to be administered and seek to tease out the potential effect of faking after the fact.

A more nuanced look into these interventions demonstrates that they target specific factors within some of the previous theoretical models of faking behavior. Preventive interventions can either focus on modifying *intent* or *ability* to fake. Warning statements are the most common approach for preventively influencing faking intent, while item transparency, forced choice formats and time limits seek to influence faking ability. The most common remedial strategies are score corrections or removing individuals due to high SDR scores. These are done after assessment data are collected, and thus do not influence actual faking behavior. The following sections will review the most common interventions, discussing both the theory behind them and the empirical support in the extant literature. The

discussion will follow the progression put forth in **Error! Reference source not found.**

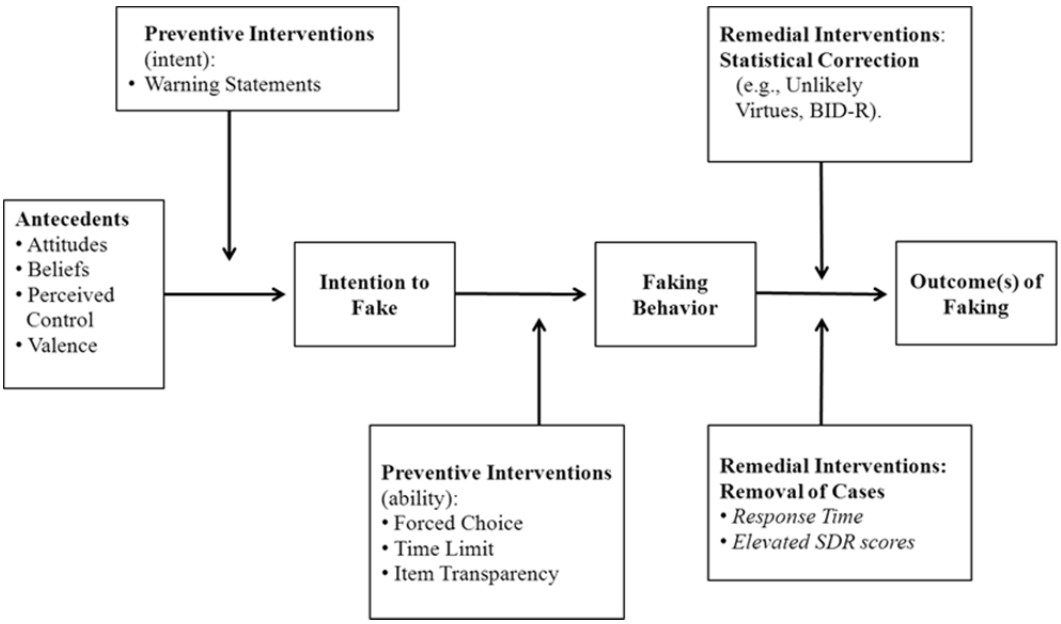

Figure 3. Theoretical Model of Faking Integrated with Proposed Interventions based on McFarland and Ryan (2000).

Preventive interventions – intent. Preventive interventions focused on limiting intention to fake can do so by modifying test-taker cognitions such as instrumentality or valence. Test-takers can be expected to engage in less faking behavior to the extent that they believe that others are no longer engaging in the behavior or that the behavior is not socially valued. The most common approach focused on intent is the use of a warning statement in the assessment instructions.

Warnings. Using this approach, test-takers are cautioned prior to taking the assessment that there are measures that can detect distorted or faked responses. This technique is effective in that it directly addresses one of the limitations of self-report personality data; namely, a lack of accountability to the answers provided (Dunning et al., 2004). Warning statements seek to provide this accountability by making the items appear verifiable. In practice, the content warning statements can take a variety of forms that typically conform to one of five categories (*cf.* Pace & Borman, 2006). In its most simple form, *detection* warnings suggest that faked responses can be detected, while *consequential* warnings suggest a negative outcome (e.g., removal from applicant pool) as a result of faking. The other three categories of warning statements (i.e., *appeal to reason*, *educational*, and *appeal to moral principles*) receive much less empirical testing than the first two warning types. Warnings looking to *appeal to reason* state that test-takers who respond more honestly will more accurately represent their personality, while warnings *appealing to moral principles* state that, as an honest and moral individual, the test-taker should not lie. Finally, *educational*

warnings state that faking on the assessment means that the researchers or test administrators cannot accurately evaluate the responses.

Warning statements are effective insofar as they influence the intention to fake (see Figure 3). This sentiment is reflected in McFarland and Ryan's (2000) model of faking behavior, where warnings are proposed to moderate the relationship between beliefs toward faking and faking intentions. The authors later found support for this proposition, finding that warnings had a direct effect on intention to fake for most of the five factor personality traits (McFarland & Ryan, 2006). The authors also unexpectedly found that warnings have a direct effect on faking behavior, even when controlling for intention to fake. This unexpected finding suggests that warnings may be a particularly robust intervention for addressing faking.

Compared to an assessment given across motivated and unmotivated conditions without a warning, the presence of a warning should result in a smaller difference in mean personality scores. Empirical studies suggest warnings are effective in lowering scores on both personality trait and SD scores. Dwight and Donovan (2003) meta-analytically demonstrated that test-taker personality scores were substantially lower (d = .23) for positive traits like conscientiousness when they were warned than when they were not. More recent primary studies support the findings of this meta-analysis and find that warnings are effective at reducing faking behavior across a variety of settings (e.g., Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Dullaghan, 2010; Griffith et al., 2006; Landers,

Sackett, & Tuzinski, 2011). It thus appears that warnings are effective at deterring faking behavior.

Mean differences in predictor scores and social desirability are useful in showing that warning statements change behaviors. However, the utility of warning statements for making accurate decisions is also an important issue. The criterion-related validity of the assessment plays a big role in these decisions. Interestingly, some studies find lower criterion-related validity coefficients in warned than in unwarned conditions (Harold et al., 2004; McFarland, 2003). Although the warning lowered self-reported faking, McFarland (2003) found lower correlations between GPA and all five factors of personality in the warned condition than in the unwarned condition. Other research suggests criterion-related validity coefficients are unchanged across warned and unwarned conditions when prediction criteria such as leadership and absenteeism (Converse et al., 2008; Fox & Dinur, 1998).

Some research uses an external rater's judgment of the test-taker's personality as a criterion for personality score accuracy. Others' ratings of a test-taker's personality are frequently used for assessing personality (e.g., Connelly & Ones, 2010) and often come in the form of peer's or acquaintances that know the test-taker well. One study found no difference between the self-observer ratings across warned and unwarned conditions (Robson, Jones, & Abraham, 2008). However, another study using self-observer ratings contradicted these findings and found that warnings moderated the relationship between self and other ratings of personality (Robie, Taggar, & Brown, 2009). This is not the only study to find

that warnings enhance criterion-related validity and selection decisions. Higher criterion-related validity estimates were observed in the warned condition than the unwarned condition when predicting behavioral procrastination (Illingsworth, 2004). Warnings can also enhance the criterion-related validity of other non-cognitive measures such as biodata (Mock, 1947).

Looking across both lab and applied settings, the presence of a warning has a meaningful influence on selection decisions (Fan et al., 2012; Lopez, 2009). The amount of fakers selected dropped 20% after a warning was given, while no drop was observed for the unwarned test-takers. Despite the mixed findings, warning statements are expected to enhance criterion-related validity.

Few studies compare the effectiveness of the different categories of warnings. In a follow-up primary study to their meta-analysis, Dwight and Donovan (2003) found that *consequential* warnings were more effective at lowering mean personality scores than *detection* alone in a motivated context. Several studies have followed suit, and tend to combine these two categories of warning into a single warning statement. Dullaghan (2010) hypothesized that the more "applicant friendly" *appeal to reason* warning would be more effective than a potentially disingenuous threat. However, results indicated that *detection/consequential* warnings were more effective than *reasoning* warnings in an applied context. The latter warning showed no significant change in mean scores compared to that of the faking condition. Additional research focused on the content of the warning statement is needed (Dilchert & Ones, 2011; Rothstein & Goffin, 2006).

The main limitation of warning statements is that the effectiveness of this intervention lies in the test-taker's trust in the warning's accuracy. It seems plausible if not likely that individuals' trust will vary significantly, making it difficult to generalize the effectiveness of one type of warning across test-takers. Some researchers also question the ethicality of providing false warning statements (Dilchert & Ones, 2011), especially considering the unconvincing literature discussed previously regarding direct faking measures. Because the detection measures are far from accurate, the warning statements about detection may be interpreted as disingenuous (Fort, 2010). Some researchers also suggest that warnings increase the cognitive difficulty of the measure (Vasilopoulos et al., 2005). If warnings make the response more complex, assessments with a warning may only be faked by test-takers with high general mental ability (Rothstein & Goffin, 2006).

In sum, the research tends to support the use of warning statements as a faking deterrent despite a few potential limitations. The practical simplicity of implementation makes it a popular intervention. A consolidation of the literature will provide insight into the effectiveness of this intervention and reveal potential avenues for future research.

Preventive interventions – ability. As opposed to warnings that focus on faking intent, other preventive interventions focus on faking ability. This category of intervention still seeks to modify faking behavior, but does so through a very different mechanism. Theories of faking suggest that even if a test-taker intends to fake, the assessment itself must allow for faking to occur. Ability interventions

thus overlook faking intentions, instead focusing efforts on limiting the "fakability" of the assessment. Research tends to focus on three specific interventions: using forced choice formats, and limiting item transparency, and imposing time limits.

Forced-choice. The ability to fake an assessment can be strongly influenced by its design. Most personality assessments present items independently from one another, and have test-takers respond to items on a Likert-style scale. In this format, test-takers can easily endorse positive items (i.e., positive traits) while ignoring or downplaying negative ones. A Forced Choice (FC) measure seeks to take this choice away from the test-taker by making it impossible to avoid endorsing a more "negative" item. Research on FC measures suggests that they correlate strongly with normative measures and offer useful information about absolute trait levels (e.g., Bowen, Martin, & Hunt, 2002; Goffin, Jang, & Skinner, 2011; Heggestad, Morrison, Reeve, & McCloy, 2006).

Most FC measures ask test-takers to respond to an item with multiple options that are balanced in terms of social desirability. Some FC items are structured in a way that asks the test-taker to rank the available options in terms of how much the items reflect their own personality. Another item may be formatted such that the test-taker chooses the option that is "most like him/her" or "least like him/her," with as many desirable options as undesirable ones. For instance, two FC items may ask the following (Christiansen, Burns, & Montgomery, 2005): "Which of the following adjectives is most true or most descriptive of you?":

(1) Practical or Imaginative

(2) Unkind or Careless

For item (2) above, it is unlikely that either adjective is socially desirable for a job applicant. On the other hand, both adjectives in item (1) could be socially desirable. Thus, test-takers are unable to strictly adhere to socially desirable responding patterns.

Regardless of format, the psychometric limitations of FC measures must be addressed. For instance, FC measures by nature reflect intraindividual differences as opposed to interindividual differences (Converse et al., 2010; Meade, 2004), and personality dimensions are not independent, such that a test-taker cannot receive a high score on all dimensions (Dilchert & Ones, 2011). However, partially ipsative FC measures, as opposed to purely ipsative measures, address these limitations. Purely ipsative measures exist when test-takers have to rank all options per item or all paired comparisons are taken into account in scoring (Converse et al., 2010). This is problematic because the total score for all participants will be the same, which highly reduces the variance and often results in negative correlations in the data (Converse et al., 2010; Meade, 2004). Partially ipsative measures, on the other hand, increase test variance by not having the total test score constant across all test-takers. With the example above, a measure could be partially ipsative if some of the response options are not scored or differential weights are applied in scoring response options (Hicks, 1970). Other ranking formats can be partially ipsative if test-takers are not required to respond to all options within an item or if scales have a varying number of items (Dilchert & Ones, 2011).

As depicted in Figure 3, FC measures seek to reduce faking by limiting test-takers' ability to fake. Test-takers cannot choose every desirable response because each item contains potentially multiple socially desirable options. Thus, the applicant's response to the item theoretically represents the option that is most like the test-taker because they have to choose between multiple desirable options (Gordon, 1951).

Within the faking literature, mean difference scores are typically compared between FC measures and non-FC measures (e.g., Likert-style). Jackson et al. (2000) found that participants completing a Likert-style personality inventory were able to increase their scores on a dependability scale nearly a full standard deviation (.95 *SD*) when instructed to do so, compared to a group that was instructed to respond honestly. This difference was over a half of a standard deviation higher than the increase for an FC measure. Martin, Bowen, and Hunt (2002) employed a slightly different method, comparing test-takers actual ratings to that of a self-rated "ideal" applicant. No difference in ratings was observed between honest and faking conditions for FC measures, while the normative (i.e., Likert-style) measures showed large differences between conditions. Consistent results were found in Bowen et al. (2002) using a similar approach and methodology. Within the five-factor framework, mean difference scores for conscientiousness and extraversion were higher across applicant and fake-good instructions for Likert-style measures than FC measures (Christiansen et al., 2005; Converse et al., 2010; Heggestad et al., 2006).

Although empirical research on mean difference scores is generally in the proposed direction, the strength of the effect varies significantly. For instance, Christiansen et al. (2005) found mean differences for conscientiousness to be much lower for FC (d = .40) than Likert (d = .68) measures when comparing honest and applicant instructions Heggestad et al. (2006), on the other hand, found much stronger mean differences for FC and Likert measures but the differences between them were minimal (d = 1.20 and 1.23, respectively). These varying effect sizes suggest that a meta-analysis may be useful for providing an estimate across studies of FC as a faking intervention.

There are two important caveats in the discussion of FC measures. First, while FC measures tend to show less response inflation than Likert-style measures, both measures show inflated scores compared to an "honest" sample. Indeed, Jackson et al. (2000) reported that the mean shift between the honest and applicant conditions for the FC measure was meaningful (d = .32). Although this is less than the Likert-style measure (d = .95), a degree of inflation is still observed. Second, FC measures are problematic in selection contexts because normative comparisons across applicants are difficult. However, recent research suggests that some FC measures (i.e., partially ipsative measures) can be used to attain normative trait standing (Heggestad et al., 2006).

Item transparency. The above section on FC measures indicated they limit faking ability in part by making the underlying trait less transparent. In typical test formats (e.g., Likert-scales), items measuring a similar construct are grouped together. This format typically demonstrates desirable psychometric

properties such as strong internal consistency and clean factor loadings (Schriesheim, Solomon, & Kopelman, 1989). However, doing so may increase faking ability. Additionally, the content of the items can influence the extent to which the test-taker is able to identify the measured construct. Measures that include the name of the construct in the item are likely especially transparent (e.g., Ten-Item Personality Inventory; Gosling, Rentfrow, & Swann, 2003). Thus, faking ability should be limited to the extent that the measured construct is less transparent. This is achievable through a variety of methods.

A scrambled or randomized format is where "proximal items on a scale load onto different constructs" (Dilchert & Ones, 2011, p. 187). This approach asserts that test-takers look for connections between items, and that the underlying constructs are less clear by mixing the items together. Despite assertions that grouping items together improves psychometric properties (e.g., Schriesheim & DeNisi, 1980), recent research suggests that internal consistency is not significantly different across scale format (Schnell et al., 2011). As opposed to the order of all the items on a scale, the content of the items may also influence faking behavior.

Integrity testing and biodata research has long focused on the use of subtle versus obvious items (e.g., Alliger, Lilienfeld, & Mitchell, 1995; Mael, 1991). The item content in this sense refers to the extent to which a test-taker can be reasonably expected to identify the trait measured by a given item. Obvious items reflect those of more common personality scales (e.g., NEO-FFI) where items are created with the intent of tapping a single construct. Subtle items, on the other

hand, are typically created through empirical keying. This means that they are not created for a construct *a priori* as obvious items are, but rather are chosen based on how well they empirically differentiate groups. Implicit measures of personality can be argued to be more subtle, but will not be included in this review because they reflect performance on tasks such as target discrimination and speed (e.g., Flipikowski, 2007).

Using contextualized items can also limit faking behavior. Due to the ambiguous nature of some general personality measures, test takers may be more prone to respond positively across all items. By providing a context, such as a specific frame-of-reference (FOR), personality assessments may provide a more accurate level of a test-taker's trait level. Previous researchers have theorized that an "at work" FOR in particular may limit the test-taker's *ability* to fake by constraining the situational context (English, 2004). Empirical tests bear this out, as a version of the NEO-FFI that used an "at work" FOR showed less response distortion across the five traits than a general version of the NEO-FFI (English, 2004; Griffith et al., 2006).

Through the above discussion and review of the literature, it is clear that these three methods seek to limit faking by reducing "item transparency." The rationale behind this decision has several key components. First, all three approaches are focused on the item level. This stands in contrast to other interventions such as warnings, which are focused on the person level. By sharing the same level of focus, these approaches are likely to influence faking behavior in similar ways. Indeed, these approaches target faking ability within the model of

faking behavior (see Figure 3). Similar to FC scales, the purpose is to modify test-taker's ability to fake or distort on the assessment. These approaches also use a similar rationale to limit transparency, and share the same mechanism for influencing behavior (i.e., limiting faking ability).

The extant literature on item transparency as an intervention for faking almost exclusively focuses on its impact on mean personality trait scores. In general, both item order and subtlety have received partial support as viable interventions. Indirect support can be gleaned from a study that included items measuring a "decoy construct" around focal items (Bernal, 1999). Mean scores across honest and faked conditions on the focal personality traits decreased with the inclusion of a decoy construct the prevalence of faking, but did not influence the rank order of participants. Although this aligns with the theoretical assertion of randomizing item placement, the inclusion of a decoy construct may contribute additional problems such as increased cognitive or reading load.

A more direct test of this intervention found that mean conscientiousness scores were higher for grouped items than for randomized items across honest, fake good, and respond as applicant conditions (McFarland, Ryan, & Ellis, 2002). The findings regarding the effectiveness of grouped and randomized items is far from consistent, as more recent research finds that only agreeableness has lower mean scores in randomized formats (Wolford, 2009), and others have found no significant mean differences for any personality traits (Schnell et al., 2011). However, even when not statistically significant, mean scores across the FFM traits tend to be smaller in randomized formats than in grouped formats.

Subtle items may provide an advantage over obvious items. Literature outside of the selection context suggests that empirically keyed scales are more resistant to faking than traditional ones (e.g., Dannebaum & Lanyon, 1993; Hsu et al., 1989). Within a selection context, however, the literature is more mixed. Item subtlety was not significantly related to item-level faking when operationalized as mean difference scores across motivated conditions (Gibby, 2004). Utilizing item response theory, item transparency led to more faking (operationalized as responses to an unlikely virtues scale) for openness and extraversion items, but not for conscientiousness (Day, 2008). Using an unlikely virtues scale stands in contrast to most other research on item transparency that uses mean difference scores to operationalize faking. Indeed, meta-analytic evidence suggests that subtle items are less likely to be faked than obvious items in terms of mean difference scores (Viswesvaran & Ones, 1999). The above literature review suggests that although the literature is inconsistent, item transparency (i.e., item order or content) may be effective at lowering faking when faking is operationalized as mean difference scores.

Time limits. Many theorists hypothesize that the amount of time it takes an applicant to answer an item can be used as an indicator of faking. Both overall response time to a measure and response latencies (i.e., the amount of time between item presentation and item response; Hsu et al., 1989) are commonly investigated within the faking literature. Time limits are proposed to limit the test-taker ability to fake on the assumption that lying takes time. Empirical literature, however, is mixed on the relationship between time and faking.

There are three main models of response time applied to faking that can be used to support opposing hypotheses. First, the Self-Schema model (Markus, 1977) suggests that lying should take longer than responding honestly. Self-schemas refer to generalizations about the self to aid in processing and organizing information. Because a distorted response is discordant with this self-schema, it takes time to cognitively process the options before providing an alternative response. It is therefore evident that this model considers faked and honest response patterns to involve very different cognitive processes. The Semantic Exercise model (Hsu, Santelli, & Hsu, 1989) suggests that faked responses will be faster because they refer to a less complex schema, namely that of an ideal respondent. This schema is more semantic in nature. The cognitive process will be less complex when "making a semantic evaluation than when making a self-referenced evaluation" (Martin, 2011; p .16). Finally, the Adopted Schema Model (e.g., Holden et al., 1992) goes a step further to suggest that faked and honest responses involve similar cognitive processes. This model suggests that both a self-schema and an adopted (i.e., "faking") schema exist while completing the assessment, where the latter represents the characteristics the test-taker wishes to display. Because the adopted schema has qualities more similar to that of an ideal candidate, the response time will be shorter for fakers responding in ways consistent with their adopted schema. Most importantly, this theory further proposes that *honest* responders will still use their self-schema during the assessment process. Fakers therefore display faster response times because of

their use of the adopted schema, as opposed to honest responders who use a self-schema (Martin, 2011).

There are two ways that organizations can use time in their personality assessment process. Time limits can be implemented before taking an assessment, or response time can be used remedially as an indicator of faking. The latter will be discussed shortly. Using time limits is done preventively with the intent of reducing test-takers' ability to fake. It follows from the Self-Schema proposition that if lying takes time, limiting the time participants have to answer the items should limit or prevent faking. However, the previous paragraph also put forth theoretical support for faked responses taking *less* time than honest responses. If time has a null or negative relationship with faking, imposing a time limit will not avert faking and may even negatively affect honest responders. Some research suggests that including time limits has no impact on a participant's ability to increase desirable trait scores compared to participants without a time limit (Holden, Wood, & Tomashewski, 2001; Robie et al., 2009; Robie et al., 2010). Other research finds that a time limit reduces socially desirable responding (Komar et al., 2010), while still other research finds that the results are mixed (Khorramdel & Kubinger, 2006). It is important to note that Komar et al. (2010) found an interaction between time limit and cognitive ability, showing that the time limit influenced SDR only for low cognitive ability participants. This assertion is consistent with self-regulatory models (e.g., Kanfer & Ackerman, 1989) or resource allocation theory (Ackerman, 1986). It follows that faking takes

up cognitive resources, and a time limit thus reduces the cognitive resources available for the test-taker to use for faking.

In general, research on time limits as a preventive faking intervention suggests that it is less effective than other interventions. The other preventive interventions focused on limiting faking ability, namely forced choice and item transparency, are likely more useful at reducing faking than time limits.

Remedial interventions. The primary mechanisms for influencing faking behavior are by limiting either intention or ability to fake. However, other interventions take a more pragmatic approach by focusing primarily on the decisions made from faking. In other words, remedial interventions seek to modify the interpretation of personality assessments after allowing potential faking behavior to occur. Common remedial interventions include partialling out variance due to socially desirable responding (e.g., SD scales, bogus items), or by removing applicants based on either elevated social desirability scores or response time.

Direct measures of social desirable responding. Using direct measures of socially desirable responding (SDR) together with personality scales a common intervention used to address faking despite its many shortcomings (Bäckström, Björklund, & Larsson, 2011; Ellingson, Sackett, & Hough, 1999). Direct measures often come in the form of social desirability scales such as the BIDR (Paulhus, 1991), lie scales, or bogus item scales (Harvel, 2012). There are typically two ways of using SDR scores in this context, and they are driven by

distinct theories for how social desirability influences the relationship between personality and a criterion.

First, high SDR scores can be considered an indicator of faking. Assessment administrators can choose to flag, retest, or even remove such test-takers from a given pool (Burns & Christiansen, 2006). At the core of this approach is the assumption that SDR acts as a moderator. Specifically, the relationship between personality and a criterion is assumed to vary as a function of SDR. When SDR scores are quite high, the relationship between personality and the criterion (e.g., performance) is believed to no longer be sufficiently similar to the rest of the sample. In other words, the criterion-related validity is sufficiently different at high levels of SDR and cannot be meaningfully compared to the rest of the sample. This approach is often advocated by commercial personality assessments, as assessment publishers often state that individuals scoring high on a SDR scale will have invalid personality profiles (Rothstein & Goffin, 2003).

Many empirical tests of this approach examine how the removal of high SDR scorers influences observations within a sample (e.g., Hough, 1998). Schmitt and Oswald (2006) found that removing applicants who scored above a cut-off on a SDR measure had little impact on mean job performance. Additional empirical research suggests that there are minimal differences, if at all, in the criterion-related validity of personality measures when removing high scorers on a SD scale (e.g., Christiansen et al., 2005). Direct tests of SDR moderating the criterion-related validity of personality assessments are inconsistent. Hough et al.

(1990) concluded that criterion-related validity did not substantially change at high levels of SDR, but White, Young, and Rumsey (2001) used the same personality inventory to find support for moderation. White et al. found that criterion-related validity was consistently lower in the high SDR group as opposed to both the moderate and low SDR group. Other support for moderation is found in predicting self-other congruence as the criterion (Borkenau & Ostendorf, 1992; Holden, 2007). A recent simulation tested the moderation hypothesis and failed to find a significant effect (Paunonen & LeBel, 2012).

Overall, the effect of SDR on criterion-related validity varies from small to nonsignificant. This stands in contrasts to other interventions, such as warnings, that tend to demonstrate more positive findings with regards to criterion-related validity (e.g., Illingsworth, 2004; Robie et al., 2009). Perhaps the main reason removing high SDR test-takers fails to enhance criterion-related validity is because, while fakers tend to score at the extreme on scales of SDR, extreme scorers on SDR are not necessarily fakers (Holden, 2007). Warning statements avoid this problem by preemptively limiting test-taker intent to fake. Regardless of the theoretical processes, providing warning statements appear to be a more viable intervention for enhancing criterion-related validity than removing cases based on high SDR scores.

SDR scores can also be used to correct an observed trait score, such that SDR is partialled out or corrected. This approach assumes that SD acts as a suppressor variable. A variable acts as a suppressor to the extent that it is related to the predictor variable yet unrelated to the outcome or criterion (Paulhus,

Robins, Trzesniewski, & Tracy, 2004). Thus, the variance attributable to the suppressor (in this case, SDR) acts as a contaminant and removing it makes the personality measure more efficient. To the extent that SDR acts as a suppressor, partialling out or correcting for it should remove construct-irrelevant variance and enhance the criterion-related validity (Burns & Christiansen, 2006). SDR scores are partialled out using either part correlation, multiple regression, or by using the residualized score from the trait-SDR correlation as the predictor (Reeder & Ryan, 2011). Using corrections in this way is a popular approach for addressing faking; over half of surveyed researchers supported the use of score corrections based on SD or other lie scales (Goffin & Christiansen, 2003).

Both primary and meta-analytic research, however, suggest that correcting for SD scores does not meaningfully enhance criterion-related validity. Five Factor traits, particularly conscientiousness and emotional stability, can actually demonstrate higher criterion-related validity in unadjusted rather than adjusted conditions (Barrick & Mount, 1996; Christiansen et al., 1994). Multiple meta-analytic studies found that criterion-related validities were roughly the same or decrease when partialling out SDR-related variance (Li & Bagger, 2006; Ones et al., 1996). One potential reason why this approach is not empirically supported is because SDR does not measure response distortion, but rather a predisposition to fake. Thus, because SDR measures a trait that is not entirely orthogonal to the criterion, the variance removed is construct-relevant and consequently lowers criterion-related validity. Just like with removing cases based on extreme SDR

scores, correcting scores based on SDR is expected to be less effective at enhancing criterion-related validity than providing warning statements.

Summary and integration. The previous sections review the most common interventions for combating faking. Preventive interventions focus on limiting actual faking behavior, either through faking ability or intention. Remedial interventions allow faking behavior to occur and instead seek to alter the interpretation of the personality assessment data.

It is evident through the review that preventive interventions are often more effective than remedial ones. Because remedial strategies fail to modify faking behavior, they are likely ineffective at addressing the faking problem. Remedial interventions are also limited because they operate on the assumption that faking behavior can be directly measured. Although faking is a measurement issue, it is unlikely that a self-report scale will sufficiently help with scale criterion-related validity. Preventive interventions are instead more focused on modifying a specific antecedent of behavior that will in turn limit the prevalence or effect of faking. The effectiveness of preventive measures over remedial ones is supported by the extant literature reviewed above (e.g., Dwight & Donovan, 2003; Jackson et al., 2000; Li & Bagger, 2006; McFarland et al., 2003; Ones et al., 1996; Robie et al., 2010).

It is also important to compare the efficacy of the different preventive interventions. Limiting faking intention (i.e., warnings) assume that the warnings are equally received by all test-takers. Additionally, warnings may be problematic because they insinuate that the assessment administrator has a direct measure of

faking. The above review of remedial interventions suggests that this claim is tenuous at best. Despite the problems, warnings are typically a useful tool to limit faking. Alternatively, modifying faking ability directly addresses aspects of the assessment that are more easily faked. Forced choice formats in particular make it impossible to endorse every positive item on the assessment. However, time limits have limited support in the literature. Very few studies directly compare intervention strategies. Converse et al. (2008) utilized a 2 (forced choice vs. Likert) x 2 (warning vs. no warning) design but did not find clear differences across conditions. Other studies similarly find inconsistent results with direct comparisons of the interventions (e.g., Day, 2008; Ramakrishnan, 2005; Vasilopoulos et al., 2005). At this point, solid conclusions have yet to be reached on the most preferable preventive intervention.

Rationale

This research investigates the effectiveness of interventions for addressing faking on personality assessments. Given that the preponderance of assessment data is collected by way of self-report, individuals have the opportunity to present themselves in a positive light with little chance of being caught. Thus, job applicants can be expected to positively distort their responses to the extent that they believe this behavior will result in obtaining a valued outcome, such as a job offer. Prevalence estimates vary, but studies tend to suggest that around 30% of applicants can be classified as fakers (Griffith & Converse, 2011).

The number of potential fakers is quite alarming given that organizations utilize personality assessments under the assumption that they are tapping into job-relevant traits. However, research suggests that faking can influence mean

trait scores (Converse et al., 2008), the rank-order of applicants (Rosse et al., 2008), and even the criterion-related validity of the test scores (Mueller-Hanson et al., 2003). The utility of the selection system is critically undermined to the extent that any or all of these factors are altered. This causes organizations to look for ways to ameliorate such deleterious outcomes, and research evidence implicates some interventions that may be viable. This dissertation seeks to identify which interventions are most useful for addressing the problem of faking in order to guide organizational decision-making and organize the faking literature to reveal important avenues for future research.

There are two main categories of interventions: preventive and remedial. The previous sections reviewed the extant literature on five specific interventions that fall into these two categories. Warning statements are theorized to influence behavior by affecting the *intent* to fake, while forced choice formats, time limits, and item transparency influence the test-taker's *ability* to fake. These interventions can influence both mean test scores as well as the ability to make inferences from these scores. An administrator could also implement remedial interventions after the scores are collected, such as score corrections or case removal based on scores on a social desirability scale. The primary purpose of remedial interventions is to try and enhance the ability to make valid inferences from the test scores.

Overall, there is far less empirical support for remedial approaches than for preventive ones. This is likely because remedial approaches fail to influence actual test-taker behavior. By not altering behavior during the assessment process,

remedial approaches are less desirable interventions than the preventive approaches. Preventive approaches, on the other hand, can influence test-taker behavior by targeting either their *intent* or *ability* to fake. Remedial approaches also depend on a direct measure of faking (i.e., a social desirability scale). These scales are more useful for measuring a predisposition to fake than actual faking behavior, and also tend to share meaningful variance with traits such as conscientiousness (Smith & McDaniel, 2011). Other methods for operationalizing faking, such as increased mean dimension scores on desirable traits, are more effective at capturing the behavior.

This dissertation offers a unique contribution to the field by comparing different faking interventions. This helps to provide a more nuanced perspective on not only which interventions are effective at addressing faking, but which interventions are *most* effective. Considering the potential for significant organizational costs due to faking, this research has strong implications for practitioners. An organization may lose over $2,000 per worker per year by selecting a less optimal candidate as a result of faking (Komar et al, 2008). The results of this research can be used by practitioners and organizational leaders to better position themselves and their selection tests to minimize the impact of faking. This dissertation also informs faking research by highlighting which interventions are most effective at limiting faking, and can drive future research examining the underlying mechanisms of these interventions in greater detail. To this end, Figure 3 depicts where interventions fit into a framework of faking behavior based predominantly on TPB (Ajzen, 1998; McFarland, 2003).

The overriding objective of this study is to consolidate the growing literature on faking and provide summary statements about the effectiveness of different interventions for limiting the behavior. This study uses meta-analysis to "quantitatively aggregate the results of multiple primary studies" (Arthur, Bennett, & Huffcutt, 2001, p. 8). This methodology allows for testing relationships outside of the scope of primary studies, such as comparing the efficacy of different interventions, and moderators such as study design (i.e., within vs. between subjects).

Statement of Hypotheses

Effect Size Hypotheses

Hypothesis I: Effect sizes for FFM trait scores between honest and faked conditions will be smaller for studies that issue a warning than studies that have no intervention.

Hypothesis II: Effect sizes for FFM trait scores between honest and faked conditions will be smaller for studies that use a forced-choice test as compared to those who use a single-stimulus (e.g., Likert) test.

Hypothesis III: Effect sizes for FFM trait scores between honest and faked conditions will be smaller for studies that use less transparent items as compared to those that use more transparent items.

Hypothesis IV: Effect sizes for FFM trait scores between honest and faked conditions will be larger for studies that use a time limit intervention compared to A) warnings, B) Forced-Choice measures, and C) less transparent items.

Hypothesis V: Effect sizes for FFM trait scores between honest and faked conditions will be smaller for preventive interventions focused on intent (i.e.,

warnings) than preventive interventions focused on ability (i.e., forced choice, item transparency, time limits).

Criterion-Related Validity Hypotheses

Hypothesis VI: The criterion-related validity for personality scores will be higher when test-takers are given a warning compared to A) correcting personality scores for social desirability or B) removing scores based on high scores to a social desirability scale.

Hypothesis VII: The criterion-related validity for personality scores will be higher when test-takers are given a forced choice scale compared to A) correcting personality scores for social desirability or B) removing scores based on high scores to a social desirability scale.

Hypothesis VIII: The criterion-related validity for personality scores will be higher when test-takers are given less transparent items compared to A) correcting personality scores for social desirability or B) removing scores based on high scores to a social desirability scale.

Method

Two separate meta-analyses were conducted for the two types of effect sizes, namely standardized mean differences (*d*'s) and correlations (*r*'s). A meta-analysis of *d*s was conducted to test the extent to which a faking intervention limited the increase of mean personality scores between honest and faked conditions. A meta-analysis of *r*'s was conducted to test the extent to which the criterion-related validity of personality test scores changed based on the presence of a faking intervention. This chapter describes the search strategy, inclusion criteria, coding of the variables, and analytical strategy.

Search Strategy

The following databases were searched to identify articles related to faking and personality: PsycINFO, PsycArticles, Business Source Complete, Social Sciences Citation Index (SSCI), Google Scholar, and ProQuest Dissertations and Theses Full Text. Databases were searched including the years 1950 to June 2013, with the exception of the SSCI that only includes 1985 to present. The following search terms were used: "faking," "personality," "response distortion," "fake good," "impression management," "social desirability," "self-presentation," "intentional distortion," "warnings," "instructions," "forced-choice," "ipsative," "response latency," "response time," "item placement," "lie scale," and "bogus items". All of the searches included at least the terms "personality" and a combination of "faking or response distortion or self-presentation or intentional distortion." This search combination provided many of the articles that did not use any intervention. Subsequent inclusions of "warnings" or "instructions" in the next search provided many of the warning intervention studies. This pattern was followed for each of the hypothesized interventions. The same pattern was followed in each of the six databases. Because most research on faking and personality did not begin until after 1950, databases were not searched prior to 1950. During the initial article consolidation, articles were screened on the surface level for two factors: empiricism and adequacy of sample. For the first factor, book chapters, summaries, or other theoretical pieces that did not include necessary meta-analytic information upon initial inspection were excluded. If the source provided any tabular representation of data, it was retained for possible

inclusion. For the second factor, any articles that were expressly done in a clinical setting based on the title or abstract were excluded from the initial list. No other review for inclusion criteria (discussed below) were performed during the initial search. The search identified 311 potential sources.

Next, key journals between January 2008 and December 2012 were electronically searched via PsycInfo or through the journal's website to assess whether the keywords failed to identify any relevant articles. I chose these years in order to focus on these recently published articles because they may have been in press in other reference lists or searches. The journals searched were those likely to include research on faking: *Academy of Management Journal, Academy of Management Review, Human Performance, International Journal of Selection and Assessment, Journal of Applied Psychology, Journal of Business and Psychology, Journal of Management, Journal of Personality and Social Psychology, Journal of Vocational Behavior, Organizational Behavior* and *Human Decision Processes, and Personnel Psychology*. No additional articles were identified through the manual search.

Searches for unpublished research were performed via two sources: online conference programs and internet searches of departmental websites for faculty and universities that regularly conduct faking research. Conference programs were searched for papers presented within the last five years (2008 to 2013) at the annual conferences of the Society of Industrial and Organizational Psychology (SIOP), International Personnel Assessment Council (IPAC), and Academy of Management (AOM). Digital programs were searched using the aforementioned

keywords. Each unpublished work was cross-referenced with the existing list of published or in-press articles to ensure that it had not been published since its presentation date. A total of 49 unpublished sources were identified for potential inclusion in the meta-analysis. Of these, copies of 11 articles could not be located through online searches (e.g., through departmental or faculty websites). Authors were contacted to request a copy of the paper if the article could not be found. Five of these articles were provided from the authors, while 6 articles were unable to be obtained. The searches of unpublished research resulted in an additional 43 sources for possible inclusion.

Finally, I searched the reference lists of existing meta-analyses and literature reviews on faking and personality (i.e., Birkeland et al., 2006; Dwight & Donovan, 2003; Jenson & Sackett, 2012; Li & Bragger, 2006; Ones et al., 1993; Stanush, 1997; Viswesvaran & Ones, 1999). This search identified 17 additional sources for potential inclusion. The majority of these were older conference presentations and unpublished technical reports. University libraries and other sources were used to obtain copies of the additional sources, and primary authors were again contacted if the article could not be obtained. Only 4 of these articles or reports were located and added to the list of eligible studies.

In total, the above search strategy identified 358 potential sources for inclusion in the meta-analysis. This included potential studies for both the meta-analysis of effect sizes (i.e., the *d* dataset) and the meta-analysis of correlations (i.e., the *r* dataset). The next step was to identify the relevance of the articles for the present analyses as outlined below.

Inclusion Criteria

In order to be included, a study had to (a) use a normal adult population (i.e., non-institutional, at least 18 years old), (b) employ a personality scale that measured at least one trait in the FFM, (c) provide an effect size or the data necessary to compute one, and (d) compare an honest to faked condition in the *d* dataset. More detail is provided on the inclusion criteria in the remainder of this section.

Participants. Because this dissertation is focused on the applicability of faking to personnel selection, studies were included only if the participants were from an adult population or were comparable to a working population. Thus, college samples were included but any samples from grade school or high school were excluded. Studies with participants from clinical settings were also excluded; for example, if they focused on the use of the personality assessment for diagnosing pathological disorders. Because measures such as the MMPI can be used in both clinical and personnel selection settings, the participants were the primary factor in determining a study's eligibility rather than the measure used.

Five-Factor Model personality scale. Each study included in the meta-analysis was required to examine faking in regards to at least one Five-Factor Model (FFM) traits. In other words, measures developed around the FFM or Big 5 *a priori* (e.g., NEO Five Factor Inventory, International Personality Item Pool, Big Five Inventory, and Hogan Personality Inventory) met this criteria. Measures that assessed personality outside of the FFM were included as well, provided the measure could be converted to the FFM using the taxonomy provided by Hough

and Ones (2002), Birkeland et al. (2006), or other research provided on the scale. The test dimension to construct mapping can be found in Appendix A. Whether factors on the scale were aligned with the FFM *a priori* and *post hoc* was coded as a potential moderator. Measures that could not be mapped onto a FFM trait were excluded from the analyses.

Effect size data. A primary study was included in the meta-analyses if it provided an effect size estimate of the amount of faking between honest and faked conditions, or sufficient information to calculate an effect size according to standard practice (e.g., Arthur, Bennett, & Huffcuttt, 2001; Morris & Deshon, 2002). Acceptable reported statistics include *t*-statistics, *F*-statistics, or means, standard deviations, and sample sizes. Effect size calculations and conversions are discussed in the Results section.

Comparing honest to faked conditions. To be included in the *d* dataset (i.e., the data measuring mean differences between faking and honest conditions), a primary study had to compare participant scores in honest and faked conditions. Both laboratory experiments (instructionally-induced faking) and applied organizational studies (comparing applicants to incumbents) were included. The comparison of an honest to faked condition was not necessary for the *r* dataset (i.e., the hypotheses assessing criterion-related validity). To be included in the *r* dataset, a study had to both (a) include a faking intervention and (b) have a faking condition (either instructionally-induced or naturally-occurring). The *r* analysis was focused on the ability to make valid inferences from test scores under applicant or simulated applicant conditions in the presence of an intervention. As

such, these hypotheses were not focused on the existence of faking (i.e., shift in mean scores from honest to faked conditions), but rather how motivated/faked scores could be used to predict organizational criteria.

The exclusion of fake bad studies. This meta-analysis is focused on the use of personality assessments for employee selection, and thus is chiefly concerned with positive self-representation. For this reason, "fake bad" studies where participants are instructed to represent themselves in a negative light were excluded from the analysis.

Results of Inclusion Criteria

Of the 358 sources identified through the search processes, 210 sources met the inclusion criteria for the *d* dataset, and 35 met the criteria for the *r* dataset. Most studies were excluded because they did not include an honest and control condition that could be compared. For example, some articles only compared test-takers who received a time limit to those who did not receive a time limit. Another reason for exclusion was a lack of sufficient information to calculate an effect size. Whenever possible, the author(s) were contacted to obtain more information. The 245 retained studies included a total of 1870 effect sizes in the *d* dataset and 179 correlations in the *r* dataset.

All effect sizes were mapped onto one of the broad Five Factor Model (FFM) traits. Some traits, such as self-monitoring or goal orientation, lacked sufficient conceptual or empirical overlap with a single trait in the FFM. Excluding these effect sizes provided a total of 1676 effect sizes distributed across the five personality traits in 210 studies in the *d* dataset, and 223 effect sizes

across 33 sources for the *r* dataset. Multiple measures of the same trait within a study were also collapsed into a single estimate by taking the sample-weighted average of the effect size. For instance, a study may have examined multiple facets of conscientiousness. Because all analyses were done at the broad trait-level, these estimates could not be considered to be independent. This further reduced the dataset to 725 effect sizes from 210 sources in the *d* dataset and 86 effect sizes from 33 sources for the *r* dataset.

Finally, the effect sizes (both *d* and *r*) had to be inspected for independence prior to testing specific hypotheses. Consistent with best practice (Arthur et al., 2002), effect sizes were considered independent if different participants contributed to the effect or if the same participants contributed to effects that represented distinct constructs. Therefore, effect sizes for different traits were considered independent even if they were collected on the same group of participants. Some sources in the *d* dataset ($k = 17$, 8%) reported effect sizes comparing the same honest condition (i.e., control) to different faking (i.e., experimental) conditions. For example, some studies compared an honest condition to a faked condition with a) no intervention or b) with a warning. In these cases, the effect sizes were considered independent as a function of the level of the moderator analysis. Effect sizes were collapsed across interventions for overall estimates such that a sample would only contribute once to the specific analysis. Dependent effect sizes were combined and represented in the independent dataset by a sample-weighted average of the effect size.

After all of the above steps, a total of 610 independent effect sizes from 210 sources were available in the *d* dataset, while the *r* dataset included 79 independent effect sizes (*r*'s) across 33 sources.

Coding Empirical Study Characteristics

A codebook was developed and can be found in Appendix B. A variety of characteristics needed to be accounted for in the proposed study based on the extant literature (e.g., Hooper, 2007; Ones et al., 1996; Viswesvaran & Ones, 1999). In addition to the hypothesized moderators (i.e., faking interventions), several other moderators were coded in the meta-analysis. Moderators are discussed in the following section.

Faking intervention. The purpose of this meta-analysis is to study the effectiveness of various interventions for a) curbing the response inflation or faking and b) enhancing the ability to make valid inferences from test scores. Thus, the most important piece of this meta-analysis was to appropriately code effect sizes within each intervention. If a study did not meet any of the criteria for the 5 interventions discussed below, it was coded as "no intervention." For preventive interventions, the difference between honest and faked mean scores was used to calculate the effect size for faking in the *d* dataset, and the correlation of a personality score with a criterion in a motivated condition (either instructionally-induced or naturally-occurring) was used in the *r* dataset. Remedial interventions were only applicable to the *r* dataset due to the study's hypotheses. In the *d* dataset, most effect sizes (517) were available for the studies with no intervention.

Preventive interventions – warnings. Studies were coded as either including a warning or no (code of "1" or "0"). Warning "type" coding followed Pace and Borman's (2006) five category taxonomy: detection, consequential, appeal to reason, educational, and appeal to moral principles. Detection warnings only stated that "faking can be detected," while consequential warnings offered more of a threat (e.g., "Your application will be removed from the pool"). Appealing to reason warnings argued that responding honestly is more appropriate (e.g., "more accurately portray your personality") while appealing to moral principles stated that faking is wrong (e.g., "as a moral person, it is wrong to distort your responses"). Finally, educational warnings provided the perspective of the test user or administrator (e.g., "We will not be able to evaluate your responses"). Based on the inclusion criteria, 97 effect sizes were included for warnings in the *d* dataset and 16 effect sizes were included in the *r* dataset.

Preventive interventions - forced-choice. Forced choice (FC) scales were reviewed to differentiate partially ipsative from fully ipsative measures, in accordance with recent reviews highlighting the theoretical and practical differences between the scales (Dilchert & Ones, 2011; Meade, 2004). Coders identified the measure as partially ipsative if it satisfied any of the seven listed characteristics provided by Hicks (1970):

- Respondents only partially order item alternatives, rather than ordering them completely
- Scales have differing number of items
- Not all alternatives ranked by respondents are scored

- Scales are scored differently for respondents with different characteristics, or are referred to different normative transformations on the basis of respondent characteristics
- Scored alternatives are differentially weighted
- One or more of the scales from the ipsative predictor set is deleted when data are analyzed
- The test contains normative sections

In most cases, however, the primary article did not provide sufficient information to make a determination on the seven factors from Hicks (1970). No further moderator analyses were done on level of ipsativity due to lack of available data. Within the FC intervention, 91 effect sizes were included in the *d* dataset and 12 were included in the *r* dataset.

Preventive interventions – time limits. Coders identified the existence of a time limit on a measure. If any time limit was reported in the article, it was recorded by the coders. Because time limits are not typically imposed on personality test, the primary article did not have to categorize the test as a speeded test versus a power test. Any report of time limit met the criteria for a time limit intervention. A dichotomous variable (i.e., "timed" or "not timed") was the primary code for the time limit intervention. To the extent that the information was provided, a continuous time limit was also coded. Because the time limit depends in part on scale length, the ratio of time limit to number of items was coded. Only three studies and 15 effect sizes that used a time limit intervention were eligible for this meta-analysis (in the *d* dataset).

Preventive interventions - item transparency. Item transparency can exist in multiple forms. Measures are less transparent with randomized item sets than blocked item sets because the underlying traits are not as easily identified (McFarland et al., 2002). Item content can also be modified to make them less transparent, often done through the use of empirical-keying. Finally, using contextualized items focuses the test-taker on the desired context (e.g., "at work") rather than a more general, ambiguous context. These three approaches to limiting item transparency were theorized to influence faking behavior in a similar manner based on the working model of faking (see Figure 3). Thus, item transparency was coded if the study employed either approach. Coders identified whether (a) item randomization, (b) subtle item content, or (c) contextualized items were implemented. Item randomization was operationalized by the primary author specifically mentioning that items on the measure were randomized. Subtle item content was operationalized by the use of empirical/criterion keyed tests that make the measured construct less verifiable. Additionally, the California Personality Inventory (CPI) satisfied the subtle item intervention because the test was developed with the purpose of including more subtle items (Dilchert & Ones, 2012). Finally, contextualized items were operationalized by an explicit statement of an "at work" frame of reference in the personality measure. For any of the above three options, the study was included in the meta-analysis as "no intervention" if there was no specific reference to any of these three types of item transparency intervention. A total of 79 effect sizes were included in the *d* dataset

within one of the three item transparency interventions. Only 9 effect sizes for item transparency were included in the *r* dataset.

Remedial interventions – corrections. Corrections made to personality trait or scale scores are typically done through social desirability responding (SDR), unlikely virtues scales (UV), or other lie scales. The first coding category for the statistical corrections intervention identified which of these scales was used to correct the score. With regards to SDR, coders identified which of the two factors was measured (impression management and self-deception; Paulhus, 1991). Next, the adjusted personality trait and criterion correlation with SD, UV, or the lie scale partialled out was recorded. If the study did not report the corrected relationship, the semi-partial correlation was obtained based on the correlation matrix reported in the article. This provided the relationship between the personality trait and the criterion with social desirability partialled out of the personality measure (*cf*. Ones et al., 1996). A total of 30 effect sizes were included for the score correction intervention in the *r* dataset.

Remedial interventions - removal of cases. Similar to score corrections, test-taker removal as a remedial intervention often uses SDR scales to inform those decisions. If a study removed cases, the scale that was used to remove cases and the name of the SDR scale was coded if applicable. The correlation between the personality trait and the criterion in this "modified" sample was coded as the effect size for studies that used this intervention. Only 12 effect sizes were available for case removal intervention in the *r* dataset.

Study characteristic - study setting. Many researchers maintain laboratory settings fail to fully replicate the conditions and incentives of employee selection (e.g., Dilchert & Ones, 2011; MacCann et al., 2011). Thus, study setting was included to identify whether systematic faking differences were observed in both settings. Studies were coded as either: (a) field, or (b) laboratory setting. Studies were coded as (b) lab if a student sample was used, even if students were instructed to fake "as if they were an applicant."

Study characteristic - study design. Most researchers argue that within-subject designs are more appropriate for testing faking because they do not rely on nonequivalent comparison groups, namely applicants and incumbents (Griffith & McDaniel, 2006). However, practical issues often necessitate the use of between-subject designs. Past meta-analytic findings (e.g., Birkeland et al., 2006; Viswesvaran & Ones, 1996) suggest that within-subject designs show stronger effect sizes for faking, but no research has looked at the efficacy of faking interventions across different study designs.

Study design was coded as (a) between-subjects or (b) within-subjects design. Additional categorizations were made for the latter to control for the order of faking versus honest responding. Three possible orders were coded: (a) honest conditions followed by faking, (b) faking followed by honest condition, or (c) a counterbalanced design. The time lag for within-subject designs was also coded in terms of number of days between the first and second administration.

Study characteristic - type of faking. Instructionally-induced faking sets the upper limit on the extent to which personality trait scores can be changed or

faked (Viswesvaran & Ones, 1996). Instructionally induced faking also rests on the strong assumption that all participants hear the instructions in the same way and that the instruction is sufficient for replicating an application scenario (Holden & Book, 2011). Thus, this study sought to identify if the effectiveness of faking interventions differed across levels of faking fidelity (i.e., instructional vs. natural faking), and faking was coded as either: (a) naturally occurring or (b) instructionally induced.

To be coded as naturally occurring, faking had to be done without any instructions to fake. For example, test-takers taking a personality test as part of a job application were considered "naturally" motivated. Alternatively, the instructionally induced code was assigned to any study where test-takers' responses are altered by the instructions from the assessment administrator or researcher. The more contextualized instructions of "responding as an applicant" may more realistically simulate the application scenario, and give test-takers a clearer frame of reference for their responses. Thus, whether test-takers were instructed to: (a) present themselves in a positive light without any context provided (e.g., Gibby, 2004), or (b) instructed to respond as if they were an applicant (e.g., Robson et al., 2008) was coded.

Following the work of Hooper (2007), an additional distinction was coded within the latter category of "responding as an applicant." Job relevance can influence faking behavior, consistent with previous research within the context of job valence (e.g., Day, 2008). It follows that the extent of faking behavior may depend in part upon how relevant the job is to the individual applicant. Thus, the

specificity of the job context was coded as either: (a) unspecified or (b) specified. Studies were coded as "unspecified job position" if they simply stated to respond as an applicant, while "specified job positions" included those in which test-takers were given a job description or other job content as a referent (Wolford, 2009).

Study characteristic – criterion. For the criterion-related validity portion of the study (i.e., the *r* dataset), different types of criteria were distinguished. Following the work of Ones et al. (1996), coders identified the type of criterion reported within the following categories: (a) school success, (b) task performance, (c) counterproductive or organizational citizenship behaviors, (d) training performance or (e) job performance.

School success was operationalized as any measure of performance in an academic setting. In this meta-analysis, the following were included in this category: ACT score, grade point average, and grades in an academic course. Job performance was most frequently measured by a performance review from a supervisor, where the reference is how the test-taker performed in the context of a work environment. This was frequently done for administrative purposes (i.e., as part of the normal performance review process), but some studies used the performance review for research purposes only. Studies in the latter were most frequently lab studies that sent performance reviews to a test-taker's supervisor. Job performance was differentiated form task performance in that the latter involved things that may be done on a job but the study was *not* in an existing work environment during the typical workday. For example, in one study task performance was measured observer ratings of performance on a task that, while

similar to clerical jobs, was done in the lab for research purposes only (Mueller-Hanson et al., 2003). Behavior on the job was most frequently measured through a counterproductive work behavior (CWB) scale, although it was also collected by self-reported and/or administratively recorded absence or lateness. Finally, training performance was less frequently observed but involved performance in a training-specific context. For example, this was observed for performance in a military training exercise (Fox & Dinur, 1988). This information on defining the criteria can also be found in Appendix A.

Coding the Articles

The dissertation author and one doctoral student in I/O Psychology coded the identified articles. The author trained the second coder on the codebook and provided a coding guidelines sheet (see Appendix A) to assist with making decisions during the coding process. After training, three articles were chosen at random, and both coders coded the three articles independently. The two coders met to identify the source of any discrepancies before moving forward. After coding an initial common set of three sources, each coder independently coded a common set of 10 articles. The 10 articles included at least one example from each faking intervention. A Kappa (K, Cohen, 1960) index was computed to examine the agreement between the two raters. Agreement across study characteristics was .96 and agreement across the 93 effect sizes was .94 for these 10 studies. For study characteristics, coding discrepancies were often slight. For instance, a different total number of items were used to calculate the ratio of seconds to items. Disparities in effect size calculations were mostly attributable to

transcription error in the honest versus control condition. Because the agreement exceeded typical standards (Landis & Koch, 1977), the remaining articles were coded independently.

After aggregating all effect sizes to be analyzed in the study, effect sizes for neuroticism were reversed so that the direction of the effect size would be in the same direction as the other traits. Positive effect sizes reflected higher scores in the faked condition than the honest condition.

Analytical Strategy

Meta-analyses typically correct for unreliability in the predictor and/or the criterion. Because the analyses are focused on the "operational use of personality inventories" (Ones et al., 1996, p. 201), no corrections were made for unreliability in the predictor (i.e., personality measure). However, to provide the operational validity estimate, corrections were made for unreliability in the outcome (i.e., criterion) in the *r* analyses. Only about 25% ($k = 8$) of studies reported criterion-related validity, and this was most often internal consistency measured by coefficient alpha. Due to restricted variability across criterion types, the average criterion-related validity was estimated for each of the moderator x trait analyses, and this estimate was used to correct the observed *r* coefficient. Both the *r* and *d* datasets computed sample-weighted statistics to correct for sampling error and corrected for the attenuating effect of unequal or unbalanced sample sizes. The latter correction was especially appropriate for some studies that had significantly different sample sizes across applicant and incumbent samples. The formulas followed those offered by Hunter and Schmidt (2004).

Credibility intervals were used to estimate the variability of the effect sizes in the meta-analytic sample. If the credibility interval is relatively large, then it is likely that moderators may be present. If the credibility is relatively small, then it is likely that moderators are not in operation. The width of a credibility interval was used as one indicator of the presence of moderator, along with the percent of variance explained.

Confidence intervals were used to estimate the accuracy of the effect size estimate, and these were used to test many of the hypotheses related to the *d* dataset. The confidence interval provides a range within which the mean effect size would likely fall if other studies were selected from the population (Arthur et al., 2002). To this end, the extent that the confidence interval bands around a given effect overlapped with the confidence interval band around a compared effect was used as a primary indicator of a significant difference. If the 95% confidence intervals around the estimates had zero overlap, they were considered significantly different. However, the inverse is not necessarily true; namely, estimates that have overlapping confidence intervals can still be significantly different from each other (*cf*. Cumming, 2009; Cumming & Finch, 2005). The "proportion overlap" (*POL*) was used to further test the effect size estimates that had overlapping confidence intervals. The POL was calculated by first taking the average distance between the two independent estimates and their respective CI_{95}. Take the following as an example as Step 1:

SWMD(1)= 0.79, CI95 = 0.72 to 0.86. (0.79 – 0.72 = 0.07)

SWMD(2) = 0.57, CI95 = 0.41 to 0.74 (0.74 – 0.57 = 0.16)

The average of these two = (0.07+0.16)/2 = 0.12. The next step was to take the distance between the overlapping confidence intervals. In this case, the lower-bound CI_{95} of the larger estimate is subtracted from the upper-bound CI_{95} of the smaller *SWMD* estimate. In the example, this would mean Step 2 is:

0.74 – 0.72 = 0.02

Finally, to obtain the POL value, the distance between the two intervals obtained in Step 2 is divided by the average width of the intervals obtained in Step 1.

0.02/0.12 = 0.17 = 17% *POL*

In line with previous research, independent estimates were considered significantly different if the obtained *POL* value was 50% or smaller (Cumming & Finch, 2005). Based on direction of the hypotheses, this ostensibly meant that less than 50% of the difference between the estimate and the lower-bound confidence interval for studies *without* an intervention could overlap with the upper-bound confidence interval for studies *with* an intervention.

This approach to determining the significance between two estimates has the potential for producing confusing values. For instance, if two estimates do not overlap at all, and thus the value obtained in Step 2 is negative, the resulting *POL* value will be negative. Similarly, if all of one estimate is completely subsumed by another, the resulting *POL* estimate will be greater than 100%. In both cases, the interpretation is more clearly represented by replacing a negative value with 0% and a value over 100% with 100%. In the former, a negative *POL* simply means that there is no overlap between the two estimates, and this meaning is not altered by replacing a negative value with 0%. In the latter, a value greater than 100%

means that there is complete overlap between the estimates, and showing a maximum of 100% does not change this interpretation. This is similar to common approaches in meta-analysis for reporting percentage of variance attributable to sampling error, as values greater than 100% are replaced with 100%.

Finally, it is important to consider when, or if, it is appropriate to combine effect sizes gathered from different study designs. Between- and within-subject effect sizes can be combined into the same meta-analytic estimate only if the study design moderator analysis is not significant (Morris & Deshon, 2002). If the moderator analysis is significant, the study design has a meaningful impact on the effect size in addition to the faking intervention in question. In such cases, a combined effect size estimate may not be appropriate. Previous studies have demonstrated that this is especially relevant to faking meta-analyses, as standardized mean differences between honest and faked conditions may differ substantially across study design (Viswesvaran & Ones, 1999). In cases where study design appears as a meaningful moderator, the meta-analytic results will be presented separately for each study design.

Within the analyses, a positive *d* suggested that the experimental (i.e., faking) condition scored higher than the control (i.e., honest condition) condition, while a negative *d* suggested the control condition scored higher. Similarly, a positive *r* indicated a positive relationship between the personality score and criterion of interest. Cohen's (1992) guidelines of small (.20, .10), medium (.50, .30), and large (.80, .50) effect sizes were used for the *d* and *r* effect sizes, respectively.

Results

Description of the Database

As mentioned previously, two distinct datasets were used for the current meta-analysis: one that examined the standardized mean difference between an honest and faked condition (meta-analysis of *d*'s), and another that examined the criterion-related validity of personality test scores (meta-analysis of *r*'s).

In the *d* dataset, conscientiousness was the most included effect size, comprising just over 25% of the included effect sizes across 162 studies. Across all of the traits, a majority of the effect sizes were obtained from studies using lab settings (k = 171, 81%) and student samples (k = 164, 78%), as opposed to field settings (k = 39, 19%) or non-student samples (k = 46, 22%). Because most of the studies measured faking in lab settings, the majority of studies manipulated faking via instructionally-induced faking. Just under half of all effect sizes represented (k = 101, 48%) faking with specific information provided about what the test-taker was faking to achieve. The context of this information ranged from a variety of positions including, but not limited to, customer service (k = 25, 12%), managerial (k = 17, 8%), police or military (k = 18, 9%), and sales (k = 12, 6%). Most of the eligible effect sizes used between-subjects designs (k = 119, 57%), while the majority of within-subjects designs presented the honest test-taking condition before faking (k = 45, 21%). The sample was generally a young working age (M_{age} = 23.16, SD = 4.87), although not all studies (k = 116) presented age data. This is consistent with the preponderance of lab-based, student sample studies included in the data set. Forced choice and warning interventions contributed the most effect sizes to the dataset (k = 30, 14%; k = 23,

11%, respectively). However, the majority of the effect sizes included in the data set were associated with a study that did not use any intervention to address faking (k = 135, 64 %).

In the *r* dataset, conscientiousness was also the most included effect size (k = 31), comprising over 40% of the independent effect sizes. Job performance was the most frequently cited criterion (k = 13, 38%). Most (k = 11) of these studies measured job performance by supervisor or 360 ratings, while the remaining two studies operationalized job performance by a self-report measure. Most (k = 6) of the supervisor and 360 ratings were for collected as part of formal, administrative processes, with the remaining (k = 5) collected for research purposes only. The second largest criteria group for the *r* dataset was academic performance (k = 10, 29%). Most of the academic criteria were measured using grade point average (k = 7), while the remaining were a measured by ACT scores and class grades. In cases when the expected relationship was negative (e.g., conscientiousness with turnover or counter-productive work behaviors), the direction of the correlation was reversed to facilitate aggregation. Across all of the traits, a majority of the effect sizes were obtained from studies using lab settings (k = 21, 62%) and student samples (k = 20, 59%), as opposed to field settings (k = 13, 38%) or non-student samples (k = 14, 41%). Because most of the studies measured faking in the lab, the majority measured faking via instructionally-induced faking. Of the entire sample of articles, just under half (k = 15, 44%) measured faking with specific information provided about what the test-taker was faking to achieve. Most of the eligible effect sizes used between-subjects designs

(k = 19, 56%). The sample was generally a young working age (M_{age} = 24.96, SD = 4.52), although not all studies presented age data. This is consistent with the preponderance of lab-based, student sample studies included in the data set. Correction for social desirability was the most frequent intervention (k = 7, 21%).

Preliminary analyses. Prior to testing the espoused hypotheses, the data on effect sizes (i.e., the d dataset) were analyzed to compare the included effect sizes to previous meta-analyses on faking. The purpose of this dissertation is not to test whether participants can fake on personality test, as this has been established by the extant literature. However, comparing the standardized mean difference between honest and faked conditions in the current meta-analytic sample to previous estimates provides an important context for testing the hypotheses. It also informs which study characteristic moderators are most relevant for further inquiry.

Table 2 shows the standardized mean difference of personality trait scores between honest and motivated conditions across the FFM traits. This table reports effect sizes independent of the presence of an intervention, and compares the results from previous meta-analyses conducted by Viswesvaran and Ones (1999), Birkeland et al. (2006), and Hooper (2007). All three meta-analyses serve as useful contextual comparisons for the current analysis. Other than excluding forced-choice inventories, the inclusion criteria for Hooper (2007) are identical to the current meta-analysis. Birkeland et al. (2007) focused only on applied samples (e.g., applicant versus incumbents), which is why only the "field" settings in the current meta-analysis were compared. Viswesvaran and Ones (1999) also

Table 2

Standardized mean difference of personality scores between honest and faked conditions compared to previous meta-analyses.

			Current Meta		Vis. & Ones (1999)		Birk. et al. (2006)		Hooper (2007)	
Trait	Setting	Design	*k*	*d*	*k*	*d*	*k*	*d*	*k*	*d*
Extraversion	Field		28	0.27	-	-	29	0.11	26	0.42
	Lab		165	0.47	-	-	-	-	99	0.63
		Between	85	0.48	15	0.63	-	-	-	-
		Within	80	0.45	10	0.54	-	-	-	-
Agreeableness	Field		25	0.21	-	-	20	0.16	18	0.52
	Lab		138	0.41	-	-	-	-	57	0.57
		Between	68	0.41	17	0.48	-	-	-	-
		Within	70	0.40	14	0.47	-	-	-	-
Openness	Field		20	0.15	-	-	20	0.13	16	0.23
	Lab		111	0.36	-	-	-	-	46	0.37
		Between	62	0.36	11	0.65	-	-	-	-
		Within	49	0.34	9	0.76	-	-	-	-
Emotional Stability	Field		24	0.61	-	-	25	0.44	24	0.75
	Lab		144	0.82	-	-	-	-	98	0.93
		Between	88	0.85	17	0.64	-	-	-	-
		Within	56	0.76	29	0.93	-	-	-	-
Conscientiousness	Field		38	0.42	-	-	27	0.45	27	0.75
	Lab		179	0.82	-	-	-	-	75	1.07
		Between	93	0.87	29	0.6	-	-	-	-
		Within	86	0.75	24	0.89	-	-	-	-

Note: Effect sizes represent the sample-weighted mean *d* between trait scores in honest versus faked conditions. Higher, positive values of *d* represent higher mean scores in the faked condition than the honest condition. Vis. & Ones (1999) = Viswesvaran & Ones (1999). Birk. et al. (2006) = Birkeland et al. (2006).

included fake-bad studies in their meta-analysis, but these effect sizes are not shown in Table 2. These previous meta-analyses computed the sample-weighted mean effect size after accounting for sampling error, and only Birkeland et al. (2006) also corrected for unreliability in the predictor.

As Table 2 shows, effect sizes in previous meta-analyses varied significantly. Likewise, the estimates in the current meta-analysis vary slightly from those in previous research. Although they vary slightly, sample-weighted mean *d*'s (*SWMD*) of honest versus motivated/faked trait scores in the current meta-analysis tend to align with the extant literature. All four meta-analytic sources report consistently larger effect sizes in lab samples than field samples. The *SWMD* for conscientiousness in lab studies (k = 179, d = .821) is twice as large as the *SWMD* in field studies (k = 38, d = .417). The *SWMD* for field studies found in the current meta-analysis are more in line with those reported by Birkeland et al. (2006) than by Hooper (2007). For example, the *SWMD* for conscientiousness in the current meta-analysis (k = 38, d = .417) is smaller than results reported by Hooper (k = 27, d = .75), but closer to those reported by Birkeland and colleagues (k = 27, d = .45).

The main inconsistency between the current meta-analysis and previous analyses is observed for study design. Viswesvaran and Ones (1999) found that *SWMD* estimates tended to be higher in within-subject designs than between subject designs. In other words, there was a larger shift in personality trait scores from honest to faked conditions for studies that used a within-subjects design versus a between-subjects design. The current meta-analysis, on the other hand,

found consistently larger effects in between-subject designs. For example, *SWMD* estimates in the current analysis are quite similar for emotional stability in between-subjects (k = 88, d = 0.85) and within-subjects (d = 56, k = 0.76). However, Viswesvaran and Ones (1999) show a large difference for effect sizes in between-subjects (k = 17, d = 0.64) and within-subject (k = 29, d = 0.93) designs. A large part of this difference is likely attributable to larger k sizes in the current meta-analysis

Table 2 showed some discrepancies in the effect sizes across study design in the current meta-analysis compared to previous analyses on faking. Tests for statistical significance based on proportion overlap (*POL*) were not possible because the necessary information was not provided in the previous meta-analysis. However, discrepancies can be observed in the table. For instance, Viswesvaran and Ones (1999) found a larger effect size for conscientiousness in within-subject designs compared to between-subject designs, while the current meta-analysis found the opposite. Therefore, study design (i.e., within-subject and between-subjects design) was inspected within each of the following hypotheses to try and gain more insight on this why this discrepancy was observed. It is important to note that the effect sizes (i.e., standardized mean difference of scores across honest and faked conditions) did not significantly differ across study design *within* the current meta-analysis. Therefore, the estimates could be meaningfully combined within the various analyses (Morris & Deshon, 2002).

Hypothesis I - Warnings

Hypothesis 1 predicted that the *SWMD* for personality scores between honest and faked conditions would be smaller for studies with a warning compared to studies without an intervention. Table 3 shows general support for this hypothesis at a high-level across each of the FFM traits. *SWMD* estimates were consistently smaller for warnings compared to no intervention. Conscientiousness and emotional stability consistently showed the largest *SWMD* across warning and no intervention compared to other traits, suggesting more faking on these traits. The *SWMD* for agreeableness, extraversion, and openness for warning studies were not only smaller than studies without an intervention, but the confidence intervals did not overlap. For instance, the upper-bound confidence interval for the observed *SWMD* for agreeableness under a warned condition ($k = 16$, $d = 0.20$, $CI_{95} = 0.11$ - .28) did not include the observed *SWMD* for agreeableness with no intervention ($k = 103$, $d = 0.40$, $CI_{95} = 0.32$ - .50), which suggests that the difference in effect sizes in such cases is particularly meaningful. Even though the confidence intervals overlapped for emotional stability for warnings ($k = 22$, $d = .57$, $CI_{95} = 0.41 – 0.74$) and no intervention ($k = 111$, $d = 0.79$, $CI_{95} = 0.72 – 0.86$), the overlap was not substantial. The proportion overlap (*POL*) was only 17% for the two estimates. Because this is less than 50%, it suggests that the difference in the estimates should be considered significant (Cumming & Finch, 2005). Conscientiousness showed the smallest difference between no intervention ($k = 135$, $d = .62$) and warning ($k = 23$, $d = .59$), and the confidence interval of the warning estimate included the estimate for no

Table 3

Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Warnings Versus No Intervention

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness											
	No Intervention	64,452	103	0.40	0.40	4%	0.32	0.48	0.12	1.06	
	Warning	1,853	16	0.20	0.17	55%	0.11	0.28	-0.21	0.79	0%*
Conscientiousness											
	No Intervention	88,778	135	0.62	0.42	4%	0.55	0.69	0.08	1.15	
	Warning	3,330	23	0.59	0.37	18%	0.44	0.74	0.07	1.07	100%
Emotional Stability											
	No Intervention	38,511	111	0.79	0.39	7%	0.72	0.86	-0.11	0.91	
	Warning	2,861	22	0.57	0.39	17%	0.41	0.74	-0.35	0.49	17%*
Extraversion											
	No Intervention	62,167	122	0.42	0.34	6%	0.36	0.48	0.28	1.29	
	Warning	2,602	21	0.22	0.22	40%	0.12	0.31	-0.07	0.50	0%*
Openness											
	No Intervention	26,921	81	0.29	0.39	7%	0.20	0.37	-0.02	0.42	
	Warning	1,717	15	0.07	0.33	25%	-0.10	0.23	-0.02	0.85	24%*

Note. *N* = total sample size across all effect sizes; *k* = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; *POL* =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for *POL* means that the confidence intervals did not significantly overlap.

intervention (*POL* = 100%). Hypothesis 1 was therefore generally supported, with the exception of conscientiousness.

The lower bound of the 80% credibility intervals did not always exclude zero. Credibility intervals for emotional stability (-0.11, -0.35) and openness (-0.02, -0.02) included zero for both no intervention and warning conditions, respectively. Further, the credibility intervals were relatively large for the effect sizes without an intervention, suggesting that the magnitude of the effect sizes varies meaningfully across studies and that moderators may be present. Although the 80% credibility intervals were more modest for effect sizes with a warning, moderators were investigated within both.

Study design as a moderator of the effect size between honest and faked scores with a warning intervention. Study design was the first moderator investigated, given the known differences observed in *SWMD*s of scores across honest and faked conditions in the current meta-analysis compared to previous studies (see Table 2). As seen in Table 4, *SWMDs* varied substantially across study design. The results for warnings across study design (Table 3) showed that the 95% confidence intervals did not overlap across warning and no intervention studies for agreeableness, extraversion, and openness, and had minimal overlap for emotional stability. However, the same observation was only observed in between-subjects designs for agreeableness and extraversion and within-subject designs for openness. In other words, the 95% confidence interval on the *SWMD* for extraversion in warned, between-subjects designs (k = 14, d = 0.14, CI_{95} = 0.03 - .25) did not overlap with the same level of moderator observed for studies

with no intervention (k = 74, d = 0.41, CI_{95} = 0.33 - .48). Similar to the findings presented in Table 3, the confidence intervals for the emotional stability estimates had minimal overlap for both between (*POL* = 41%) and within-subject (*POL* = 53%) designs. In general, the study design moderator did not illuminate any new findings with regards to the usefulness of warnings. The difference in effect sizes between warnings and no interventions observed within each trait remained relatively consistent. This moderator showed continued support for Hypothesis 1, although the differences were still not strong or significant for conscientiousness. Although not as illuminating for warning effectiveness, this moderator analysis yielded useful insights in regards to the previous meta-analysis comparisons observed in Table 2. The effect sizes of scores between honest and faked conditions across study design reported in Viswesvaran and Ones (1999) more closely match the effect sizes from studies without any intervention in the current meta-analysis. For example, between-subjects (k = 74, d = 0.77) and within-subject (k = 37, d = 0.91) with no intervention present are much closer in magnitude to those reported in Viswesvaran and Ones for between-subjects (k = 17, d = 0.64) and within-subject (k = 29, d = 0.93) designs. This suggests that, outside of the large difference in k sizes, at least part of the large differences observed in Table 2 may be attributable to the inclusion of effect sizes with a faking intervention (in this case, a warning).

Type of warning as a moderator of the effect size between honest and faked scores. Recent theory suggests that more attention should be paid to the

Table 4

Study Design as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions Between Warning and No Intervention

Trait	Intervention	Study Design	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
								Lower	Upper	Lower	Upper	
Agreeableness												
	No Intervention											
		Between	57,511	59	0.40	0.42	3%	0.29	0.5	-0.13	0.93	
		Within	6,941	44	0.40	0.24	32%	0.33	0.47	0.1	0.7	
	Warning											
		Between	1,311	11	0.12	0.14	65%	0.04	0.2	-0.05	0.29	0%*
		Within	543	5	0.38	0.08	84%	0.31	0.46	0.27	0.49	100%
Conscientiousness												
	No Intervention											
		Between	78,836	76	0.60	0.40	2%	0.51	0.69	0.09	1.12	
		Within	9,942	59	0.71	0.50	9%	0.58	0.84	0.07	1.36	
	Warning											
		Between	2,298	17	0.57	0.41	15%	0.37	0.77	0.04	1.1	100%
		Within	1,032	6	0.64	0.23	31%	0.45	0.82	0.34	0.94	100%
Emotional Stability												
	No Intervention											
		Between	33,790	74	0.77	0.37	6%	0.69	0.86	0.29	1.25	
		Within	4,721	37	0.91	0.50	12%	0.75	1.07	0.28	1.55	
	Warning											
		Between	1,830	16	0.53	0.44	16%	0.32	0.75	-0.04	1.1	41%*
		Within	1,032	6	0.64	0.27	26%	0.42	0.85	0.3	0.98	53%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

Table 4

Study Design as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions Between Warning and No Intervention (continued)

Trait	Intervention	Study Design	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
								Lower	Upper	Lower	Upper	
Extraversion												
	No Intervention											
		Between	55,278	74	0.41	0.33	5%	0.33	0.48	-0.02	0.83	
		Within	6,889	48	0.49	0.38	17%	0.38	0.6	0	0.98	
	Warning											
		Between	1,547	14	0.14	0.20	47%	0.03	0.25	-0.12	0.4	0%*
		Within	1,056	7	0.33	0.19	42%	0.19	0.48	0.08	0.58	79%
Openness												
	No Intervention											
		Between	22,951	51	0.27	0.39	6%	0.16	0.38	-0.23	0.77	
		Within	3,970	30	0.41	0.36	20%	0.28	0.53	-0.05	0.86	
	Warning											
		Between	1,175	10	0.01	0.38	19%	-0.23	0.24	-0.48	0.49	46%*
		Within	543	5	0.20	0.06	90%	0.14	0.26	0.12	0.28	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

text of the warning, and not just the presence versus absence of a warning (Pace & Borman, 2006). Most of the existing literature on warnings has focused on the latter. In other words, most empirical studies examine the extent of faking for participants who are given a warning versus those who are not. As mentioned earlier, the type of warning may play a role in the effectiveness of a warning statement. Some warnings simply alert the test-taker to the *detection* of faking (i.e., faked responses can be detected), others offer a *consequence* of faking (i.e., being removed from the applicant pool), while still others offer some *educational* warning (i.e., faking does not allow for accurate measurement). The taxonomy put forth by Pace and Borman (2006) also identified *appeal to moral principles* and *appeal to reason* as other warning types, but there were not sufficient empirical studies to be included in the current meta-analysis. Table 5 reports the results of the warning type moderator, and the analysis is limited to the trait-level due to the available study size.

There are two main findings from this moderator analysis. First, with the exception of the *detection* warning for conscientiousness, all warning types across the FFM reported smaller *SWMDs* (i.e., smaller mean score differences between honest and faked conditions) than studies without any intervention. This suggests that warning types, with the exception of detection warning are generally effective at limiting response distortion. However, the 95% confidence intervals for warning effect size estimates tend to have significant overlap with those without any intervention (e.g., all *POL* > 100% for conscientiousness. The effect sizes for *educational* warning in emotional stability and extraversion represent an

Table 5

Warning Type as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions

Trait	Warning Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness											
	No Interv.	64,452	103	0.4	0.4	4%	0.32	0.48	0.12	1.06	
	Consequence	595	5	0.34	0.17	54%	0.19	0.49	0.12	0.56	100%
	Detection	800	7	0.13	0	100%	0.13	0.13	0.13	0.13	0%*
	Educational	459	4	0.14	0.21	44%	-0.07	0.35	-0.13	0.41	21%*
Conscientiousness											
	No Interv.	88,778	135	0.62	0.42	4%	0.55	0.69	0.08	1.15	
	Consequence	1,528	7	0.54	0.31	17%	0.31	0.77	0.15	0.94	100%
	Detection	964	9	0.75	0.3	31%	0.55	0.94	0.37	1.13	100%
	Educational	742	6	0.53	0.5	12%	0.13	0.93	-0.11	1.17	100%
Emotional Stability											
	No Interv.	38,511	111	0.79	0.39	7%	0.72	0.86	-0.11	0.91	
	Consequence	1,084	6	0.58	0.28	23%	0.35	0.8	0.22	0.94	55%
	Detection	1,036	10	0.74	0.48	15%	0.44	1.03	0.12	1.35	100%
	Educational	742	6	0.33	0.25	35%	0.13	0.53	0.01	0.65	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap; Interv. = intervention.

Table 5

Warning Type as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions (continued)

Trait	Warning Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		POL
							Lower	Upper	Lower	Upper	
Extraversion											
	No Interv.	62,167	122	0.42	0.34	7%	0.36	0.48	0.28	1.29	
	Consequence	1,084	6	0.24	0.25	26%	0.04	0.44	-0.09	0.57	62%
	Detection	1,060	11	0.23	0.23	43%	0.09	0.37	-0.07	0.53	10%*
	Educational	459	4	0.15	0	100%	0.15	0.15	0.15	0.15	0%*
Openness											
	No Interv.	26,921	81	0.29	0.39	7%	0.2	0.37	-0.02	0.42	
	Consequence	595	5	0.25	0.37	20%	-0.08	0.58	-0.23	0.73	100%
	Detection	664	6	-0.21	0.13	68%	-0.32	-0.11	-0.38	-0.05	0%*
	Educational	459	4	0.24	0.06	92%	0.19	0.3	0.17	0.32	100%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap. Interv. = intervention.

exception, as the confidence intervals around these estimates do not overlap at all. The effect size of emotional stability with an *educational* warning ($k = 6$, $d = 0.33$, $CI_{95} = 0.13$ - .53) does not overlap with the same trait without an intervention ($k = 111$, $d = 0.79$, $CI_{95} = 0.72$ - .86). Although the trend of the data suggests that *educational* warnings are more effective than other warning types, the small *k* size for this warning type suggests more research is needed.

The second main finding from this moderator analysis is that effect sizes are inconsistent between the *consequences* warning compared to *detection* warning. *Consequence* effect sizes were smaller than *detection* for conscientiousness ($d = 0.54$ and 0.75, respectively) and emotional stability ($d = 0.58$ and 0.74, respectively). This finding makes intuitive sense, as a consequence warning both acknowledges the ability to detect faking as well as offers an undesirable corrective action to the test-taker. However, the opposite relationship was observed for agreeableness, extraversion, and openness, where *consequence* effect sizes were larger than *detection.* This inconsistent pattern may be attributable to the very small *k* size within the warning type moderator, as they ranged from 4 to 11 studies. Based on limited research, warning "type" does not have a consistent effect on faking behavior across personality traits.

Summary of warning hypothesis. Hypothesis I predicted that the standardized mean difference of personality scores between honest and faked conditions would be smaller (i.e., less faking) for studies with a warning than studies without any intervention. In general, this hypothesis was supported. Effect sizes tended to be smaller in the presence of a warning, and this relationship was

observed across study design as well as the type of warning supplied. Conscientiousness was one exception among the FFM, as the estimate in warning studies tended to significantly overlap with the estimate in studies without any intervention. Additional exceptions were observed with *detection* warnings and between-subject designs for conscientiousness, but the majority of the evidence supports the hypothesis.

Hypothesis II – Forced-Choice

Hypothesis II predicted that effect sizes (representing the difference between scores in honest and faked conditions) obtained from a forced-choice (FC) measure would be smaller than effect sizes obtained in Likert-style measures. Table 6 compares the *SWMDs* for forced choice versus no intervention at the trait-level. Similar to the results presented for Hypothesis 1, the largest *SWMD* estimates for FC measures were observed for conscientiousness ($k = 30$, $d = 0.70$) and emotional stability ($k = 14$, $d = 0.40$). This suggests that conscientiousness was the most faked trait. Further, the *SWMD* for conscientiousness was larger than the *SWMD* for studies without an intervention ($k = 135$, $d = 0.62$). In contrast, the 95% confidence interval on the *SWMD* for emotional stability in FC measures ($CI_{95} = 0.22$ to .57) did not overlap with the effect size for no intervention ($d = 0.79$, $CI_{95} = 0.72$ to .86). The confidence interval around the estimate for agreeableness had minimal overlap (*POL* = 46%) with the effect size obtained for studies without an intervention ($k = 103$, $d = 0.40$, $CI_{95} = 0.32$ to .48), suggesting that the effect size was meaningfully smaller than

Table 6

Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Forced-Choice versus No Intervention

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness											
	No Intervention	64,452	103	0.40	0.56	8%	0.32	0.48	-0.32	1.12	
	Forced Choice	2,951	19	0.15	0.56	8%	-0.11	0.40	-0.57	0.86	46%*
Conscientiousness											
	No Intervention	88,778	135	0.62	0.42	4%	0.55	0.69	0.08	1.15	
	Forced Choice	4,753	30	0.70	0.49	10%	0.52	0.87	0.07	1.32	100%
Emotional Stability											
	No Intervention	38,511	111	0.79	0.39	7%	0.72	0.86	0.28	1.29	
	Forced Choice	2,589	14	0.40	0.33	17%	0.22	0.57	-0.03	0.82	0%*
Extraversion											
	No Intervention	62,167	122	0.42	0.34	7%	0.36	0.48	-0.02	0.85	
	Forced Choice	3,165	25	0.30	0.35	21%	0.16	0.43	-0.15	0.74	77%
Openness											
	No Intervention	26,921	81	0.29	0.39	7%	0.20	0.37	-0.21	0.79	
	Forced Choice	2,591	19	0.31	0.42	15%	0.13	0.50	-0.22	0.85	100%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

the analog effect size without an intervention. As with the previous findings with warnings, however, there is a stark contrast in the included *k* size for the FC intervention. Even though FC had sufficient *k* size to investigate meta-analytically (e.g., $k = 19$ for agreeableness), it is difficult to directly compare to an effect size drawn from over 100 studies. Results should be interpreted with caution. Overall, these mixed results do not offer strong support for Hypothesis 2. Certain study moderators, namely lab versus field samples, were investigated in order to provide a more meaningful comparison for FC measures.

Lab studies as a moderator of the effect size between honest and faked scores with a forced choice intervention. Very little research to date has studied faking on FC measures in an applied setting. Indeed, only one study in the current meta-analysis used a field setting or non-student sample to test faking on a FC measure. This is likely in part due to difficulties in implementing FC measures for interpersonal comparisons, a central purpose for personnel selection. However, this meant that comparing FC measures to the entire range of studies without an intervention provided an unrepresentative comparison group. In other words, the comparator for FC measures should be limited to lab studies, considering all but one study used a lab, student sample. This is especially impactful given known differences in the effect sizes (i.e., honest versus faked scores) between lab and field studies (see Table 2). Table 7 compares the effect sizes of scores across honest and faked conditions of FC versus no intervention (i.e., Likert scales) for lab studies only. These results provide a very different story than the results at the

Table 7.

Standardized Mean Difference of Lab Personality Scores Between Honest and Faked Conditions for Forced-Choice versus No Intervention in Lab Studies

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness – Lab											
	No Intervention	17,870	85	0.48	0.39	12%	0.40	0.56	-0.02	0.97	
	Forced Choice	2,639	18	0.08	0.56	8%	-0.18	0.34	-0.64	0.80	0%*
Conscientiousness – Lab											
	No Intervention	23,552	106	0.87	0.46	8%	0.78	0.96	0.27	1.46	
	Forced Choice	4,442	29	0.72	0.50	10%	0.53	0.90	0.08	1.36	88%
Emotional Stability – Lab											
	No Intervention	18,929	94	0.93	0.46	10%	0.84	1.02	0.34	1.52	
	Forced Choice	2,589	14	0.40	0.33	17%	0.22	0.57	-0.03	0.82	0%*
Extraversion – Lab											
	No Intervention	19,525	102	0.51	0.35	15%	0.44	0.58	0.07	0.95	
	Forced Choice	2,853	24	0.33	0.35	22%	0.19	0.47	-0.12	0.78	24%*
Openness – Lab											
	No Intervention	13,716	67	0.42	0.33	16%	0.35	0.50	0.01	0.84	
	Forced Choice	2,279	18	0.31	0.45	14%	0.10	0.51	-0.27	0.88	100%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

overall trait-level presented in Table 6. There is consistently less of a difference between honest and faked scores (i.e., smaller effect sizes) across the five traits in lab studies, and the magnitude of the differences is quite a bit larger, than those presented in Table 6. For instance, the effect size for emotional stability is much larger in lab studies (k = 94, d = 0.93) than all studies (k = 111, d = 0.79). Although the small to moderate effect size for emotional stability (k = 14, d = 0.40) from lab studies using a forced choice intervention is significantly smaller than the comparison in both cases (see Tables 6 and 7), the magnitude is much larger when the sample is limited to lab studies. As with the previous hypotheses, it is important to note the significant difference in k size for the effect sizes in question. Namely, results should be interpreted with caution when comparing effect sizes drawn from 94 versus 14 studies.

The 95% confidence interval on the effect size for agreeableness in FC studies (k = 18, d = 0.08, CI_{95} = -0.18 to .34) did not overlap with the effect size in studies without an intervention (k = 85, d = 0.48, CI_{95} = 0.40 to .56). These results tend to provide more support of Hypothesis 2, even though the effect sizes for conscientiousness and openness significantly overlapped across FC and Likert studies (*POL* = 88% and 118%, respectively). Indeed, the effect size for conscientiousness in FC measures (k = 29, d = 0.72) is surprisingly large relative to other FC effect sizes. It is critical to note, however, that this effect is smaller than the effect size for Likert-scales without any intervention (k = 106, d = 0.87).

Study design as a moderator of the effect size between honest and faked scores for lab studies with a forced choice intervention. Study design

(i.e., between vs. within subject) was analyzed as a potential moderator for FC interventions using lab/student samples. This was examined due to the large 80% credibility intervals around the *SWMD*s presented in Table 7, even despite the more focused approach to limiting only lab studies. It was also examined due to previously reported differences in faking across study design (e.g., Viswesvaran & Ones, 1999). The differences in estimates across study design in the current meta-analysis are smaller than those reported in previous meta-analyses, although part of that is dependent upon the presence of a faking intervention (see Tables 2 and 3). That is, when looking across all studies included in the current meta-analysis, the inclusion of studies that used an intervention (e.g., warnings, forced-choice) that were excluded from previous meta-analyses attenuated the differences across study design. Considering study design (i.e., between group and within group) can help provide a more nuanced understanding of where FC measures show the largest effects.

Table 8 shows the effect size for FFM scores across honest and faked conditions for forced choice and no intervention studies across study design. This analysis is limited to lab samples based on the limited field studies presented earlier. This moderator analysis revealed that *SWMD* estimates were consistently larger for studies using Likert measures (i.e., no intervention) than those with a forced choice measure in between-subject designs. Only openness had significant overlap between FC and Likert effect size estimates. The results were less consistent for within-subject designs. Compared to Likert/no-intervention studies, forced choice studies reported larger *SWMDs* for agreeableness (d = 0.39 and

Table 8

Study Design as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Forced-Choice versus No Intervention in Lab Studies

Trait	Intervention	Study Design	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
								Lower	Upper	Lower	Upper	
Agreeableness – Lab												
	No Intervention											
		Between	11,126	43	0.53	0.44	8%	0.4	0.66	-0.04	1.1	
		Within	6,744	42	0.39	0.24	31%	0.32	0.47	0.09	0.7	
	Forced-Choice											
		Between	1,851	6	-0.08	0.48	5%	-0.46	0.31	-0.69	0.54	0%*
		Within	788	12	0.45	0.55	17%	0.14	0.76	-0.25	1.16	100%
Conscientiousness – Lab												
	No Intervention											
		Between	14,924	55	0.92	0.43	8%	0.81	1.03	0.37	1.47	
		Within	8,628	51	0.77	0.50	9%	0.63	0.91	0.13	1.42	
	Forced-Choice											
		Between	2,776	11	0.67	0.42	9%	0.42	0.92	0.13	1.21	61%
		Within	1,666	18	0.80	0.60	12%	0.52	1.07	0.04	1.56	100%
Emot. Stab. – Lab												
	No Intervention											
		Between	14,405	59	0.94	0.44	9%	0.82	1.05	0.37	1.5	
		Within	4,524	35	0.91	0.51	12%	0.75	1.08	0.26	1.56	
	Forced-Choice											
		Between	1,965	7	0.44	0.25	19%	0.25	0.62	0.12	0.75	0%*
		Within	624	7	0.27	0.49	16%	-0.09	0.63	-0.36	0.89	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; *POL* =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for *POL* means that the confidence intervals did not significantly overlap.

Table 8

Study Design as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Forced-Choice versus No Intervention in Lab Studies (continued)

Trait	Intervention	Study Design	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		POL
								Lower	Upper	Lower	Upper	
Extraversion – Lab												
	No Intervention											
		Between	12,833	56	0.53	0.34	14%	0.44	0.62	0.1	0.96	
		Within	6,692	46	0.48	0.36	18%	0.37	0.58	0.01	0.94	
	Forced-Choice											
		Between	1,804	7	0.26	0.24	21%	0.08	0.44	-0.06	0.57	0%*
		Within	1,049	17	0.45	0.45	25%	0.24	0.66	-0.12	1.03	100%
Openness – Lab												
	No Intervention											
		Between	9,943	39	0.44	0.31	14%	0.34	0.53	0.04	0.83	
		Within	3,773	28	0.39	0.36	19%	0.26	0.52	-0.06	0.85	
	Forced-Choice											
		Between	1,598	7	0.29	0.27	20%	0.09	0.49	-0.05	0.63	100%
		Within	681	11	0.34	0.70	12%	-0.07	0.75	-0.56	1.23	100%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; *POL* =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for *POL* means that the confidence intervals did not significantly overlap.

0.45, respectively) and conscientiousness (d = 0.77 and 0.80, respectively). However, within-subject forced choice studies had a significantly smaller *SWMD* for emotional stability (k = 7, d = 0.27, CI_{95} = -0.09 to 0.63) than studies with Likert/no-intervention (k = 35, d = 0.91, CI_{95} = 0.75 to 1.08). Outside of the similar caveat on sample size differences, especially with regard to small (k < 10) available studies, these mixed findings suggest that the efficacy of forced choice measures to address faking may be largely dependent upon the study design.

Summary of forced choice hypothesis. Hypothesis II predicted that there would be less faking (i.e., smaller sample weighted mean d) for studies that used a forced-choice measure than those that used single-stimulus scale (e.g., Likert-style scale) without any other faking intervention. This hypothesis was supported, although the results were limited to lab studies due to the absence of field studies in the forced choice literature. Indeed, lab studies showed consistently significantly less faking for emotional stability and agreeableness on FC measures than Likert measures. The amount of faking is largely influenced by study design, as large differences were observed in between- vs. within-subject designs on lab studies using FC measures.

Hypothesis III – Item Transparency

Hypothesis III predicted that effect sizes representing the difference between scores in honest and faked conditions obtained from less transparent items will be smaller than effect sizes obtained from those that do not use any faking intervention. Table 9 shows mixed support at the trait-level across the different FFM traits. The effect size for conscientiousness with an item

transparency intervention (k = 22, d = 0.71) is actually larger than the effect size for studies without an intervention (k = 135, d = 0.62). However, the effect sizes for all other traits are smaller for studies that used an item transparency intervention than those that did not use any intervention. For example, the effect size for emotional stability with an item transparency intervention (k = 14, d = 0.61, CI_{95} = 0.35 to 0.87) was smaller than the effect size without an intervention (k = 111, d = 0.79, CI_{95} = 0.72 to 0.86), although the difference was not significant (*POL* = 86%). However, the confidence intervals overlapped significantly. This suggests that while the effect sizes fell in the expected direction, the differences between effect sizes with an item transparency intervention and no intervention were generally small.

Student/Lab sample as a moderator of the effect size between honest and faked scores with an item transparency intervention. Inspection of the study characteristics for studies that implemented an item transparency intervention yielded a similar finding to that of FC measures. Namely, all but two of the studies (k = 20) that used an item transparency intervention did so in a lab or student sample. Given the known disparities between naturally-occurring and instructionally-induced faking (see Table 2; Hooper, 2007), it is unsurprising that results at the trait-level provided largely mixed findings. The lab/student moderator was included in Table 9. Similar to Hypothesis II, effect sizes were larger (i.e., larger standardized mean difference of personality scores between honest and faked conditions) within the subset of item transparency studies that used lab studies than the overall item-transparency sample. Only two studies that

Table 9

Lab/Student Sample as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Item Transparency versus No Intervention

Trait	Intervention Type	Sample/ Setting	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
								Lower	Upper	Lower	Upper	
Conscientiousness												
	No Intervention		88,778	135	0.62	0.42	4%	0.55	0.69	0.08	1.15	
		Student/Lab	23,552	106	0.87	0.46	8%	0.78	0.96	0.27	1.46	
	Transparency		2,712	22	0.71	0.54	11%	0.49	0.94	0.02	1.41	86%
		Student/Lab	2,494	20	0.77	0.54	11%	0.53	1.00	0.08	1.45	100%
Agreeableness												
	No Intervention		64,452	103	0.40	0.40	4%	0.32	0.48	-0.11	0.91	
		Student/Lab	17,870	85	0.48	0.39	12%	0.40	0.56	-0.02	0.97	
	Transparency		2,579	19	0.34	0.47	12%	0.13	0.55	-0.26	0.94	100%
		Student/Lab	2,361	17	0.36	0.49	11%	0.13	0.59	-0.27	0.99	100%
Emotional Stability												
	No Intervention		38,511	111	0.79	0.39	7%	0.72	0.86	0.28	1.29	
		Student/Lab	18,929	94	0.93	0.46	10%	0.84	1.02	0.34	1.52	
	Transparency		1,902	14	0.61	0.50	11%	0.35	0.87	-0.03	1.24	100%
		Student/Lab	1,684	12	0.67	0.50	11%	0.38	0.95	0.03	1.31	63%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; *POL* =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for *POL* means that the confidence intervals did not significantly overlap.

Table 9

Lab/Student Sample as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Item Transparency versus No Intervention (continued)

Trait	Intervention Type	Sample/ Setting	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
								Lower	Upper	Lower	Upper	
Openness												
	No Intervention		26,921	81	0.29	0.39	7%	0.20	0.37	-0.21	0.79	
		Student/Lab	13,716	67	0.42	0.33	16%	0.35	0.50	0.01	0.84	
	Transparency		1,624	11	0.09	0.39	15%	-0.14	0.32	-0.41	0.59	100%
		Student/Lab	1,406	9	0.10	0.43	13%	-0.18	0.37	-0.45	0.64	19%*
Extraversion												
	No Intervention		62,167	122	0.42	0.34	7%	0.36	0.48	-0.02	0.85	
		Student/Lab	19,525	102	0.51	0.35	15%	0.44	0.58	0.07	0.95	
	Transparency		2,332	18	0.45	0.34	21%	0.29	0.61	0.01	0.89	100%
		Student/Lab	2,114	16	0.48	0.35	20%	0.31	0.65	0.03	0.93	100%

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

used item transparency interventions were done in the field and therefore not included in this analysis Although only two studies were excluded in this moderator, the effect sizes from the field studies were much smaller than those observed in lab studies. This is consistent with previous observations in this analysis.

Hypothesis III received more consistent support when examining only lab studies as the comparison group. Although the effect size of personality scores between honest and faked conditions remained large for conscientiousness in studies that used an item transparency intervention (k = 20, d = 0.77, CI_{95} = 0.53 to 1.00), it was smaller than the effect size for lab studies that did not use any intervention (k = 106, d = 0.87, CI_{95} = 0.54 to 0.69). The confidence intervals overlapped substantially for these effect sizes, so although the difference between observed effects is in the expected direction, it is not significant. Indeed, with the exception of openness, all of the confidence intervals around the effect size estimates significantly overlapped when comparing student/lab samples in studies without an intervention to those that used an item transparency intervention.

Type of item transparency as a moderator of the effect size between honest and faked scores. The item transparency intervention was unique in the sense that it combined multiple types of interventions. Randomization of items, the use of subtle items, and the use of contextualized items were combined because they all theoretically influenced faking by limiting item transparency and thus ability to fake (see Figure 3). The type of item transparency was examined as

Table 10

Type of Item Transparency Intervention as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Item Transparency versus No Intervention

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness											
	No Intervention	64,452	103	0.40	0.40	4%	0.32	0.48	-0.11	0.91	
	Subtle	663	5	0.31	0.00	100%	0.31	0.31	0.31	0.31	0%*
	Randomized	396	4	0.53	0.18	57%	0.36	0.71	0.31	0.76	100%
	Contextualized	1,061	8	0.57	0.41	16%	0.29	0.85	0.05	1.09	100%
Conscientiousness											
	No Intervention	88,778	135	0.62	0.42	4%	0.55	0.69	0.08	1.15	
	Subtle	636	7	0.35	0.01	100%	0.35	0.36	0.35	0.36	0%*
	Randomized	396	4	0.95	0.50	15%	0.45	1.44	0.30	1.59	100%
	Contextualized	1,221	9	0.71	0.58	9%	0.33	1.09	-0.03	1.45	100%
Emotional Stability											
	No Intervention	38,511	111	0.79	0.39	7%	0.72	0.86	0.28	1.29	
	Subtle	469	5	0.13	0.00	100%	0.13	0.13	0.13	0.13	0%*
	Randomized	396	4	0.85	0.50	15%	0.36	1.34	0.21	1.49	100%
	Contextualized	478	2	0.55	0.16	40%	0.32	0.77	0.34	0.75	37%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

Table 10

Type of Item Transparency Intervention as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions for Item Transparency versus No Intervention (continued)

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Extraversion											
	No Intervention	62,167	122	0.42	0.34	7%	0.36	0.48	-0.02	0.85	
	Subtle	416	4	0.17	0.00	100%	0.17	0.17	0.17	0.17	0%*
	Randomized	396	4	0.57	0.34	27%	0.24	0.91	0.14	1.01	100%
	Contextualized	1,061	8	0.55	0.41	16%	0.26	0.84	0.02	1.08	100%
Openness											
	No Intervention	26,921	81	0.29	0.39	7%	0.20	0.37	-0.21	0.79	
	Subtle	416	4	0.14	0.00	100%	0.14	0.14	0.14	0.14	0%*
	Randomized	396	4	0.38	0.00	100%	0.38	0.38	0.38	0.38	100%
	Contextualized	353	1	0.16	0.00	.	-	-	-	-	-

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

a potential moderator. Similar to the warning type moderator within Hypothesis I, this moderator was only applicable to the trait-level due to limited sample size.

Table 10 shows the sample weighted mean *d* of scores between honest and faked conditions across the different types of item transparency interventions. The most notable result from this table is that the effectiveness of the item transparency intervention varies greatly between subtle and randomized items. Subtle items consistently demonstrate smaller *SWMD*s (i.e., less score inflation across conditions) across the five factors, and even have non-overlapping confidence intervals for every trait. For example, studies that used subtle items to measure conscientiousness reported less score inflation ($k = 7$, $d = 0.35$, CI_{95} = 0.35 to .37) than those with no intervention ($k = 135$, $d = 0.62$, CI_{95} = 0.55 to .69). Non-overlapping confidence intervals for all of the FFM trait estimates indicate this is a useful intervention, but caution should be taken on these results due to limited *k* size for studies with an item transparency intervention.

The opposite is true for randomized items, where the *SWMD* (i.e., personality score across honest and faked conditions) is consistently larger for studies that reported randomizing the items compared to those that did not report any intervention for faking. This is contrary to Hypothesis III, as it suggests that test-takers are able to inflate their scores *more* when test items are randomized compared to test-takers that do not receive any faking intervention.

Finally, there were mixed results in terms of the magnitude of the *SWMD* of personality scores across honest and faked conditions for contextualized items compared to studies without any faking intervention. Conscientiousness,

agreeableness, and extraversion had higher *SWMD*'s for contextualized item conditions, while emotional stability and openness reported smaller *d*'s. Only the confidence interval for the effect size estimate emotional stability with contextualized items ($k = 2$, $d = 0.55$, $CI_{95} = 0.32$ to .77) did not overlap with the interval for the estimate without any intervention ($k = 111$, $d = 0.79$, $CI_{95} = 0.72$ to .86). However, the small *k* size ($k = 2$) limits generalizable conclusions from this result.

This moderator analysis highlights that these three interventions are not equivalent with regards to their ability to limit faking, even though they theoretically influence faking in the same way (i.e., limiting faking ability at the item level; Ellingson & McFarland, 2011). The hypothesis was partially supported – namely, subtle items seem to reduce mean differences between honest and faked conditions, but not randomized or contextualized items.

Summary of item transparency hypothesis. Hypothesis III predicted that the item transparency intervention would be effective at limiting response inflation. Support for this hypothesis was predominantly limited to subtle items, as randomizing items or using contextualized items (e.g., "at work" FOR) showed similar *SWMD* estimates to studies without any intervention. Indeed, the latter two interventions at times reported larger *SWMDs* than studies without an intervention, although not significantly so as the 95% confidence intervals had substantial overlap.

Hypotheses IVa, IVb, and IVc – Time Limit

The fourth set of hypotheses predicted that studies with a time limit would report larger effect sizes for the scores between honest and faked conditions (i.e., more faking) than (a) warnings, (b) FC measures, and (c) item transparency interventions. Table 11 shows the trait-level results of this hypothesis. First, it is important to note that only three ($k = 3$) studies were eligible for the time limit intervention in this meta-analysis. No additional moderators were analyzed, and the k size also limits the ability for any firm conclusions to be made for the time limit intervention. Analyses with $k < 5$ produce questionable meta-analytic results (*cf*. Arthur et al., 2001), although this is not a universally accepted cut-off. Indeed, as mentioned in the previous hypotheses, caution should be taken when interpreting effect sizes with $k > 5$, as there is still the potential for greater variability when based on a smaller study sample. However, even greater caution should be taken when k sizes are especially small. For this reason, the results for Hypotheses IVa-c cannot be fully tested in the present study. The following discussion is limited to high-level trends observed across the available studies.

Warnings consistently showed smaller *SWMD*s (i.e., smaller differences in mean scores between honest and faked conditions) than time limit strategies across the five traits. Indeed, the magnitude of the effect size for personality scores between honest and faked conditions was typically twice as large with time limit interventions as with warning interventions. For example, the effect size for conscientiousness with a warning ($k = 23$, $d = .59$) was a third of the size of the effect size for time limits ($k = 3$, $d = 1.97$). FC measures also tended to report

Table 11

Intervention Type (Warning, Forced Choice, Transparency, and Time Limits) as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval	
							Lower	Upper	Lower	Upper
Agreeableness										
	Warning	1,853	16	0.20	0.17	55%	0.11	0.28	-0.02	0.42
	Forced Choice	2,951	19	0.15	0.56	8%	-0.11	0.4	-0.57	0.86
	Transparency	2,579	19	0.34	0.47	12%	0.13	0.55	-0.26	0.94
	Time Limit	204	3	0.47	0	100%	0.47	0.47	0.47	0.47
Conscientiousness										
	Warning	3,330	23	0.59	0.37	18%	0.44	0.74	0.12	1.06
	Forced Choice	4,753	30	0.70	0.49	10%	0.52	0.87	0.07	1.32
	Transparency	2,712	22	0.71	0.54	11%	0.49	0.94	0.02	1.41
	Time Limit	204	3	1.97	0	100%	1.97	1.97	1.97	1.97
Emotional Stability										
	Warning	2,861	22	0.57	0.39	17%	0.41	0.73	0.07	1.07
	Forced Choice	2,589	14	0.40	0.33	17%	0.22	0.57	-0.03	0.82
	Transparency	1,902	14	0.60	0.50	11%	0.34	0.86	-0.03	1.24
	Time Limit	204	3	1.05	0.50	21%	0.48	1.62	0.41	1.69

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies.

Table 11

Intervention Type (Warning, Forced Choice, Transparency, and Time Limits) as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions (continued)

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval	
							Lower	Upper	Lower	Upper
Extraversion										
	Warning	2,602	21	0.22	0.22	40%	0.12	0.31	-0.06	0.5
	Forced Choice	3,165	25	0.29	0.35	21%	0.16	0.43	-0.15	0.74
	Transparency	2,332	18	0.45	0.34	21%	0.29	0.61	0.01	0.89
	Time Limit	204	3	0.59	0	100%	0.59	0.59	0.59	0.59
Openness										
	Warning	1,717	15	0.07	0.33	25%	-0.1	0.23	-0.35	0.49
	Forced Choice	2,591	19	0.31	0.42	15%	0.13	0.5	-0.22	0.85
	Transparency	1,624	11	0.09	0.39	15%	-0.14	0.32	-0.4	0.59
	Time Limit	204	3	0.31	0.14	76%	0.15	0.46	0.13	0.48

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies.

smaller effect sizes than time limit strategies. Openness reported the same effect size for FC (k = 19, d = .31) and time limit strategies (k = 3, d = .31), but the confidence intervals significantly overlapped. The smaller confidence interval for time limit studies is likely due in large part to the smaller k size from which to draw variability. Finally, item transparency interventions showed smaller effect sizes than time limits across all five traits.

Effect sizes of scores between honest and faked conditions for time limits versus no intervention. Although not hypothesized, the effect sizes for time limit strategies were also compared to those without any intervention. Across the five traits, the effect size for studies with a time limit was consistently larger than for studies without any intervention. For example, the effect size was three times larger for conscientiousness in studies with a time limit (k = 3, d = 1.97) than studies without any intervention (k = 135, d = 0.62). The three studies included in this analysis were collected on small, between-subject, student samples, with only around 30 participants in the honest and faked conditions. The participants completed either the Personality Research Form (Jackson, 1984) or the NEO-FFI (Costa & McCrae, 1992), with the former measuring conscientiousness through academic orientation. Despite the unexpectedly large effect sizes observed for this intervention, it is important to reiterate that the effect sizes for time limit intervention are not generalizable, and more research should be done before making solid conclusions regarding the efficacy of the intervention. That being said, initial results do not appear promising for this as a useful intervention to reduce faking.

Summary of time limit hypothesis. The fourth set of hypotheses could not be tested based on limited available data. However, trends from the data presented in Table 11 show that time limit strategies are not an effective way of reducing faking behavior, and may actually have the opposite effect for important traits such as conscientiousness.

Hypothesis V – Intention vs. Ability Interventions

Hypothesis V predicted that the *SWMD* for personality scores between honest and faked conditions would be smaller in studies with preventive interventions focused on intent than in studies with preventive interventions focused on ability. In other words, this hypothesis predicted that warnings would be most effective at reducing response inflation. Hypothesis I, II, and III showed that these interventions were generally useful in reducing faking compared to studies without an intervention, and Hypothesis IV showed that warnings tend to be different than the other interventions.

Table 12 shows the effect sizes for scores between honest and faked conditions with warnings (i.e., intent intervention) compared to the combined effect size for FC, item transparency, and time limit (i.e., ability interventions). Partially supporting Hypothesis V, most of the effect sizes were smaller for studies using a warning than studies using one of the other, ability-focused interventions. The confidence intervals around the effect size for conscientiousness and agreeableness significantly overlapped across intent and ability, while the overlap in confidence intervals was minimal for extraversion. Emotional stability, on the other hand, reported a larger *SWMD* for the scores

Table 12

Intervention Type as a Moderator of the Standardized Mean Difference of Personality Scores Between Honest and Faked Conditions

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		*POL*
							Lower	Upper	Lower	Upper	
Agreeableness											
	Ability (FC/Trans/Time)	5,734	41	0.24	0.52	10%	0.09	0.4	-0.42	0.91	
	Intent (Warning)	1,853	16	0.20	0.17	55%	0.11	0.28	-0.02	0.42	100%
Conscientiousness											
	Ability (FC/Trans/Time)	7,670	55	0.74	0.54	10%	0.59	0.88	0.05	1.43	
	Intent (Warning)	3,330	23	0.59	0.37	18%	0.44	0.74	0.12	1.06	100%
Emotional Stability											
	Ability (FC/Trans/Time)	4,695	31	0.51	0.44	12%	0.35	0.66	-0.06	1.07	
	Intent (Warning)	2,861	22	0.57	0.39	17%	0.41	0.73	0.07	1.07	100%
Extraversion											
	Ability (FC/Trans/Time)	5,701	46	0.37	0.35	21%	0.27	0.47	-0.08	0.81	
	Intent (Warning)	2,602	21	0.22	0.22	40%	0.12	0.31	-0.06	0.5	42%*
Openness											
	Ability (FC/Trans/Time)	5,701	46	0.37	0.35	21%	0.27	0.47	-0.08	0.81	
	Intent (Warning)	1,717	15	0.07	0.33	25%	-0.1	0.23	-0.35	0.49	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap; "FC/Trans/Time" = a combination of three faking interventions: Forced Choice/Transparency/Time Limit.

between honest and faked conditions for studies using an intervention focused on intent versus an intervention focused on ability. The effect size for intent (i.e., warning) interventions (k = 22, d = 0.57, CI_{95} = 0.41 to .73) was slightly larger than the effect size for ability interventions (k = 31, d = .51, CI_{95} = 0.35 to .66). Because of the inconsistent effect sizes for the FFM traits across ability and intent interventions, these findings offer only partial support for Hypothesis V.

Specific faking intervention as a moderator of the effect size between honest and faked scores. Hypothesis V compared intent and ability interventions. However, given the known differences between interventions observed in the previous hypotheses, all interventions should be compared separately in order to more clearly identify one particular intervention's effectiveness. It is also important to compare ability interventions to each other (notably FC and item transparency) given their separation in previous hypotheses.

This moderator analysis re-examines the data presented in Table 11. The sample weighted mean d's across the traits and provides partial support for Hypothesis V. Warnings show the smallest effect size (i.e., less faking) for conscientiousness, extraversion, and openness. The most traditionally job-relevant trait, conscientiousness, showed a moderate effect size in warned studies (d = 0.59), compared to large effect sizes for forced- choice (d = 0.70) and item transparency (d = 0.71) studies. The time limit study also showed a larger effect size for conscientiousness (d = 1.97), but the small sample of studies (k = 3) makes interpreting this effect size tenuous. However, FC measures appear to lessen faking for agreeableness and emotional stability. Table 11 shows that the

effect size for agreeableness (d = 0.15, 0.20) and emotional stability (d = 0.40, 0.57) are smaller in FC studies than studies with a warning, respectively.

Summary of Intent vs. Ability Hypothesis. Similar to the previous hypotheses, Hypothesis V received modest support. Although the trend was in the expected direction, the relationship was not consistent enough across all of the traits to make a definitive statement about the intervention's effectiveness. With the exception of emotional stability, interventions that limit intention to fake (i.e., warnings) tend to be more effective at limiting response inflation than interventions that limit ability to fake (i.e., forced choice, item transparency, time limits). Although not predicted in Hypothesis V, the analyses showed that (with the exception of openness) FC interventions are consistently more effective at producing less response distortion than item transparency interventions. This was unexpected because they both theoretically impact faking in similar ways, namely by limiting test-takers ability to fake the assessed construct.

Hypotheses VI to VIII – Remedial interventions

A meta-analysis of r's was used to test Hypotheses VI-VIII that focused on the criterion-related validity of the personality test scores in a motivated condition. Specifically, these hypotheses predicted that personality test scores would show stronger criterion-related validity with a range of outcomes (e.g., contextual performance, task performance) under preventive rather than remedial interventions. Due to limited available data, results were only interpreted when $k \geq 3$.

Table 13

Intervention Type as a Moderator of the Relationship Between Personality Scores and Performance Criteria

Trait	Intervention	*N*	*k*	Sample-weighted mean *r*	Corr. Mean *r* (*ρ*)	Corr. SD (SD*ρ*)	Variance due to Artifacts	95% Confidence Interval		80% Credibility Interval	
								Lower	Upper	Lower	Upper
Agreeableness											
	Correction	1,651	5	0.1	0.11	0.07	49%	0.05	0.15	-0.02	0.24
	Removal	17,172	3	0.18	0.2	0.03	26%	0.16	0.19	0.14	0.26
	Warning	648	3	0.05	0.06	0	100%	-0.03	0.13	0.05	0.05
	Forced Choice	471	2	-0.04	-0.05	0.1	44%	-0.13	0.05	-0.24	0.15
	Transparency	281	2	0.14	0.16	0	100%	0.03	0.26	0.16	0.16
Conscientiousness											
	Correction	4,080	9	0.18	0.2	0.11	22%	0.15	0.21	-0.01	0.41
	Removal	18,034	3	0.18	0.2	0.01	75%	0.17	0.2	0.18	0.22
	Warning	1,449	5	0.06	0.07	0.05	60%	0.01	0.11	-0.03	0.17
	Forced Choice	2,719	6	0.28	0.31	0.1	40%	0.25	0.31	0.12	0.5
	Transparency	641	4	0.27	0.3	0.05	75%	0.2	0.35	0.21	0.39
Emotional Stability											
	Correction	2,199	6	0.12	0.14	0.03	76%	0.08	0.17	0.07	0.2
	Removal	17,353	3	0.15	0.16	0.02	38%	0.13	0.16	0.12	0.2
	Warning	1,005	4	0.12	0.13	0.08	43%	0.05	0.18	-0.03	0.28
	Forced Choice	471	2	0.05	0.06	0	100%	-0.04	0.14	0.06	0.06

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; corr. Mean *r* (ρ) = corrected population correlation rho; corr. SD (SD ρ) = corrected standard deviation of the population correlation rho; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies

Table 13

Intervention Type as a Moderator of the Relationship Between Personality Scores and Performance Criteria (continued)

Trait	Intervention	*N*	*k*	Sample-weighted mean *r*	Corr. Mean *r* (ρ)	Corr. SD (SDρ)	Variance due to Artifacts	95% Confidence Interval		80% Credibility Interval	
								Lower	Upper	Lower	Upper
Extraversion											
	Correction	2,164	5	0.03	0.03	0.15	13%	-0.01	0.07	-0.26	0.32
	Removal	17,853	3	0.12	0.13	0.08	4%	0.10	0.13	-0.03	0.29
	Warning	1,005	4	0.06	0.07	0.00	100%	0.00	0.12	0.07	0.07
	Forced Choice	1,613	2	0.03	0.04	0.15	17%	-0.01	0.08	-0.26	0.33
	Transparency	281	2	0.11	0.13	0.08	57%	0.00	0.23	-0.03	0.28
Openness											
	Correction	983	3	0.08	0.09	0.08	41%	0.02	0.15	-0.07	0.26
	Removal	16,672	2	0.19	0.21	0.06	6%	0.18	0.20	0.08	0.33
	Warning	648	3	0.03	0.03	0.00	100%	-0.05	0.10	0.03	0.03
	Forced Choice	203	1	0.13	0.14	0.00	.	-	-	-	-

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; corr. Mean *r* (ρ) = corrected population correlation rho; corr. SD (SD ρ) = corrected standard deviation of the population correlation rho; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies

First, Hypothesis VI predicted that the criterion-related validity of test scores would be higher when using a warning compared to (a) correcting scores or (b) removing scores based on social desirability. The results presented in Table 13 fail to support either Hypothesis VIA or VIB. For each of the FFM traits, the criterion-related validity was lower in studies that used a warning versus either of the remedial interventions. For instance, the relationship between criteria and conscientiousness scores was smallest for warnings ($k = 5$, $\rho = .07$, 95% $CV_L = .05$) compared to both correcting for social desirability ($k = 9$, $\rho = .18$, 95% $CV_L = -.01$) and removing test-takers based on social desirability ($k = 3$, $\rho = .18$, 95% $CV_L = .18$). Most notably, the 95% confidence intervals did not overlap between warning and score correction conditions.

Next, Hypothesis VIIA and VIIB predicted that the criterion-related validity of test scores would be higher for measures using a forced choice method than those using a remedial faking intervention. The results provided mixed support for this hypothesis. As seen in Table 13, conscientiousness had a stronger relationship with the criteria in FC tests ($k = 6$, $\rho = .31$, 95% $CV_L = .12$) compared to both correcting for social desirability ($k = 9$, $\rho = .18$, 95% $CV_L = -.01$) and removing test-takers based on social desirability ($k = 3$, $\rho = .18$, 95% $CV_L = .18$). The combination of mixed support and limited data fail to provide sufficient evidence to support the pair of hypotheses.

Finally, Hypotheses VIIIA and VIIIB predicted that the criterion-related validity would be stronger for studies that used an item transparency intervention than those that used a remedial intervention (e.g., either score correction or case

Table 14.

Criterion-Related Validity Estimates in the Current Meta-Analysis Compared to Previous Personality Meta-Analyses

Trait	Barrick & Mount (1991) Job Performance	Ones et al. (1996) Correct for Social Desirability	Hurtz & Donovan (2000) Operational Validity	McCabe & Oswald (2013) GPA	Current Meta-Analysis				
					Score Correction (k =9)	Score Removal (k =3)	Warning (k =5)	Forced Choice (k =6)	Transp. (k =4)
Agreeableness	0.06	0.06	0.11	0.07	0.11	0.20	0.06	-0.05	0.16
Conscientiousness	0.23	0.23	0.20	0.26	0.18	0.18	0.06	0.28	0.27
Emotional Stability	0.07	0.07	0.13	0.00	0.12	0.14	0.12	0.05	-
Extraversion	0.10	0.10	0.09	-0.03	0.03	0.13	0.07	0.04	0.13
Openness	-0.03	-0.03	0.06	0.08	0.10	0.21	0.03	0.14	-

Note. All values reflect estimated criterion-related validity for test scores within the specified FFM trait; k = number of effect sizes included in the current meta-analyses

removal). In support of Hypotheses VIIIA and VIIIB, results based on limited data presented in Table 13 suggest that item transparency interventions may offer a useful avenue for improving the criterion-related validity of personality assessments. The relationship between the criteria and conscientiousness was stronger for item transparency studies ($k = 4$, $\rho = .30$, 95% $CV_L = .21$) than either corrections ($k = 9$, $\rho = .18$, 95% $CV_L = -.01$) or case removals ($k = 3$, $\rho = .18$, 95% $CV_L = .18$). A similar result was observed for agreeableness and openness for score corrections, but not case removal. In other words, the criterion-related validity was stronger for agreeableness and openness for case removal than item transparency interventions. No data were available for emotional stability or openness to contribute to the current meta-analysis. Because the relationship was enhanced under the most traditionally job-relevant trait (i.e., conscientiousness), the hypothesis was partially supported. However, a limited number of studies preclude any confident conclusions.

Additional analyses. The limited data for the hypotheses on criterion-related validity did not allow for any further moderator analyses. For example, many of the lab studies used a school success criterion (37%). The number of available studies examining different criteria within each intervention was consistently less than 5 and often less than 3. This did not allow for meaningful results to be presented for most FFM traits. However, it is important to frame the current findings around the extant literature on the criterion-related validity of personality test scores. An analysis of the criterion-related validity of the available literature on personality test scores was beyond the scope of this meta-analysis.

Instead, Table 14 presents the results of the current meta-analysis compared to previous meta-analyses on personality, namely Barrick and Mount (1991), Hurtz and Donovan (2000), Ones et al. (1996), and McCabe and Oswald (2013). These meta-analyses were chosen for a variety of reasons. The first two were chosen because they are the most frequently cited meta-analyses on the criterion-related validity of personality test scores. Finally, McCabe and Oswald (2013) was chosen because it is the most recent meta-analysis on the relationship between personality and GPA. Because a meaningful subset of the current data examined school performance, the comparison to McCabe and Oswald seemed prudent.

Similar to previous meta-analyses, the current study found that criterion-related validity tended to be highest for conscientiousness across interventions. However, warnings (ρ = .06), score corrections (ρ = .18), and score removal (ρ = .18) interventions reported lower criterion-related validity for conscientiousness scores compared to previous meta-analyses. Thus, there is no support that these interventions improve the criterion-related validity of personality traits.

Some of the differences in the observed validity coefficients in the current meta-analysis and those reported in previous analyses may be attributable to the mix between job and academic performance criteria. As seen in Table 14, traits such as emotional stability vary greatly across these criteria (Hurtz & Donovan, 2000; McCabe & Oswald, 2013). Due to the limited available data for the current analysis, no moderator analyses were possible within the criteria.

Summary of criterion-related validity hypotheses. In general, there is little support for Hypotheses VI. These hypotheses predicted that preventive

faking interventions would enhance the ability to make valid inferences from test scores in a motivated condition compared to remedial interventions. Indeed, the criterion-related validity observed in warning studies was actually smaller across many traits compared to the remedial interventions. It is important to also keep the larger context in mind; namely, how these estimates compare to the extant literature on the criterion-related validity of personality test scores. The estimates presented in the current analysis (Table 14) are inconsistent relative to previous meta-analyses. Some estimates are larger than previously reported estimates, while others are smaller. These inconsistencies are at least partially attributable to the need to collapse across criteria due to limited data (k = 34 across the 5 personality factors). Combined with significant overlap across moderators (e.g., criterion type and study design), the lack of data did not allow for further analysis within moderators.

Discussion

The findings presented in the Results section generally supported the hypotheses surrounding faking (i.e., Hypotheses I-V), but not the hypotheses on criterion-related validity (i.e., Hypotheses VI-VIII). In other words, faking (operationalized as the standardized mean difference between honest and faked conditions) tends to be reduced in the presence of various interventions, but there is limited evidence that interventions enhance the criterion-related validity of the assessment scores. The majority of the discussion will focus on response inflation rather than criterion-related validity because the preponderance of studies included in the meta-analysis measured the former.

The main findings can be summarized in three main points. First, faking still exists even if steps are taken to minimize it. This is evidenced by moderate to large effect sizes on FFM traits between honest and faked conditions even in the presence of a faking intervention. Although test-takers can still fake in the presence of an intervention, the second main finding is that the interventions are generally effective at *limiting* faking on FFM traits. However, the magnitude of this difference is not always practically or statistically significant. Finally, there is insufficient evidence that interventions improve the criterion-related validity of assessment scores taken in a motivated context. The limited available evidence, interventions do not appear to enhance the ability to make valid inferences from motivated test scores. These points are detailed below.

Main Findings

Faking still exists. One of the main conclusions from this study is that test-takers are able to distort their responses even when given one of the various faking interventions. While some of the faking interventions were useful at limiting faking, faked scores still tend to be moderately higher than honest scores. This is consistent with previous literature. Modest effect sizes were observed for conscientiousness in both warnings (d = .42; Mueller-Hanson et al., 2003) and item transparency (d = .57; Cucina et al., 2010) interventions. The findings around forced-Choice (FC) interventions were also consistent with the existing literature. For example, Christiansen et al. (2005) reported moderate effect sizes for conscientiousness across honest and faked conditions (d = .40) when collected via a FC measure. Although this was smaller than the effect size for a traditional,

Likert-style measure of conscientiousness (d = .68), test-takers were still able to inflate their scores to a moderate degree on the FC measure. The difference in faking between FC and Likert scales was even more muted in Heggestad et al. (2006), where the faking effect size for conscientiousness in FC measures (d = 1.20) was only nominally different than Likert measures (d = 1.23).

Given the extant empirical research on faking, it is unsurprising that test-takers were still able to increase their assessment scores even in the presence of a faking intervention. This is consistent with the key assumption of the faking literature, and one that researchers and practitioners alike will readily acknowledge: namely, that test-takers will fake their responses to the extent that they are motivated to do so, or believe that doing so will increase their odds in obtaining a valued outcome (Ellingson et al., 2011). Although presenting a warning against faking or using an alternative test method may curb the severity of the distortion, it is unlikely that the intervention will completely override the situational and motivational factors at play during the test-taking process. Researchers or practitioners looking for the "answer" to the faking problem will instead have to accept that score inflation is likely an unavoidable situation for self-report tests in motivated contexts. That does not mean, however, that nothing can be done about the issue. This study identified various interventions that, in the right contexts, were particularly useful for limiting response distortion.

It should be noted that a fair amount of the extant literature on intervention effectiveness compares faked scores in the presence versus absence of an intervention. For instance, McFarland (2003) and Robson et al. (2008) provided a

large amount of the empirical support for the effectiveness of warnings as a faking intervention, while Converse et al. (2008) was instrumental in building the theory around FC interventions. However, these primary studies were not included in the meta-analysis because they did not compare an honest to a faked condition. The finding that test-takers are still able to raise their FFM trait scores in a faked condition compared to an honest condition helps refine the focus of the faking literature. Specifically, scores should always be compared to an honest group (whether the design is between- or within-subjects) because it more directly answers the research question of "do test-takers fake." Only collecting faked responses and comparing across the presence or absence of an intervention oversimplifies the subject of faking on personality assessments by failing to take into account the movement relative to an "honest" or "baseline" score.

Intervention effectiveness – score inflation. There are two main approaches to determine the effectiveness of a faking intervention based on the analyses presented in this study. The first approach is to inspect the lower-bound 95% confidence interval. If the interval includes the point estimate in studies that used an intervention, this means that the intervention was generally effective in reducing response inflation. This was observed for some traits within many of the interventions, but the relationship was not consistent across the traits and interventions. For instance, forced choice interventions were generally effective at reducing response inflation for agreeableness. The lower-bound 95% confidence interval around the *SWMD* for agreeableness included zero, and also included zero within many of the study moderators such as lab studies and between-

subjects designs. Based on these findings, forced choice designs seem to be useful in inhibiting test-takers *ability* to fake their agreeableness score on personality tests (Ellingson & McFarland, 2011). While some confidence intervals included zero for other interventions, this finding was inconsistent across interventions and traits. Thus, generalizable conclusions of intervention efficacy for a given trait were difficult to make.

Identifying if the confidence interval includes zero is a useful indicator of an intervention's effectiveness. It helps researchers and practitioners identify if, in the presence of a given intervention, test-takers are able to meaningfully inflate their responses to a personality test. However, it is not surprising that few of the confidence intervals included zero. The previous section detailed the existing literature that has found consistently higher scores when test-takers are motivated to present themselves in a positive light. Further, this method of judging the faking intervention fails to take into account the relative inflation of the trait with *no* intervention present. Comparing conscientiousness and openness serves as a useful example in this case. Conscientiousness demonstrated much more response inflation ($k = 135$, $d = .62$) than openness ($k = 81$, $d = .29$) in studies without an intervention, and the lower-bound confidence intervals differed substantially (.545 and .202 for conscientiousness and openness, respectively). The findings related to Hypothesis I demonstrated general support for less faking with the presence of a warning than when no intervention was present. It therefore is unsurprising that more of the lower-bound 95% confidence intervals included zero for traits with less faking to begin with, such as openness (and to a lesser extent, agreeableness).

The current study employed a more practical approach to examining an intervention's usefulness. Specifically, the amount of overlap of confidence intervals was a primary determinant of whether an intervention was effective as compared to a no intervention condition (Cumming, 2009; Cumming & Finch, 2005). For example, as demonstrated in Hypothesis I, warnings against faking appeared effective for emotional stability even though the lower-bound confidence interval did not approach zero. Despite the moderate effect size (d = 0.57), it was significantly smaller (POL = 17%) relative to faking in studies without any intervention (d = 0.79). Although the effect sizes for interventions were not always significantly smaller to the appropriate comparison for no intervention (e.g., conscientiousness for warnings), the pattern was consistent enough to suggest that the interventions were generally effective at limiting distortion.

Criterion-related validity. Faking interventions are especially helpful for organizations that use top-down selection, as faking can influence the rank-order of test-takers (e.g., Rosse et al., 1998). However, faking interventions are perhaps most impactful if they alter or enhance the criterion-related validity of the assessment for predicting performance. Job performance ratings were the most frequently used criterion in the current study (k =13). This was most frequently operationalized by supervisor ratings, and was split between being used for administrative and research purposes. Another subset of the sample in the r dataset (k = 10) used academic performance, most frequently operationalized by GPA.

The extant literature is mixed on faking's impact on the criterion-related validity of personality assessments, and the current meta-analysis was unsuccessful in making firm conclusions. Although the interventions may be helpful in limiting response inflation, they do not appear to influence the ability to make valid inferences about future or concurrent performance from the scores on the assessment. In many ways, the lack of a significant finding is not surprising. Many researchers have found only modest differences in the criterion-related validity between honest and faked conditions, suggesting that faking itself may not impact the criterion-related validity of personality test scores (e.g., Hough, 1998; Ones et al., 1993). It is also important to point out that the extant literature on faking interventions and criterion-related validity is much less mature than the literature on response inflation. The limited number of studies did not allow for as much power or moderator analyses in the validity analyses as it did in the response inflation analyses. More empirical and theoretical research is needed before generalizable conclusions are made about the impact of a faking intervention on a test score's criterion-related validity. As mentioned in the previous section, Figure 3 represents a potential starting point for framing continued research in this area.

Practical Implications

Practitioners will likely focus on the modest difference in effect sizes between the presence versus absence of a faking intervention. If organizational leaders are truly concerned that faked scores are no longer representative of the true, underlying trait, the results of this meta-analysis do little to quell their

concerns. However, the analyses found that scores are consistently less inflated in the presence of a warning. Warnings are easy to implement: there are not many costs associated with using a warning, and they do not require sophisticated development strategies. It also appeared, on review of the literature, that some instructions included a warning about detection of faking even without expressly testing the effectiveness of a warning. This suggests that warnings may already be implemented in the instructions of some tests.

Although the cost of development is low, organizational leaders wishing to add a warning against faking into the instructions of their personality assessment have some important decisions to make. First, leaders need to be comfortable with providing a false warning. The extant literature suggests that there is little reason to believe that there are readily available methods for detecting faking behavior in a consistent, reliable manner (*cf*. Griffith et al., 2006). Providing a warning without a way of verifying a "faked" response may come across disingenuous to applicants. Although extant literature has not shown a large difference in applicant reactions as the result of a warning (e.g., Converse et al., 2008; Mitchell & Adair, 2014), organizational leaders should consider if this is consistent with their values.

Another consideration before implementing warnings is the *type* of warning provided. Many warnings include a statement of consequence (e.g., "if you fake, you will be removed from the applicant pool") while others are more positively framed (e.g., "faking inhibits our ability to make valid assessments"). The differences in warning type are reviewed by Pace and Borman (2006), and

despite a dearth of primary research directly comparing warning type, the current meta-analysis was able to compare the amount of faking under different types of warnings. Table 5 shows that the effect sizes were consistently smaller in the presence of a warning compared to no warning, although differences between warning types were inconsistent. Consequential warnings showed less response inflation on conscientiousness and emotional stability, the two traits more traditionally related to job performance (Hurtz & Donovan, 2000). However, educational warnings had some significantly smaller effect sizes compared to other warnings. Interestingly, one primary study found that *appeal to moral reason* warnings were most effective at reducing faked responses to conscientiousness scores on the Big Five Inventory (BFI; Mitchell & Adair, 2014). This warning type was not included in the current analysis, as there were insufficient primary studies to include it in a meta-analysis. More research using real applicant data is needed prior to making definitive judgments about the best "type" of warning.

Interventions focused on limiting faking *ability* also limited response inflation. Forced-Choice (FC) measures reported less faking on most traits than Likert measures without any other intervention. These studies were generally limited to lab/student samples, but the findings were generally consistent. Item transparency interventions were most effective when using subtle items (see Table 10). Practitioners looking to curb faking should consider investing resources in developing such measures that go beyond standard, single-stimulus personality measures.

However, FC measures or subtle items are difficult to implement in practice. They require a fair amount of effort to develop as well as validate. For FC measures in particular, efforts must be taken to address ipsativity (*cf.* Hicks, 1970). These interventions may not be as attractive to organizations wishing to test personality, especially considering that the difference in faking effect size is generally small and actually in the *opposite* direction for conscientiousness. Organizations may also incur similar costs if wishing to develop a test with more subtle items. This typically requires a larger item pool, and test-taker reactions tend to be lower for longer assessments. All the same, great strides have been made in implementing FC measures in part due to their less "fakeable" nature. For instance, the Tailored Adaptive Personality Assessment System (TAPAS) used by the United States Army uses item-response theory (IRT) to construct and score its multidimensional forced choice measure (Drasgow, Stark, Chernyshenko, Dye, Hulin, & White, 2012). This test has been used widely to select and promote entry-level Army cadets, and results suggest that there is little difference between scores in honest and motivated conditions.

Implications for Research

One of the core implications from this research is that, although faking still exists in the presence of faking interventions, the interventions *limited* faking in most cases. This suggests that the interventions are generally working, and that faking can be decreased by further calibrating the interventions. For instance: does the type of warning depend on the context of the job? Is it possible to develop good, subtle conscientiousness items, considering that test-takers were

able to fake subtle items more than transparent items on this trait? This meta-analysis is crucial for faking researchers because it provides the first comparison of various faking interventions. Although there is a significant base of primary research, this study offers an essential starting point for consolidating the literature and guiding future faking research.

Another implication for faking research in general, and research on faking interventions specifically, is the importance of sample and study design. Contrary to Hooper (2007), the current meta-analysis observed consistently larger effect sizes in lab than field studies. This implies that research done on lab studies may not be generalizable to the field, and highlights that lab studies likely represent the upper limit of faking (Viswesvaran & Ones, 1999). It is also important to highlight the incredibly modest number of field studies compared to lab studies, as the large differences observed between these samples may be due to the quantity of empirical research. This was most noticeable in FC and item transparency interventions, where almost all studies were performed on student samples in the lab.

Study design (i.e., between- vs. within-subject) also highlighted important trends in the faking literature. Most studies used between-subject designs, despite arguments in the literature that a within-subjects design is more appropriate (e.g., Griffith & McDaniel, 2006). Between-subject designs are easier to implement, and indeed are frequently a requirement for field research. In other words, field studies often compare applicants to incumbents to identify faking. One way to incorporate within-subjects designs in field settings is to compare applicants at

Time 1 to the same employees after they are incumbents at Time 2. One potential problem with this design is that it may result in range restriction, as only "favorable" respondents to the assessment at Time 1 will be available for comparison in Time 2.

Researchers that are limited to student samples should strive to employ within-subject designs whenever possible. Recent literature in warnings is a good example of how researchers can employ these designs. For instance, Fan et al. (2012) and Ellingson et al. (2011) both used an honest group and then offered a warning to half of the group, with the other half continuing to the faked condition without a warning. This provides a more robust test of the warning effectiveness than comparing independent groups.

As mentioned earlier, a more robust test of the TPB-based model in Figure 3 is needed in order to understand exactly how the interventions function. For example, the model suggests that warnings influence faking behavior primarily through their impact on *intent* to fake. Some empirical research supports this claim. Mitchell and Adair (2014) found that the presence of a warning influenced pre- to post-intentions to fake, but did not influence perceived *ability* to fake. More empirical research is needed to test other parts of the model, especially the *ability* to fake interventions. This is especially important considering the large differences observed within this group. For instance, randomizing items actually increased faking behavior relative to no intervention, while subtle items appeared to consistently reduce faking. Testing the underlying mechanisms of these interventions is a crucial step to extending the research.

One final result may be interesting for researchers, and deserves further inquiry. Effect sizes across many of the interventions closely approximated the effect size of that trait in the field without any intervention. For example, the *SWMD* for emotional stability scores across honest and faked conditions with a warning ($k = 22$, $d = 0.57$) was quite similar to the effect size for field studies without an intervention ($k = 17$, $d = 0.65$). This was similar across several other interventions and traits. Researchers should investigate if an intervention (in this case, a warning) helps approximate the applicant context by increasing test-taker accountability to a similar level as in true applicant scenarios. In other words, faking in lab studies is often regarded as the "upper limit" of faking (Viswesvaran & Ones, 1999). If implementing a warning results in faking estimates that are more consistent with "real-life" faking, this may help further researchers' understanding of applicant faking without the costs and difficulties often associated with conducting research in the field.

Limitations

Faking in the field. One of the first limitations of the current study was that the final dataset was disproportionately lab-based, especially within the studies that used a faking intervention. Only around 10% of the total studies were measured in field samples, with only seven out of 75 (9.3%) of studies in warning, FC, and item transparency studies. This severely limits the generalizability of these findings to a field setting because of the large differences in faking observed in the field versus lab. Effect sizes from field studies were consistently smaller than lab studies measuring the same trait at the same level of

a given moderator. For instance, studies without an intervention reported larger effect sizes in the lab than the field (d = 0.87 and 0.52, respectively), and a similar trend was observed for warning studies in the lab and field (d = 0.62 and 0.43, respectively). Even with a sufficient k size, the current meta-analysis is more reflective of faking in the lab than in the field. Applying these findings to the field should be done cautiously. The largely lab-based sample in the current meta-analysis reinforces the call for increased field studies in faking research (e.g., Dilchert & Ones, 2011). The representation was so small in FC and item transparency studies in the d dataset that the 3 field settings in these cases were ostensibly removed from the analyses to help create a more representative comparison to studies without an intervention.

FFM measurement: Facets and post hoc mapping. One potential limitation with the current meta-analysis was that some studies were included and mapped to one of the FFM traits even though they were developed and validated outside of that model. Personality is not limited to the five traits measured in this paper, and meaningful insights can be gathered by traits outside of the FFM. Indeed, effect sizes were at times quite different between FFM traits measured from an *a priori* measure compared to *post hoc* measures. The mapping process was well documented and supported by empirical testing, but that does not change some of the differences observed. Despite being mapped to a FFM trait, it is possible that some of the *post hoc* scales share minimal construct-relevant variance with the *a priori* scale measuring the same trait. This may be most evident for measures of aggression, depression, and irritability that were mapped

to the FFM trait of emotional stability. While one part of emotional stability (the opposite of neuroticism) includes a lack of control over emotions and behavior that is similar to aggression, an argument could be made that these are distinct constructs. Table 15 displays the effect sizes for emotional stability across the *a priori* and *post hoc* measures, and inspection of the POL shows that the effect size estimates within each of the interventions do not overlap between the two types of measurement. This finding suggests that *a priori* and *post hoc* measures may be measuring different constructs, or there is some systematic difference in the way these measures are applied that tends to make *post hoc* measures less fakable.

Another potential issue of the observed effect sizes within the FFM is that it ignores a growing literature on the importance of facet traits (e.g., Fisher, Bell, Dierdorff, & Belohlav, 2012). Test-takers applying for a sales job may be more prone to fake on a measure of assertiveness than a measure of excitement seeking. The decision to include effect sizes at the "global" trait level was necessary in order to achieve necessary *k* size. Additionally, very few studies reported effect

Table 15

A-priori/Post-Hoc Measurement as a Moderator of the Standardized Mean Difference of Emotional Stability Scores Between Honest and Faked Conditions Across Types of Faking Interventions

Intervention	A-priori/ Post-Hoc	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		POL
							Lower	Upper	Lower	Upper	
No Intervention											
	A-priori FFM	24,101	64	0.90	0.37	8%	0.81	0.99	0.43	1.37	
	Post hoc FFM	14,719	49	0.60	0.36	10%	0.50	0.70	0.13	1.06	0%*
Warning											
	A-priori FFM	1,121	9	0.31	0.27	30%	0.14	0.49	-0.04	0.66	
	Post hoc FFM	1,740	13	0.74	0.37	19%	0.54	0.94	0.27	1.21	0%*
Forced-Choice											
	A-priori FFM	801	2	0.56	0.00	100%	0.56	0.56	0.56	0.56	
	Post hoc FFM	1,682	11	0.30	0.37	16%	0.08	0.52	-0.18	0.77	0%*
Item Transparency											
	A-priori FFM	1,495	9	0.73	0.50	9%	0.40	1.05	0.09	1.37	
	Post hoc FFM	407	5	0.15	0.00	100%	0.15	0.15	0.15	0.15	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

sizes on applicable facet traits such as achievement motivation or dutifulness. The focus of the current analysis was not on the appropriateness of FFM measurement, so this was not included as a hypothesis or focus of analysis. However, future primary studies may wish to incorporate the aforementioned evidence regarding "job-relevant" faking along with the dearth of literature on facet traits to investigate faking within FFM facets compared to global traits.

Future Directions

The results of this meta-analysis revealed fruitful areas for future research. First, the results can guide research on faking interventions by highlighting which interventions are most effective, and the conditions (e.g., study design, sample) in which this effect is maximized. As no research to date has compared faking interventions, this research was sorely needed. Second, results can also guide theory development on faking and the integration of faking interventions with TPB. Finally, new questions emerged based on a deeper dive into the results of this meta-analysis. For example, future research can examine the job-relevance of some interventions as well as the impact of faking interventions on the psychometric properties of the personality assessment. More detail on these potential avenues for future research is described below

Guiding future research on faking theory. The results of the current study contribute to the extant literature on faking by illuminating areas for research within faking theory. For instance, an intervention's effectiveness should be further examined within the framework of Theory of Planned Behavior (TPB). This theory was used as the basis for understanding how the interventions

influenced faking behavior (see Figure 3). Support for Hypothesis I and V indicated that warnings tended to be more effective at limiting faking than other interventions, and warnings also influence faking behavior the earliest according to theory by influencing faking intentions. These results suggest that the "earlier" the intervention, the better for reducing score inflation. However, the opposite appeared to be the case with the criterion-related validity results. Indeed, warnings appeared to do little to enhance (and, in fact, lessened) the ability to make valid inferences from test scores. More robust, empirical tests of this model are needed in order to make stronger claims about how faking interventions are best represented within TPB, but the current study offers useful guidance for future theoretical work.

Curvilinear faking and the job-relevance of faking interventions. The current study found that faking interventions decreased the amount of faking when operationalized as the standardized mean difference between honest and motivated conditions. Some may argue that this operationalization oversimplifies faking, as it relies on two strong assumptions. First, it assumes that faking should be considered as linear, such that higher scores are always better. The relationship between personality and various criteria is likely curvilinear (*cf.*, Converse & Oswald, 2014), so it stands to reason that faking should be considered curvilinear as well. Indeed, research suggests that item desirability ratings are not linear across response options and tend to vary as a function of occupational context (Dunlop, Telford, & Morrison, 2012). Further answers to this question of non-linearity are best provided by primary studies, rather than meta-analyses, in order

to capture the nuances necessary to report a curvilinear relationship. The current study, along with several other meta-analyses (e.g., Hooper, 2007; Birkeland et al., 2007), have clearly established that mean scores in motivated conditions are higher than in honest conditions. While the current study shows that faking interventions are useful for reducing faking in this operationalization, the next step for intervention research is to examine how the interventions influence faking or item desirability in non-linear ways.

The second main assumption in this operationalization is that faking operates independent of important contextual cues. Although the results and previous discussion focus more on conscientiousness than openness, the question of job-relevant faking is raised. In other words, how effective are interventions at reducing job-relevant faking? The above analysis suggests that test-takers are able to fake more traditionally job-relevant traits (e.g., conscientiousness) than less job-relevant traits (e.g., openness). It therefore stands to reason that a faking intervention may also be influenced by the larger context. For instance, faking may differ when the personality assessment is used as the only predictor or based on multiple factors because of the perceived weight given to personality test scores (Ziegler et al., 2011). In a similar vein, applicants may be more inclined to fake on a "pre-test" that is given as a first hurdle in a selection system, as they are unaware of what (if any) future steps await in the selection process (Reeder & Ryan, 2011).

The type of job that the applicant is seeking represents another important contextual factor. Although conscientiousness is commonly regarded as the most

effective trait at predicting job performance (e.g., Hurtz & Donovan, 2000), there are certainly job positions for which other traits are also important. For instance, extraversion is often considered an important trait in sales positions (Barrick & Mount, 1991), although recent research again suggests that the relationship is curvilinear (Grant, 2013). All the same, warnings may not be as effective at deterring applicants for sales jobs that potentially reward cut-throat behavior and Machiavellianism as they are in other, more "white-color" jobs. Because the current study offers useful information about faking interventions across contexts, the next step for future research is to examine the interventions within specific jobs to identify if the effectiveness is dependent upon contextual factors of the job.

In order to provide some context for this future research, the current *d* dataset was examined to assess whether intervention effectiveness varied across job position. The job position (including both naturally-occurring and instructionally-induced faking) provided by primary studies varied significantly, so they were collapsed into a meaningful sub-set that provided sufficient *k* sizes to provide additional analyses. Table 16 shows the primary study context and the mapping used for this analysis. Customer service was the most consistently used context (*k* =27, 13%) across all of the included studies, but none of the warning studies were collected in this context.

The job position context was examined as a supplementary analysis to provide a preliminary test of whether faking interventions differ by job position. Table 17 illustrates some interesting findings, albeit on limited data. Faking is

Table 16

Examples of Specific Job Positions Within Collapsed Position Areas

Collapsed Position	Specific job position information (provided to test taker)
College or University	Scholarship selection, student organization, university applicant
Customer Service	Customer Service Representative
Government	"Sensitive government position"
Librarian	Librarian
Managerial	Entry-level manager, store manager
Misc. Blue Collar	Gardener, manufacturing
Misc. White Collar	Bank Teller, Administrative Assistant, Accountant
Nursing	Nurse
Police or Military	Patrolman, Army, Border Patrol Agent
Sales	Salespeople
Teacher	Teacher, Professor

Note: Misc = Miscellaneous.

Table 17

Applicant Job Position as a Moderator of the Standardized Mean Difference of Extraversion Scores Between Honest and Faked Condition for Different Faking Interventions

Trait	Intervention Type	*N*	*k*	Sample Weighted Mean *d*	Corrected SDδ	Variance due to Sampling Error	95% Confidence Interval		80% Credibility Interval		POL
							Lower	Upper	Lower	Upper	
Extraversion - Sales											
	No Intervention	1,055	8	0.64	0.33	23%	0.41	0.86	0.22	1.05	
	Warning	367	4	0.55	0.27	38%	0.28	0.82	0.20	0.90	100%
	Forced Choice	412	2	0.48	0.00	100%	0.48	0.48	0.48	0.48	65%
	Transparency	412	2	0.48	0.00	100%	0.48	0.48	0.48	0.48	65%
Extraversion - Misc. White Collar											
	No Intervention	3,309	7	0.56	0.18	21%	0.42	0.69	0.33	0.79	
	Warning	431	3	0.05	0.00	100%	0.05	0.05	0.05	0.05	0%*

Note. N = total sample size across all effect sizes; k = number of effect sizes included in the meta-analyses; SDδ = sample size weighted standard deviation of the mean effect size; confidence intervals = the variability around the mean effect size; credibility intervals = the variability of the mean effect size in the population of studies; POL =proportion overlap of 95% confidence intervals, where the referent is the effect size for "no intervention" in that trait. * for POL means that the confidence intervals did not significantly overlap.

relatively prevalent among sales jobs, with a sample-weighted mean *d* of 0.64 without any intervention. The effect size decreases across each of the interventions, but not nearly as much as it does in other jobs where extraversion may be less critical for job performance. These "miscellaneous white collar" jobs include accounting, administrative assistant, and human resources positions. Applicants for these positions may feel less pressured to appear extraverted on selection tests, so a warning against faking may be more effective at limiting response inflation. More research should investigate the extent to which faking interventions vary as a function of job type or the relevance of a given trait for performance.

Faking and other psychometric properties. The analyses presented in this meta-analysis focused primarily on mean scores and the criterion-related validity of these scores. However, faking can affect other test properties. In particular, the internal consistency and construct validity can be meaningfully altered by faking.

Reliability. Test scores are more reliable to the extent that the measure is internally consistent. No research to date has expressly examined the influence of faking on internal consistency, although many authors suggest that faked scores have greater variance (e.g., Heggestead, 2011). To provide a baseline for future research, the reliability coefficients for the obtained scores were compared across honest and faked conditions. Table 18 shows the sample-weighted reliability across the available studies in honest and faked conditions. This analysis suggests that faked scores tended to be more reliable than honest scores, although the

Table 18.

Reliability Across Honest and Faked Conditions among Types of Faking Interventions

Intervention	Trait	Alpha (Honest)	Alpha (Faked)
No Intervention			
	Agreeableness	0.76	0.71
	Conscientiousness	0.74	0.82
	Emotional Stability	0.77	0.76
	Extraversion	0.78	0.77
	Openness	0.74	0.73
Warning			
	Agreeableness	0.78	0.79
	Conscientiousness	0.82	0.83
	Emotional Stability	0.80	0.84
	Extraversion	0.82	0.79
	Openness	0.73	0.75
Item Transparency			
	Agreeableness	0.71	0.73
	Conscientiousness	0.83	0.86
	Emotional Stability	0.73	0.64
	Extraversion	0.76	0.74
	Openness	0.71	0.60
Forced Choice			
	Agreeableness	0.75	0.64
	Conscientiousness	0.76	0.74
	Emotional Stability	0.74	0.68
	Extraversion	0.81	0.77
	Openness	0.75	0.72

Note: All estimates reflect the internal consistency (alpha).

opposite is true for FC measures. The largest shifts are observed for conscientiousness. A large shift in the reliability between honest and faked conditions is consistent with the literature on the Semantic Exercise Model of cognitive processing (e.g., Hsu et al., 1989; Martin, 2011). This model suggests that faked responses reflect a less complex schema than a self-referenced schema and should therefore take less time because it is easier to reference an ideal or faked schema than it is to make a complete self-evaluation. Along those lines, it seems that the faked schema results in a more consistent measure. Table 17 further shows that the reliability tends to be higher in warning studies compared to other interventions. This is interesting in light of the reduced criterion-related validity for scores with this intervention, and further highlights that more reliable scores do not necessarily make the test more valid. While these findings serve as a useful starting point for future research, more primary, empirical research is needed to test the nuanced effects of faking interventions on various test properties. Primary research can also address some of the limitations of the current meta-analysis by examining test properties at the facet trait level.

Construct Validity. Faking can influence the factor structure underlying the personality scale, which in turn can result in a personality assessment that no longer measures job-relevant traits. Some studies suggest that traditional 5-factor personality inventories offer the best fit for non-applicants (e.g., incumbents), but not for job applicants. Specifically, a sixth-factor termed the "ideal employee" factor, may maximize the fit for job applicants (e.g., Cellar, Miller, Doverspike, & Klawsky, 1996; Schmit & Ryan, 1993). Confirmatory factor analyses by Zickar

and Robie (1999) found that applicants and non-applicants differed in the number of latent factors, error variance, and factor correlations. Item response theory (IRT) also shows that items function differently across an independent sample of applicants and non-applicants (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). However, not all research shows that faking affects construct validity. Smith, Hanges, and Dickson (2001) found that the scale structure remained relatively stable across applicant, incumbent, and student samples, suggesting that faking does not have a strong influence on construct validity coefficients.

The literature on construct validity and faking is relatively scarce and the findings are somewhat inconsistent. More research is needed on the influence of faking on construct validity as well as the ability of various faking interventions to maintain or improve the factor structures of personality measures.

Conclusion

The current meta-analysis judged the efficacy of various faking interventions with the goal of furthering both practice and research on the use of personality for employee selection. The limited data on faking and criterion-related validity suggests that more research is needed before solid conclusions can be reached on the impact of faking on making inferences for selection. Additionally, the limited research in field settings suggests that more research on actual job applicants is needed before results can be generalized to real-world faking. This caveat is even stronger considering observed differences between lab and field studies when the data were available. Despite some of these limitations, the extensive literature on mean difference scores in faking allows for specific

recommendations. The below 5 recommendations offer a summary of the analyses.

1. Warnings are as effective as other methods for limiting response inflation.
2. However, warnings appear to limit the criterion-related validity, so organizationally-specific validation tests should be done to ensure the warning does not alter predictive validity and other important test properties.
3. Randomizing items does not reduce faking behavior and therefore has limited utility as a faking intervention.
4. To limit faking ability, both a) FC measures and b) using subtle items are effective interventions.
5. Given the null effects on criterion-related validity and the potential for removing or altering valid test scores, correcting or removing responses due to a test-taker's score on a social desirability measure is not recommended.

Test-takers are generally able to inflate their scores from honest to faked conditions, regardless of the presence of a faking intervention. However, the current meta-analysis succeeded in demonstrating that faking interventions can *reduce* score inflation on personality tests. This is most important for organizations using top-down selection methods on personality assessments, as any influence to the rank order of applicants can alter a selection decision. All the same, top-down selection systems still have the potential for selecting a mix of faked and honest responses to the extent that an intervention reduces all faked

scores to a similar degree. More research is needed to test if interventions function differently for “fakers” versus “honest responders.” Additional research is also needed on non-student samples and the effect of faking interventions on the ability to make valid inferences on organizational criteria from personality scores. However, the current analysis is instrumental in addressing what can be done to address the problem of faking.

References

* Reference included in meta-analysis

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckman (Eds.), *Action-control: From cognition to behavior*, (pp. 11-39). Heidelberg, Germany: Springer.

Ajzen, I. (1991). Theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

* Alliger, G.M., Lilienfeld, S.O., & Mitchell, K.E. (1995). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32-39.

* Anguiano-Carasco, C., Vigil-Colet, A., & Ferrando, P.J. (2013). Controlling social desirability may attenuate faking effects: A study with aggression measures. *Psicothema*, *25*, 164-170.

Arthur, W. Jr., Bennett, W., & Huffcuttt, A.I. (2001). *Conducting a meta-analysis using SAS*. Mahwah, NJ: Lawrence Erlbaum Associations.

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet–based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, 18, 1-21.

Bäckström, M.B., Björklund, F., & Larsson, M.R. (2011). Social desirability in personality test: Outline of a model to explain individual differences. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality test* (pp. 201-213). New York, NY: Oxford University Press.

* Bagby, R.M., & Marshall, M.B. (2003). Positive impression management and its influence on the revised NEO personality inventory: A comparison of analog and differential prevalence group designs. *Psychological Assessment, 15*, 333-339.

* Bagby, R.M., Rogers, R., Nicholson, R.A., Buis, T., Seeman, M.V., & Rector, N.A. (1997). Effectiveness of the MMPI-2 validity indicators in the detection of defensive responding in clinical and nonclinical settings. *Psychological Assessment, 9*, 406-413.

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

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.

Beck, L., & Ajzen, I. (1991). Predicting dishonest actions using the theory or planned behavior. *Journal of Research in Personality, 25*, 285-301.

* Bernal, D.S. (1999). *The hybrid scaling technique: Faking out the fakers with a new method of scale construction* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 9925138).

* Bing, M.N., Whanger, J.C., Davison, H.K., & VanHook, J.B. (2004). Incremental validity of frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89*, 150-157.

* Bing, M.N., Kluemper, D., Davison, H.K., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes, 116*, 148-162.

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.

Boldero, J. (1995). The prediction of household recycling of newspapers: The role of attitudes, intentions, and situational factors. *Journal of Applied Social Psychology, 25*, 440-462.

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6*, 199-214.

* Bott, J.P., O'Connell, M.S., Ramakrishnan, M., & Doverspike, D. (2007). Practical limitations in making decisions regarding the distribution of applicant personality tests scores based on incumbent data. *Journal of Business Psychology, 22*, 123-134.

* 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.

* Boyce, A. (2005). *An investigation of faking: Its antecedents and impacts in applicant settings* (Master's Thesis). Retrieved from ProQuest Dissertations and Theses Database. (UMI 1428909).

* Braun, J.R. (1963a). Effects of positive and negative faking sets on the survey of interpersonal values. *Psychological Reports*, *13*, 171-173.

* Braun, J.R. (1963b). Fakability of the Gordon Personality Inventory: Replication and extension. *The Journal of Psychology*, *55*, 441-444.

* Braun, J.R. (1965). Effects of specific instructions to fake on Gordon Personal Profile scores. *Psychological Reports, 17*, 847-850.

* Braun, J.R., & Costantini, A. (1968, April). *Fakability and control for social desirability of the Survey of Personal Values*. Paper Presented at the 76th meeting of the American Psychological Association, San Francisco, CA.

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

* Braun, J.R., & Farrell, R.M. (1974). Re-examination of the fakability of the Gordon Personal Inventory and profile: A reply to Schwab. *Psychological Reports, 34*, 247-250.

* Braun, J.R., & Gomez, B.J. (1966). Effects of faking instructions on the Eysenck Personality Inventory. *Psychological Reports, 19*, 388-390.

* Braun, J.R., & Iervolino, A. (1972). Faking and faking detection on the Comrey Personality Scales. *Psychological Reports, 30*, 636.

* Braun, J.R., & La Faro, B. (1968a). Effects of salesman faking instructions on the Contact Personality Factor test. *Psychological Reports, 22*, 1245-1248.

* Braun, J.R., & La Faro, B. (1968b). Fakability of the Sixteen Personality Factor Questionnaire, form C. *The Journal of Psychology, 68*, 3-7.

* Braun, J.R., & La Faro, B. (1969a). A further study of the fakability of the Personal Orientation Inventory. *Journal of Clinical Psychology, 18,* 101-105.

* Braun, J.R., & La Faro, B. (1969b). Faking and faking detection on the 16 PF-Form A. *The Journal of Psychology, 71*, 155-158.

* Braun, J.R., & Smith, M. (1973). Fakability of the Self-Perception Inventory: Further investigation. *Psychological Reports, 32*, 586.

* Braun, J.R., & Tinley, J.J. (1972). Fakability of the Edwards Personality Inventory booklets IA, II, and III. *Journal of Clinical Psychology, 28,* 375-377.

* Brown, R.D., & Harvey, R.J. (2003, April). *Detecting personality test faking with appropriateness measurement: Fact or fantasy?* Poster Presented at the 18th annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

* Burkevich, S.M., Jenkins, M., & Griffith, R.L. (2007). *Lying down on the job: Applicant faking and dependability.* Paper Presented at the 22nd annual meeting of the Society for Industrial and Organizational Psychology, New York, NY.

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selec- tion and assessment. In R. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 115–150). Greenwich, CT: Information Age.

* Caldwell-Andrews, A., Baer, R.A., & Berry, D.T.R (2000). Effects of response sets on NEO-PI-R Scores and their relations to external criteria. *Journal of Personality Assessment, 74*, 472-488.

Chiaburu, D., Oh, I., Berry, C., Li, N., & Gardner, R. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96(6),* 1140-1166.

* Christiansen, N.D., Burns, G.N., & Montgomery, G.E. (2005). Reconsidering forced choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307.

* Christiansen, N.D., Goffin, R.D., Johnston, N.G., & Mitchell, M.G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*, 847-860.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Connelly, B.S., & Ones, D.S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092-1122.

* Converse, P.D., Oswald, F.L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on

criterion-related validity and test taker reactions. *International Journal of Selection and Assessment, 16*, 155-169.

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

* Crandall, B.D. (1998). *Item response latency in computerized personality assessment and the effect of socially desirable responding*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI 9827505).

Crowne, D. P., & Marlowe, D. (1964). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology,* 24, 349-354.

* Cucina, J.M., Hunter, A., Martin, N., & Vasilopoulos, N.L. (2010, April). *Empirical keying of personality scales to reduce faking.* Poster Presented at the 24th meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28*, 205-220.

* Cunningham, M.R., Wong, D.T., & Barbee, A.T. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies of the Reid Report. *Journal of Applied Psychology, 79*, 643-658.

* Dalen. L.H., Stanton, N.A., & Roberts, A.D. (2001). Faking personality questionnaires in personnel selection. *The Journal of Management Development, 20*, 729-741.

Dannenbaum, S. E., & Lanyon, R. I. (1993). The use of subtle items in detecting deception. *Journal of Personality Assessment, 61*, 501-510.

Day, N.T. (2008). *Item and person characteristics as predictors of faking.* (Unpublished master's thesis). Wright State University, Dayton, OH.

Dilchert, S., & Ones, D.S. (2011). Application of preventative strategies. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 177-200). New York, NY: Oxford University Press.

* Dodaj, A. (2012). Social desirability and self-reports: Testing a content and response-style model of socially desirable responding. *Europe's Journal of Psychology, 8*, 651-666.

Donovan, J.J., Dwight, S.A., & Schneider, D. (2008, April). *Faking in the real world. Evidence from a field study.* In R.L. Griffith & M.H. Peterson (Chairs), Examining faking using within-subjects designs and applicant data. Symposium conducted at the 23rd Annual conference for the Society for Industrial and Organizational Psychology, San Francisco, CA.

* Douglas, E.F., McDaniel, M.A., & Snell, A.F. (1996, August). The validity of non-cognitive measures decays when applicants fake. *Proceedings of the Academy of Management*, Cincinnati, OH.

Drasgow, F., Stark, S., Chernyshenko, O.S., Dye, C.D., Hulin, C.L., & White, L.A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical report 1311). Ft. Belvoir, VA. U.S. Army Research Institute.

Driskell, J. E., Hogan, R., & Salas, E. (1987). Personality and group performance. In C. Hendrick (Ed.), *Group processes and intergroup relations: Review of personality and social psychology* (Vol. 9, pp. 91-112). Newbury Park, CA: Sage.

* Dullaghan, T.R. (2010). *The effect of a reasoning warning on faking in personality testing for selection and the perception of procedural justice*. (Unpublished masters thesis). University of South Florida, Miami, FL.

Dunlop, P.D., Telford, A.D., & Morrsion, D.L. (2012). Not too little, but not too much: The perceived desirability of responses to personality items. *Journal of Research in Personality, 46,* 8-18.

* Dunnett, S., Koun, S., & Barber, P.J. (1981). Social desirability in the Eysenck Personality Inventory. *British Journal of Psychology, 72*, 19-26.

* Dunnette, M., McCartney, J., Carlson, H.C., & Kirchner, W.K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*, 13-24.

Dunning, D., Heath, C., & Suls, J.M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69-106.

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

Ellingson, J.E. (2011). People only fake when they *need* to fake. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 19-33). New York, NY: Oxford University Press.

* Ellingson, J.E., Heggestad, E.D., & Makarius, E.E. (2012). Personality restesting for managing intentional distortion. *Journal of Personality and Social Psychology, 102*, 1063-1076.

Ellingson, J.E., & McFarland, M. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance*, *24*, 322-337.

* Ellingson, J.E., Sackett, P.R., & Connelly, B.S. (2007). Personality test across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology*, *92*, 386-395.

* 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.

* Elliot, S., Lawty-Jones, M., & Jackson, C. (1996). Effect of dissimulation on self-report and objective measures of personality. *Personality and Individual Differences, 3*, 335-343.

* Ellis, R.A., & Leitner, D.W. (1980). Social desirability as a variable affecting responses on the California Psychological Inventory. *Psychological Reports, 47*, 1223-1226.

Erdheim, J., Wang, M., & Zickar, M.J. (2006). Linking the big five personality constructs to organizational commitment. *Personality and Individual Differences, 41*, 959-970.

* Exner, J.E., McDowell, E., Pabst, J., Stackman, W., & Kirk, L. (1963). On the detection of willful falsifications in the MMPI. *Journal of Consulting Psychology, 27*, 91-94.

* Eysenck, S.B.G., & Eysenck, H.J. (1974). The modification of personality and lie scale scores by special 'honesty' instructions. *British Journal of Clinical Psychology, 13*, 41-50.

* 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 assessments Within Selection Contexts. *Journal of Applied Psychology*, *97*, 866-880.

Farley, F.H., & Goh, D.S. (1976). PENmanship:: Faking the P-E-N. *British Journal of Clinical Psychology, 15*, 139-148.

* Ferrando, P.J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*, 997-1012.

* Fineran, K.R.J. (2009). *Response distortion in normal personality assessment: Investigating proposed validity scales for the NEO-PI-R in a college student sample.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI 3383706).

* Fleisher, M.S., Woehr, D.J., Edwards, B.D., & Cullen, K.L. (2011). Assessing within-person personality variability via frequency estimation: More evidence for a new measurement approach. *Journal of Research in Personality, 45,* 535-548.

Flipikowski, J.N. (2007). *Measuring conscientiousness with explicit and implicit measures* (Unpublished master's thesis). Wright State University, Dayton, OH.

* Fox, S., & Dinur, Y. (1998). Validity of self-assessment: A field study. *Personnel Psychology, 41*, 581-592.

* Frei, R.L., Peterson, M.H., Isaacson, J.A., Griffith, R.L., & Jenkins, M. (2007, April). *Exploring the relationship between academic dishonesty and applicant dishonesty.* Paper presented at the 22nd Annual Convention of the Society for Industrial and Organizational Psychology, New York, NY.

* Furnham, A.F. (1997). Knowing and faking one's Five-Factor personality score. *Journal of Personality Assessment, 69*, 229-243.

* Furnham, A., & Craig, S. (1987). Fakeability and correlates of the perception and preference inventory. *Personality and Individual Differences, 8*, 459-470.

* Furnham, A., & Henderson, M. (1981). The good, the bad, and the mad: Response bias in self-report measures. *Personality and Individual Differences, 3*, 311-320.

* Galic, Z., Jerneic, Z., & Kovacic, M.P. (2012). Do applicants fake their personality questionnaire responses and how successful are their attempts?

A case of military pilot cadet selection. *International Journal of Selection and Assessment*, *20*, 229-241.

* Gammon, A.R. (2010). *Personality measurement and the impact of typical frame of reference modifications for personnel selection.* (Doctoral Dissertation) Retrieved from ProQuest Dissertations and Theses Database. (UMI 3525543).

* Gammon, A.R., Griffith, R.L., & Kung, M. (2011, April). *How do real applicants who are fakers compare to nonfakers?* Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Gibby, R.E. (2004). *Identifying fakers on personality assessments and the properties that make personality items fakable* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (AAT 3135430).

Goffin, R.D., & Christiansen, N.D. (2003). Correcting personality assessments for faking: A review of popular personality assessments and an initial survey of researchers. *International Journal of Selection and Assessment, 11*, 340-344.

* Goffin, R.D., & Woods, D.M. (1995). Using personality testing for personnel selection: Faking and test-taking instructions. *International Journal of Selection and Assesesment, 3*, 227-236.

Goffin, R.D., Jang, I., & Skinner, E. (2011). Forced choice and conventional personality test: Each may have unique value in pre-employment testing. *Personality and Individual Differences, 51*, 840-844.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

* Gordon, L.V., & Stapleton, E.S. (1956). Fakability of a forced-choice personality test under realistic high school employment conditions. *The Journal of Applied Psychology, 40*, 258-262.

Gosling, S.D., Rentfrow, P.J., & Swann, W. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

Grant, A. (2013). Rethinking the extraverted sales ideal: The ambivert advantage. *Psychological Science, 24*, 1024-1030.

* Green, R.F. (1951). Does a selection situation induce testees to bias their answers on interest and temperament tests? *Educational and Psychological Measurement, 11*, 503-515.

* Griffith, R.L. (1998). *Faking of noncognitive selectin devices: Red herring is hard to swallow*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI 9813613).

* 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.

Griffith, R.L., & Converse, P.D. (2011). The rules of evidence and the prevlance of applicant faking. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). New York, NY: Oxford University Press.

* Griffith, R.L., Malm, T., English, A., Yoshita, Y., & Gujar, A. (2006). Applicant faking behavior: Testing apart the influence of situational variance, cognitive biases, and individual differences. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 151-178). Greenwich, CT: Information Age.

Griffith, R.L., & McDaniel, M. (2006). The nature of deception and applicant faking behavior. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior*. Greenwich, CT: Information Age Publishing.

* Griffith, R.L., & Peterson, M.H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 308-311.

* Grubb, W.L., & McDaniel, M.A. (2007). The fakability of Bar-On's Emotinal Quotient inventory short form: Catch me if you can. *Human Performance, 20*, 43-59.

Guion, R. M., & Cranny, C. J. (1982). A note on the concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 67*, 239–244.

* Gujar, A.R., Griffith, R.L., & Yoshita, Y. (2006). *The effect of temporal context on personality measures used in personnel selection.* Paper presented at the 21st Annual Convention of the Society for Industrial and Organizational Psychology, Dallas, TX.

* Haaland, D.E. (1999). *Self-assessment of interpersonal competency: Development and validation of a forced-choice method to minimize response distortion in job applicant contexts*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI 9948365).

* Hakstian, A.R., & Ng, E.L. (2005). Employment-related motivational distortion: Its nature, measurement, and reduction. *Educational and Psychological Measurement, 65,* 405-441.

Harvel, J.L. (2012). *Using the bogus knowledge scale to detect individual differences in faking: Examining the impact of variance in applicant faking* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 3114678).

* Hastey, K.C. (2003). *Would-do, could-do, and should-do: Comparisons among typical, maximal, and faking patterns of personality responding*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 3499830).

* Hedberg, R. (1962). More on forced-choice test fakability. *Journal of Applied Psychology, 46,* 125-127.

Heggestad, E.D. (2011). A conceptual representation of faking: Putting the horse back in front of the cart. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality test* (pp. 87-101). New York, NY: Oxford University Press.

* 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.

Hicks, L.E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184.

* Hirsch, J.B., & Peterson, J.B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality*, *42,* 1323-1333.

* Hoft, H. (unpublished manuscript). *The impact of the test-taking context on responding to personality measures.*

* Holden, R..R. (1995). Reponse latency detection of fakers on personnel tests. *Canadian Journal of Behavioural Science*, *27*, 343-355.

* Holden, R.R. (1998). Detecting fakers on a personnel test: Response latencies versus a standard validity scale. *Journal of Social Behavior and Personality, 13*, 387-398.

* Holden, R.R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioral Science, 39*(3), 184-201.

* Holden, R.R., (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences*, *44*, 311-321.

* Holden, R.R., & Book, A.S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences, 47,* 185-190.

Holden, R.R., & Book, A.S. (2011). Faking does distort self-report personality test. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality test* (pp. 71–84). New York, NY: Oxford University Press.

* Holden, R.R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality, 29*, 362-372.

Holden, R. R., & Jackson, D. N. (1981). Subtlety, information, and faking effects in personality test. *Journal of Clinical Psychology, 37*, 379-386.

* Holden, R.R., Wood, L.L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity. *Journal of Personality and Social Psychology, 81*(1), 160-169.

Hooper, A.C. (2007). *Self-presentation on personality measures in lab and field settings: A meta-analysis* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

* Hough, L.M. (1998). Personality at work: Issues and evidence. In M.D. Hakel (Ed.)*, Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131-166). Mahwah, NJ: Lawrence Erlbaum Associates.

* 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., & Ones, D. S. (2002). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In Anderson, N., Ones, D.S., Sinangil, H.K., & Viswesvaran, C. (Eds.) *Handbook of industrial, work and organizational psychology, Volume 1: Personnel Psychology* (pp. 233-277). Thousand Oaks, CA: Sage Publications Ltd.

Hough, L.M., & Oswald, F.L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(3), 272-290.

Hough, L.M., Oswald, F.L., & Ployhart, R.E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures:

Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Hsu, L.M., Santelli, J., & Hsu, J.R. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. *Journal of Personality Assessment, 53*, 278-295.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (Second Edition). Newbury Park, CA: Sage.

* Hurd, J.M. (2002). *Moving beyond social desirability: A practical method for measuring faking in applied settings*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 3060656).

* Hurtz, G.M., & Alliger, G.M. (2002). Influence of coaching on integrity test performance and unlikely virtues scale scores. *Human Performance, 15,* 255-273.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85,* 869-879.

* Huws, N., Reddy, P.A., Talcott, J.B. (2009). The effects of faking on non-cognitive predictors of academic performance in University students. *Learning and Individual Differences, 19*, 476-480.

* Illingsworth, A.J. (2004). *The effect of warnings and individual differences on the criterion-related validity of selection tests* (Unpublished doctoral dissertation). University of Akron, Akron, OH.

* Impelman, K., Carson, M.A., & Klabzuba, A. (2013, April). *The moderating impact of cognitive ability on faking selection measures*. Paper presented

at the 28th Annual Convention of the Society for Industrial and Organizational Psychology, Houston, TX.

* Isaacson, J.A., Frei, R.L., Quist, J.S., & Griffith, R.L. (2007, April). *The effects of behavioral intentions and opportunity to fake.* Paper presented at the 22nd Annual Convention of the Society for Industrial and Organizational Psychology, New York, NY.

* Jackson, C.J., & Francis, L.J. (1999). Interpreting the correlation between neuroticism and lie scale scores. *Personality and Individual Differences, 26,* 59-63.

* 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.

* Jenson, C.E., & Sackett, P.R. (2012, April). *Faking to the max: Do ceiling effects constrain faking?* Poster presented at the 27th annual meeting for the Society for Industrial and Organizational Psychology, San Diego, California.

* Jeske, J.O, & Whitten, M.R. (1975). Motivational distortion of the Sixteen Personality Factor Questionnaire by persons in job applicants' roles. *Psychological Reports, 37*, 379-382.

* Johnson, J.A. (1987, August). *Dissembling on the Hogan Personality Inventory during simulated personnel selection*. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY.

* Kantrowitz, T., & Robie, C. (2011, April). *Estimates of faking on computer adaptive and static personality assessments.* In Kantrowitz, T. (Chair), Innovations in mitigating faking on personality assessments. Symposium presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

* Khorramdel, L., & Kubinger, K.D. (2006). The effect of speediness on personality questionnaires: An experiment on applicants within a job recruiting procedure. *Psychology Science, 48*, 378-397.

* Komar, J.A. (2013). *The faking dilemma: Examining competing motivations in the decision to fake personality tests for personnel selection*. (Master's thesis). (Unpublished masters thesis). University of Waterloo, Waterloo, Ontario.

* Komar, S., Komar, J.A., Robie, C., Taggar, S. (2010). Speeding personality measures to reduce faking: A self-regulatory model. *Journal of Personnel Psychology, 9*, 126-137.

Komar, S., Brown, D.J., Komar, J.A., & Robie, C. (2008). Faking and the validity of Conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*, 140-154.

* Konradt, U., Syperek, S., & Hertel, G. (2011). Testing on the internet: Faking in a web-based self-administered personality measure. *Journal of Business and Media Psychology, 2*, 1-10.

* 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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

* Lavine, K.A. (2003). Job applicant faking on a personality inventory: Implications for subsequent well-being on the job. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 3096549).

Lawler, E.E., & Suttle, J.L. (1973). Expectancy theory and job behavior. *Organizational Behavior and Human Performance*, *9*, 482-503.

* LeBreton, J.M., Barksdale, C.D., Robin, J., & James, L.R. (2007). Measurement issues associated with conditional reasoning tests: Indirect measurement and test faking. *Journal of Applied Psychology, 92*, 1-16.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion-related validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*, 131-141.

* Liguori, E.W., Taylor, S.G., Choi, S., Kluemper, D.H., & Sauley, K.S., (2011). Testing measures of equity sensitivity for resistance to response distortion. *Journal of Managerial Issues, 13*, 46-61.

* Lopez, F.J. (2009). *Testing a new method designed to manage applicant dissimulation on non-cognitive measures* (Unpublished doctoral dissertation). Hofstra University, Hempstead, NY.

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

Mael, F. (1991). A conceptual rational for the domain and attributes of biodata items. *Personnel Psychology, 44*, 763-792.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*, 63-78.

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.

* Martin, N.R. (2011). *Influence of self-schema on applicant faking* (Unpublished doctoral dissertation,). The George Washington University, Washington, D.C.

McCabe, S.T., & Oswald, F.L. (2013). The criterion-related validity of personality measures for predicting GPA: A meta-analytic validity competition. *Psychological Assessment, 25*, 532-544.

McCrae, R.R., & Costa, P.T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

McDainel, M. A., & Timm, H. (1990). *Lying takes time: Predicting deception in biodata using response latencies*. Paper presented at the 98th Annual Convention of American Psychological Association, Boston, MA.

* McFarland, L.A. (2003). Warning against faking on a personality test: Effects of applicants' reactions and personality test scores. *International Journal of Selection and Assessment, 11*, 265-276.

* McFarland, L.A., & Ryan, A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

McFarland, L.A., & Ryan, A.M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Pscyhology, 36*, 979-1016.

* McFarland, L.A., Ryan, A.M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. *Journal of Personality Assessment, 78*, 348-369.

Meade, A.W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-552.

Meehl, P.E., & Hathaway, S.R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30*, 525-564.

* Merlini, P., Sudduth, M.M., Ricci-Twitchell, M., Kung, M.C., & Griffith, R. (2010, April). *The smart or right choice: Exploring job-related intelligence and faking*. Poster presented at the 25th Annual Society for Industrial and Organizational Psychology, Atlanta, GA.

* Mersman, J.L., & Shultz, K.S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences, 24*, 217-227.

* Miller, C.E. (2000). *The susceptibility of personality selection tests to coaching and faking*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (UMI 9980371).

Mitchell, T., & Adair, C.K. (2014). *An examination of warning type on personality faking*. Poster Presented at the 29th meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI, April 2014.

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

Morgeson, F.P, Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality assessments for personnel selection. *Personnel Psychology, 60*, 1029-1049.

Mount, M.K., & Barrick, M.R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In G.R. Ferris (Ed.), *Research in personnel and human resources management, Vol. 13*. Greenwich, CT: JAI Press.

* Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348–355.

Mueller-Hanson, R., Heggestad, E.D., & Thornton, G.C. III (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science, 3,* 288-312.

* Neubauer, A.C., & Malle, B.F. (1997). Questionnaire response latencies: Implications for personality assessment and self-schema theory. *European Journal of Psychological Assessment, 13*, 109-117.

* Nguyen, D. (2008). *Assessing personality through interviews: Examining the role of self monitoring and impression management in faking behavior.* (Master's thesis). Retrieved from ProQuest Dissertations and Theses Database (UMI 1454513).

* 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.

* O'Brien, E., & LaHuis, D.M. (2011). Do appliccants and incumbents respond to personality items similarl? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment, 19*, 109-118.

* O'Connell, M.S., Kung, M.C., & Tristan, E. (2011). Beyond impression management: Evaluating three measures of response distortion and their relationship to job performance. *International Journal of Selection and Assessment, 19*, 340-351.

* O'Neill, T. A., Lewis, R. J., Carswell, J. J., & Law, S. (2013, April). *Pre-employment personality test faking and the forced-choice method*. Poster presented at the 28th annual meeting of the Society for Industrial and Organizational Psychology, Houston, TX.

Ones, D.S., & Viswesvran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245-269.

Ones, D.S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 589-609.

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego: Academic Press.

Paulhus, D.L., & Reid, D.B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60*, 307-371.

Paulhus, D.L., Robins, R.W., Trzesniewski, K.H., & Tracy, J.L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research, 39,* 303-328.

* 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.

Paunonen, S.V., & LeBel, E.P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Applied Personality and Social Psychology,* 103, 158-175.

* Peterson, M.H. (2010). A theory of applicant faking: The interaction of dispositional characteristics, item characteristics, and respondent cognitive processes. (Doctoral dissertation). Received from ProQuest Dissertations and Theses Database. (UMI 3405008).

* Peterson, M.H., Burkevich, S.M., Merlini, P., & Griffith, R.L. (2007). *Locus of control and applicant faking: Direct and convergent evidence*. Poster presented at the 22nd annual meeting of the Society for Industrial and Organizational Psychology, New York, NY.

* Peterson, M.H., Griffith, R.L., & Converse, P.D. (2009). Examining the role of applicant faking in hiring decisions: Percentage of fakers hired and hiring discrepancies in single- and multiple-predictor seletion. *Journal of Businss and Psychology, 24*, 373-386.

* Peterson, M.H., Griffith, R.L., Converse, P.D., & Gammon, A.R. (2011, April). *Using within-subjects designs to detect applicant faking*. Paper presented

in Converse, P.D. (2011, April) Detecting deception: Techniques for assessing applicant faking on personality measures. Symposium conducted at the 26th Annual Conference for the Society for Industrial and Organizational Psychology: Chicago, IL.

* Peterson, M.H., Griffith, R.L., O'Connell, M.S., & Isaacson, J.A. (2008, April). *Examining faking in real job applicants: A within-subjects investigation of score changes across applicant and research settings*. Paper presented at the 23rd annual meeting for the Society for Industrial and Organizational Psychology, San Francisco, California.

* Peterson, M.H., Griffith, R.L., Isaacson, J.A., O'Connell, M.S., & Mangos, P.M. (2011). Applicant faking, social desirability, and prediction of counterproductive work behaviors. *Human Performance, 24*, 270-290.

Ployhart, R.E., McFarland, L.A., & Ryan, A.M. (2002). Examining applicants' attributions for withdrawal from a selection procedure. *Journal of Applied Social Psychology, 32*, 2228-2252.

* Quist, J.S. (2010). Differential item weighting: Improving assessment validity and reducing the impact of applicant faking. (Doctoral Dissertation). Retreived from ProQuest Dissertations and Theses. (UMI 3406632).

* Quist, J.S., Arora, S., & Griffith, R.L. (2007, April). *Social desirability and applicant faking behavior: A validation study*. Paper presented at the 22nd annual meeting for the Society for Industrial and Organizational Psychology, New York, New York.

* Ramakrishnan, M. (2005). *Moving beyond traditional warnings: Effects of alternative instructions on faking and applicant reactions* (Unpublished Doctoral dissertation). University of Akron, Akron, OH.

Reeder, M.C., & Ryan, A.M. (2011). Methods for correcting for faking. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 201-213). New York, NY: Oxford University Press.

* Reid-Seiser, H.L., & Fritzsche, B.A. (2001). The usefuleness of the NEO PI-R positive presentation management scale for detecting response distortion in employment contexts. *Personality and Individual Differences, 31*, 639-650.

* Robie, C., Curtin, P.J., Foster, C., Phillips, H.L., Zbylut, M., & Tetrick, L.E. (2000). The effect of coaching on the utility of response latencies in detecting fakers on a personality measure. *Canadian Journal of Behavioral Science, 32*, 226-233.

* Robie, C., Komar, S., & Brown, D.J. (2010). The Effects of Coaching and Speeding on Big Five and Impression Management Scale Scores. *Human Performance, 23*, 446-467.

* Robie, C., Taggar, S., & Brown, D.J. (2009). The effects of warnings and speeding on scale scores and convergent validity of conscientiousness. *Human Performance, 22*, 340-354.

Robie, C. Zickar, M.J., & Schmit, M.J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

* Robson, S.M., Jones, A., & Abraham, J. (2008). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance*, *21*, 89-106.

* Ross, S.R., Bailley, S.E., & Millis, S.R. (1997). Positive self-presentation effects and the detection of defensiveness on the NEO PI-R. *Assessment, 4*, 395-408.

* 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.

Rothstein, M.G., & Goffin, R.D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155-180.

* Rusmore, J.T. (1956). Fakability of the Gordon Personal Profile. *The Journal of Applied Psychology, 40*, 175-177.

Sackett, P. (2011). Integrating and prioritizing theoretical perspectives on applicant faking of personality measures. *Human Performance, 24*, 379-385.

Sanchez, R.J., Truxillo, D.M., & Bauer, T.N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology*, *85*, 739-750.

Schnell, K. F., Oswald, F. L., Scobel, E. L., Mitchell, M. A., Boronow, E., Marfisi, A. E., Glutz, M. J., Tran, D., & Hartman, M. (2011, April). *Item grouping and item randomization effects in personality measurement*. Paper presented at the 26th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Schmidt, F.L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206-224.

* Schmit, M.J., Ryan, A.M., Stierwalt, S.L., & Powell, A.B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*, 607-620.

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.

* Schnure, K.A. (2009). *Response distortion and social desirability in high-level executives.* (Unpublished masters thesis). Georgia Institute of Technology, Atlanta, GA.

Schriesheim, C.A., & DeNisi, A.S. (1980). Item presentation as an influence on questionnaire validity: A field experiment. *Educational and Psychological Measurement, 40*, 175-182.

Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement, 13*, 19-32.

* Schwab, D.P. (1971). Issues in response distortion studies of personality inventories: A critique and replicated study. *Personnel Psychology, 24*, 637-647.

* Shores, E.A., & Carstairs, J.R. (1998). Accuracy of the MMPI-2 computerized Minnesota report in identifying fake-good and fake-bad response sets. *The Clinical Neuropsychologist, 12*, 101-106.

* Sisco, H., & Reilly, R.R. (2007). Five factor biodata inventory: Resistance to faking. *Psychological Reprts, 101*, 3-17.

* Sliter, K.A., & Christiansen, N.D. (2012). Effects of targeted self-coaching on applicant distortion of personality measures. *Journal of Personnel Psychology, 11*, 169-175.

* 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.

* Smith, D.B., Hanges, P.J., & Dickson, M.W. (2001). Personnel selection and the Five Factor Model: Reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology, 86*, 304-315.

Smith, D.B., & McDaniel, M. (2011). Questioning old assumptions: Faking and the personality-performance relationship. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 53-70). New York, NY: Oxford University Press.

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.

* Star, K.H. (1962). Ideal-self response set and maudsley personality inventory scores. *Psychological Reports, 11*, 708.

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

Stanush, P.L. (1997). *Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.

* 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.

* Stober, J. (2001). The social desirability scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment, 17,* 222-232.

Tett, R.P. Anderson, M.G., Ho, C.L., Yang, T.S., Huang, L., & Hanvongse, A. (2006). Seven nested questions about faking on personality assessments. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 43 - 83). Greenwich, CT: Information Age.

Tett, R.P., & Christiansen, N. (2007). Personality tests at a crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. *Personnel Psychology, 60*, 967-993.

* Tett, R.P., Freund, K.A., Christiansen, N.D., Fox, K.E., & Coaster, J. (2012). Faking on self-report emotional intelligence and personality assessments: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences, 52*, 195-201.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

* Thumin, F.J., & Barclay, A.G. (1993). Faking behavior and gender differences on a new personality research instrument. *Consulting Psychology Journal, 45*, 11-22.

* Tonkovic, M. (2012). Are there personality traits that predispose applicants to fake noncognitive measures in personnel selection? *Review of Psychology, 19*, 26-36.

* Topping, G.D., & O'Gorman, J.G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23*, 117-124.

Tristan, E. (2009). *Measuring applicant faking with job desirability: Prevalence, selection, and measurement issues in an applied sample* (Unpublished doctoral dissertation). Wright State University, Dayton, OH.

* Tryba, B.A., & Griffith, R.L. (2012). *The next step: How do people fake*? Poster presented at the 26th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

* Tryba, B., Griffith, R.L., Jackson, P. Lilly, S., Wells., S., Mochinushi, Y., Peterson, M.H., & Gammon, A. (2013). *End of the world as we know it? Faking and organizational criterion*. In Robie, C. (Chair), Advances in the use of personality to predict organizational criteria. Symposium presented at the 28th annual meeting of the Society for Industrial and Organizational Psychology, Houston, TX.

* 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 computer adaptive personality measure. Technical Report No. NPRST-TR-08-2, Millington, TN: NPRST.

* van Hooft, E. A. J., & Born, M. P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology, 97,* 301-316.

* van Iddekinge, C.H., Raymark, P.H., & Roth, P.L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology, 90*, 536-552.

* Vasilopoulos, N.L. (1999). *The impact of job familiarity and warning of response verification on the relationship between response latency and impression management: A field investigation.* (Doctoral dissertation). Retreived from ProQuest Dissertations and Theses Database. (UMI 9903562).

* Vasilopoulos, N.L., Cucina, J.M., & McElreath, J.M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90*, 306-322.

* Vasilopoulos, N.L., Reilly, R.R., & Leaman, J.A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50-64.

* Velicer, W.F., & Weiner, B.J. (1975). Effects of sophistication and faking sets on the Eysenck Personalkty Inventory. *Psychological Reports, 37*, 71-73.

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

Van Eerde, W., & Thierry, H. (1996). Vroom's expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology*, *81*, 575-586.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Widhiarso, W. (2011). *Typology of ideal personality: Applicant perspective in job selection setting.* Paper presented at the 1st international conference on Information Systems for Business Competitiveness, Semarang, Indonesia.

* Wilta, N.E., Meyer, R.D., & Collins, B.J. (2011). *Developing and validating a faking detection scale for the CRT-A*. Paper presented at the 28th annual conference of the Society for Industrial and Organizational Psychology, Houston, TX.

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

White, L.A., Young, M.C., & Rumsey, M.G. (2001). ABLE implementation issues and related research. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 525–558). Mahwah, NJ: Erlbaum.

* Wolford, K.A. (2009). *Effects of item randomization and applicant instructions on distortion on personality measures* (Unpublished master's thesis). Bowling Green State University, Bowling Green, OH.

* Worthy, R.L. (2011). *Faking in personality assessments: An investigation of a method factor measure of faking*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI 1492311).

* Yang, W. (2006). *Effect of applicant faking on measurement properties of the Global Personality Inventory*. (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Yu (2008). *A process model of applicant faking on overt personality assessments* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.

* Zalinski, J.S., & Abrahams, N.M. (1979). The effects of item context in faking personnel selection inventories. *Personnel Psychology, 32*, 161-165.

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

Zickar, M., Rosse, J., & Levin, R. (1996, April). *Modeling the effects of faking on personality scales*. Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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

Appendix A

Coder Instructions & Guide

Personality Faking Interventions Meta-Analysis

Coding Information & Guide

1. Type of publication
    a. Very few of the articles should be coded as unpublished manuscripts (6). Many of them are either from conference presentations (3) or are theses/dissertations (4/5).

2. Big Five Factors measured
    a. Each personality variable will get a different code sheet. Most likely, conscientiousness and neuroticism will receive more code sheets than others. Be sure to keep all code sheets together with the same focal article, even though they will ultimately be separated into distinct meta-analyses for each trait.
    b. *a priori vs. post hoc*
        i. Some measures of the Big 5 are direct measures of these traits (1). Examples include:
            1. NEO (FFI or PI-R), IPIP, HOGAN PERSONALITY INVENTORY (HPI), BFI
        ii. Others will measure one of the Big 5 indirectly. These types of measures will be converted to the Big 5 *post hoc* (2). Use the table at the end of this sheet to identify how to code one of these scales.

3. Type of faking
    a. Naturally Occurring
        i. Code a study as naturally occurring faking if participants are not explicitly told to fake. This is most likely the case when comparing job incumbents to job applicants, where job applicants have "naturally occurring" faking.
    b. Instructionally Induced
        i. Code a study as instructionally induced if the participant is told to respond in a certain pattern. If there is a monetary incentive given to "top performers," code as instructionally induced. If participants are instructionally induced to "fake bad" or respond poorly, do not include in the analysis.
        ii. Fake good – respond as applicant
            1. If the instructions explicitly state for the participant to "respond as applicant", code as (1). If there is no explicit statement of a job context, but participants are still instructed to respond positively, code as (2).
            2. Note whether a job context is provided. A job context could be applying for a sales, nursing, management, etc. position.

4. Monetary Incentive

    a. Again, if monetary incentives are offered to "top performers", the study should be coded as "instructionally induced" (2). Note the amount given per participant if reported.

5. Criterion
    a. School Success
        i. This involves students, classes, grades, school environment, or learning in a school setting
    b. Job Performance
        i. Some clear measure of job performance (most often managerial performance review). The setting is at an actual work place, participants are workers, in work environment. Likely to be a field setting, doing things that would be done on the job.
    c. Task Performance
        i. The task involves things that may be done on a job but the study is not in the participant's work environment during the "typical" workday. Participants may or may not be actual workers/employees, could be students. (i.e., students acting as stoker brokers, group decision making on work issue) or workers in a simulation.
        ii. The task can not be a training task designed for a job (that will be a separate code), but can be a task done in a job setting.
        iii. Key differentiator between (b) and (c) is that: (b) is a measure of job performance used at the individual's place of work.
    d. Behavior on the job
        i. Some studies examine counterproductive work behaviors (CWBs) such as theft, deviance, absence from work, etc. This is often measured through actual behaviors.
        ii. Organizational Citizenship Behaviors (OCBs) represent "work behaviors that support the broader organizational, social, and psychological environment."
        iii. Either CWBs or OCBs may be measured via scale (e.g., Dineen et al., 2006 – based off of a critical incidents technique).
    e. Training performance
        i. This criterion involves performance within the context of training and may include the acquisition of skills and/or knowledge. This training may simulate a normal workday, but it is not done during work hours and is not a final measure of job performance.

6. Study Context
    a. Lab: creating a setting to conduct research, generally able to impose more experimental control as to participants and setting.

b. Applied/Real World: Research is in a more natural setting, people do not perceive the setting to have been created to conduct the research, may be using an existing group.
c. Simulation: No actual participant data collected, a simulation is run based on explicitly defined parameters.

7. Study Design
    a. Between Subjects: Two distinct, independent groups participate under different conditions.
    b. Within Subjects: Same participants are exposed to multiple conditions. This is most likely the case when participants take a test once without an intervention and then once with it.

8. Interventions
    a. Score Correction
        i. Primary studies may partial out the effect of a social desirability or lie scale from the correlation between personality and a criterion. Code the article as Score Correction (1) if the article explicitly states that the correlations accounted for or corrected for the effect of SD.
    b. Removal of Cases
        i. Code a study as Removal of Cases (2) if the correlation or validity coefficient reflects a limited number of the participants based on their responses to Social Desirability or a lie scale.
    c. Warning Statement
        i. Code a study as Warning Statement (3) if there is/are explicit warning(s) against faking. This may come in five categories:
            1. Detection (faking can be detected)
            2. Consequential (penalties will be enforced if you are identified as a faker)
            3. Appeal to moral principles (as a moral person, faking is wrong)
            4. Appeal to reason (responding accurately will more accurately portray your personality characteristics)
            5. Educational (faking a response will not allow researchers to evaluate the responses)
    d. Time Limit
        i. If there was any mention of a time limit, make a note of that.
        ii. There may not always be a “timed” and “untimed” group.
        iii. If there was no mention of a time limit, mark that there was no time limit given
    e. Forced Choice

i. Code an article as Forced Choice if it meets any of the characteristics listed in the codebook.
ii. Similar to time limit, there may not be a forced choice vs. likert condition.
iii. If no mention of forced choice, standard likert format applies.

f. Item Transparency
i. Both item transparency and item order are applicable to this category
ii. Blocked items means that the article explicitly states that all items measuring the same construct appear together.
iii. Randomized items means that there is no order to the items *a priori*
iv. Subtle items means that a technique was used to make the underlying construct less visible
1. Empirically-keyed items satisfy the "subtle items" code

Effect Size Coding

- NOTE: There are separate coding sheets for within vs. between subjects designs
- Define the condition (e.g., fake-good, warned, etc.)
- Record the number of participants for each condition
- Record the means and standard deviation of the personality variable for each condition
- If there are multiple iterations of the intervention, report the mean, SD, and N for each (these will be averaged).
    - For instance, if multiple warning "types" are administered.
- Record the mean and SD of the social desirability (if applicable). If social desirability is not broken down between impression management/self-deception, just record one and cross out the IM/SD distinction.
- Record the correlation between personality and a criterion (if applicable)
- Multiple studies with different subjects (Study 1, Study 2) get their own effect sizes

Guide for Identifying Traits within Big 5 Framework (Birkeland et al., 2006)

Extraversion		
Number	Inventory	Scale
1	Adjective Checklist	Assertiveness
2	16 Personality Factors (16 PF)	Reserved (R) Submissive (R) Sober (R) Shy (R)

		Dominance Liveliness Social Boldness Privateness (R)
3	Gordon Personal Profile Inventory (GPI)	Sociability Ascendancy
4	Hogan Personality Inventory (HPI)	Sociability Likes parties Entertaining Experience setting Likes crowds
5	Assessment of Background and Life Experiences (ABLE)	Dominance Energy Level
6	Edwards Personal Preference Schedule (EPPS)	Exhibition Affiliation
7	Matthew temp	Gregariousness
8	Psychotisim, Extraversion, Neuroticism (PEN)	E
9	California Psychological Inventory (CPI)	Sociability Social presence Dominance Capacity for status Empathy Self-acceptance
10	PPI	Influence

Agreeableness		
Number	Inventory	Scale
1	Adjective Checklist	Cooperativeness
2	16 Personality Factors (16 PF)	Tough minded Natural Privateness (R) Warmth Sensitivity
3	Hogan Personality Inventory (HPI)	Likeability
4	Assessment of Background and Life Experiences (ABLE)	Cooperativeness
5	Edwards Personal Preference Schedule (EPPS)	Deference Abasement Succorance Nuturance

Openness		
Number	Inventory	Scale
1	Adjective Checklist	Imaginativeness
2	16 Personality Factors (16 PF)	Less intelligent (R) Practical (R) Conservative (R) Openness Abstractness
3	Hough	Adaptability Openness
4	California Psychological Inventory (CPI)	Intellectual efficiency
5	Global Personality Inventory (GPI)	Openness to change
6	Edwards Personal Preference Schedule (EPPS)	Change

Neuroticism / Emotional Stability		
Number	Inventory	Scale
1	Adjective Checklist	Calmness
2	16 Personality Factors (16 PF)	Emotionally unstable Trusting Self-assured Relaxed NPF Emotional stability Tension
3	Gordon Personal Profile Inventory	Emotional stability
4	Guilford Martin	I N D C M
5	Hough	Resilient self-esteem
6	Assessment of Background and Life Experiences (ABLE)	Self-esteem Emotional stability
7	Matthew temp	Maladjustment
8	PSYCHOTICISM, EXTRAVERSION, NEUROTICISM (PEN)	N
9	California Psychological Inventory (CPI)	Well-being IndePsychoticism, Extraversion, Neuroticism

		(PEN)dence Self-control Neuroticism
10	PPI	Emotional control Self-tolerance

Conscientiousness		
Number	Inventory	Scale
1	Adjective Checklist	Conscientiousness
2	16 Personality Factors (16 PF)	Expedient (R) Poorly integrated (R) Rule-conscientiousness Vigilance Perfectionism
3	Gordon Personal Profile Inventory	Responsibility
4	Hough	DePsychoticism, Extraversion, Neuroticism (PEN)dability Achievement Detail Mindedness
5	Assessment of Background and Life Experiences (ABLE)	Work orientation Conscientiousness
6	Edwards Personal Preference Schedule (EPPS)	Endurance Achievement Order Consistency
7	Matthew temp	Thoughtfulness
8	California Psychological Inventory (CPI)	Responsibility Ach via conformance Socialization Flexibility Self-acceptance
9	PPI	Work-focus

Appendix B

Code Sheet

Personality Faking Interventions Meta-Analysis

Coder Name: _____ **Article Code: _____**

Study Number: _____ **Publication Year: _____**

Authors: _____

Type of Publication (choose one):

_____ Journal Article (1) **_____ Book Chapter (2)** **_____ Conf Paper/Presentation (3)**

_____ Masters Thesis (2) **_____ Doctoral Dissertation (5)** **_____ Unpublished Manuscript (6)**

DEFINITION OF FAKING

_____ Naturally Occurring (1)
_____ Instructionally-Induced (2)
- **_____ Respond as Applicant (3)**
  - **_____ Specific info given (4)**
  - **_____ Unspecified context (5)**
  - **_____ Can't be determined (6)**
- **_____ No context given (7)**
- **_____ Fake to the Max (8)**
- **_____ Cannot be determined (9)**

_____ Cannot be determined (99)

STUDY CONTEXT (choose one):

_____ Lab (1)
_____ Applied/Real World (2)

Did the study use a student sample?

_____ YES (1) **_____ NO (2)**

DEMOGRAPHICS:

_____ Percent Male

_____ Percent White

_____ Average Age

_____ Average Hours Worked

STUDY DESIGN (Choose one)

_____ Between-subject (1)
_____ Within-subjects (2)
- **____ Honest followed by faking (1)**
- **____ Faking followed by honest (2)**
- **____ Counterbalanced (3)**

_____ Predictive Validity (3)

_____ Concurrent Validity (4)
____ Cannot be determined (99)

If within-subjects or predictive, what is time lag between administrations?

Was there incentive to perform well?

_____ Yes (1)
Name incentive:
_____ No (2)

Was GMA measured?

_____ Yes (1)
How was it measured? _____
_____ No (2)
(*need to code specific values at end of codebook)*

Is there any dependent data in the study?

_____ Yes (1)
_____ No (2)

If YES, Explain below:

INTERVENTION CODE SHEET (Use a separate sheet for each trait/intervention)

____ Score Correction (1) OR
____ Removal of Cases (2)

____ Social Desirability Scale (1)
_____ Edwards (1)
_____ BIDR (2)
_____ Impression Management (1)
_____ Self-deception (2)
_____ Both (3)
_____ Cannot be determined
_____ K scale on MMPI (3)
_____ Marlowe-Crowne (4)
_____ OTHER (5): _____
____ Unlikely Virtues Scale (2)
____ ABLE Validity Scale (3)
____ Bogus Item Scale (4)

_____ Warning Statement (3)
_____ Detection (1)
_____ Consequential (2)
_____ Appeal to Moral Principles (3)
_____ Appeal to Reason (4)
_____ Educational (5)
_____ Cannot be determined (99)

_____ Item Transparency (4)
_____ Randomized Items (1)
_____ Subtle content (2)
_____ Cannot be determined (99)
Note: Empirical/criterion keyed satisfies "subtle" code

_____ Time Limit (5)
Time given in seconds: _____
Number of items: _____
Ratio of seconds to items: _____

Once for each study, use the following indicators of Item Transparency (circle all that apply):
1. Section Headers: Yes = 1 No = 0 DK
2. Blocked Items: Yes = 1 No = 0 DK
3. Construct name in item stem: Yes = 1 No = 0 DK
4. Rationally keyed items: Yes = 1 No = 0 DK
TOTAL: _____

_____ Forced Choice (6)

Put an X if any of the following are true.
Satisifes the partially ipsative if any have an X. If the scale identifies as Forced Choice, with no X's, it is fully ipsative

_____ Respondents only partially order item alternatives, rather than ordering them completely
_____ Scales have differing number of items
_____ Not all alternatives ranked by respondents are scored
_____ Scales are scored differently for respondents with different characteristics, or are referred to different normative transformations on the basis of respondent characteristics
_____ Scored alternatives are differentially weighted
_____ One or more of the scales from the ipsative predictor set is deleted when data are analyzed
_____ The test contains normative sections

Based on checklist above to Forced Choice, what type of scale?
_____ Fully Ipsative (1) **_____ Partially Ipsative (2)**

What criterion was used?

____ Personality trait (1) (a measure of personality, often comparing mean levels with and without intervention)
____ Self-reported faking (2) (a self-report measure, often a direct question of faking).
____ School Success (3) (e.g., grades, school environment, or learning in a school setting)
____ Task Performance (4) (things that may be done on a job but the study is **not** in an existing work environment during the "typical" work day).
____ Behavior on job (5)
_____CWB (theft, absence from work, etc. This could be measured through actual behaviors or through the use of a test/measure).
_____ OCB (prosocial work behaviors – support the broader organizational, social, and psychological environment).
_____ Job Performance (4) (some clear measure of job performance **in work setting** – e.g., number of widgets made, performance review)
____ Training Performance (6) (performance during the training, **not** during the "typical" work day)
____ Other (7): _____
IF 3 through 7 below, what is the reliability of the criterion? _____
Is the reliability coefficient measured via coefficient alpha? Y N If N, name:

Big Five Factors measured (choose one, separate code sheet for each trait within study)
NOTE: If *Post Hoc, complete this section after coding. In this codesheet, note the trait as identified in the primary article.*

_____ Conscientiousness (1) **_____ Agreeableness (2)** **_____ Extraversion (3)**

_____ Emotional Stability (4) **_____ Openness (5)** **_____ Other (6) _____**

How was the trait measured?

***A Priori* FFM Scales**

_____ NEO (1) **_____ IPIP (2)** **_____ HPI (3)** **_____ BFI (4)**
_____ Golberg (5) **____ Other *a priori*:**

***Post Hoc* FFM Scales**

_____ EPI (6) **_____ MMPI (7)** **_____ 16PF (8)** **_____ CPI (9)**
_____ Adj Checklist (10) **_____ Other *post hoc*:**

If Post Hoc (6 – 10), name trait here: _____

Reliability of personality scale marked above for specified trait: _____
Is the reliability coefficient measured via coefficient alpha? Y N If N, name:

Between Groups Design

	Control (Honest)	Experimental (Faked)
N		
Mean Personality (or SRF)		
Std. Dev. Personality (or SRF)		
Crit-Related Validity (personality)		

EFFECT SIZE (d):

Within Subjects Design

	Control (Honest)	Experimental (Faked)
N		
Mean Personality (or SRF)		
Std. Dev. Personality (or SRF)		
Crit-Related Validity (personality)		

EFFECT SIZE (d):

Correlations - For studies w/ correlations that don't fit in previous table (e.g., correlation of time w/ traits)					
Criterion	N	Control (honest)	Experimental (faked)	z (diff in r's)	d (effect size)