Journal of Personality and Social Psychology

0022-3514/12/$12.00 DOI: 10.1037/a0028165

Socially Desirable Responding and Its Elusive Effects on the Validity of Personality Assessments

Sampo V. Paunonen and Etienne P. LeBel
University of Western Ontario

Past studies of socially desirable self-reports on the items of personality measures have found inconsistent effects of the response bias on the measures' predictive validities, with some studies reporting small effects and other studies reporting large effects. Using Monte Carlo methods, we evaluated various models of socially desirable responding by systematically adding predetermined amounts of the bias to the simulated personality trait scores of hypothetical test respondents before computing test–criterion validity correlations. Our study generally supported previous findings that have reported relatively minor decrements in criterion prediction, even with personality scores that were massively infused with desirability bias. Furthermore, the response bias failed to reveal itself as a statistical moderator of test validity or as a suppressor of validity. Large differences between some respondents' obtained test scores and their true trait scores, however, meant that the personality measure's construct validity would be severely compromised and, more specifically, that estimates of those individuals' criterion performance would be grossly in error. Our discussion focuses on reasons for the discrepant results reported in the literature pertaining to the effect of socially desirable responding on criterion validity. More important, we explain why the lack of effects of desirability bias on the usual indicators of validity, moderation, and suppression should not be surprising.

AQ: 1 *Keywords:* ●●●

The purpose of this article is to report on the results of a study in which we evaluated the well-known desirability bias often observed in people's self-reports, commonly referred to as socially desirable responding (SDR). Our particular interest with this bias is its potential effects on the predictive validity of conventional personality assessments. Different opinions have been expressed about the extent to which desirability-biased self-reports present a problem for the measurement of personality (and other) characteristics, with some researchers believing the problem to be trivial and others believing it to be substantial. Our strategy for providing some resolution to this issue was to design a Monte Carlo simulation procedure that allowed us to add predetermined amounts of desirability bias into distributions of personality questionnaire scores and to determine the ensuing effects on the validity of those scores.

The results of our study revealed that whether desirability bias in self-reports has a large or small effect on personality assessment validity depends on the type of validity evaluated. Specifically, applying a correlation-based validity-estimation procedure could lead to the erroneous conclusion that desirability is not a problem, despite the fact that the sample of personality trait scores is severely compromised by distorted responses. Before presenting those results, we review some of the ways in which desirability bias has been conceptualized. We also summarize some past theorizing and data regarding the effects of SDR on test validity.

Conceptualization of Desirability Bias

Self-report measures of personality have long been criticized for their susceptibility to various types of response distortion (Bernreuter, 1933; Edwards, 1957, 1970; Meehl & Hathaway, 1946; Vernon, 1934). The suggested problems stem from the fact that common personality inventories are measures of typical performance (Cronbach, 1960, p. 29), where the scales' items have no right or wrong answers in any universal sense. Unlike maximal performance measures, such as achievement or aptitude tests, the conventional personality item does not have an inherent correct response. This means that, regardless of which of an item's response alternatives is endorsed by a respondent, the test administrator generally cannot be certain if that choice is in fact the "correct" one or whether some other response is a better description of the person.

The primary response distortion that has been studied in the personality assessment literature is a motivated and directional misrepresentation by the respondent of his or her characteristics. That distortion is commonly referred to as SDR. SDR has a very simple and straightforward manifestation—the test respondent is predisposed or biased to select as self-descriptive the response options for items that are more desirable than warranted by his or her corresponding traits or behaviors. The person may be con-

AQ: 10 Sampo V. Paunonen and Etienne P. LeBel, Department of Psychology, University of Western Ontario, London, Ontario, Canada.

This research was supported by Social Sciences and Humanities Research Council of Canada Research Grant 410-2010-2586 to Sampo V. Paunonen and Post-Doctoral Fellowship 756-2011-0090 to Etienne P. LeBel.

Correspondence concerning this article should be addressed to Sampo V. Paunonen, Department of Psychology, University of Western Ontario, London, Ontario N6A 5C2, Canada. E-mail: paunonen@uwo.ca

sciously engaging in a deliberate strategy of misrepresentation to make a good impression on those who might eventually see his or her personality profile, or the misrepresentation could occur at an unconscious level and be motivated by a latent need for self-enhancement and ego maintenance (Paulhus, 1984; Paulhus & John, 1998).[1] Fn1

The tendency to engage in SDR is conceived of as an individual-difference variable, representing a continuum with a positive and a negative end (Edwards, 1970; Jackson & Messick, 1962). Like a conventional bipolar personality trait (see Paunonen & Hong, in press), there is a distribution of SDR levels in the population, with some people at the positive pole, some at the negative pole, and most somewhere in the middle (but biased toward the positive end). With regard to the negative pole of the desirability dimension, it must be acknowledged that individuals exist who are predisposed to choose the undesirable response options when presented to them (e.g., McGrath, Mitchell, Kim, & Hough, 2010; Winder, O'Dell, & Karson, 1975). Such people might be low in self-esteem and have unduly negative self-evaluations, or they might be high in humility and feel compelled to downgrade themselves on their good qualities, or they might be deliberate malingerers who have a particular incentive for making a poor impression (e.g., avoiding conscription to military service).

SDR is a drive or need of the person, and like any personality trait, it is elicited by presses within the situation. One aspect of the situation that is important in this respect is the behavior domain under consideration—some domains are more evaluative and elicit more desirability responding than others. For example, being asked to describe one's integrity would likely have a higher press for SDR than would describing one's orderliness. Note that the interpersonal context in which the assessment is made can moderate the press for SDR for any given behavior domain. Describing one's integrity to an employer, for instance, would presumably have a higher press for desirability than would describing integrity to a casual acquaintance. Also, a person's true level of the trait will affect his or her motivation to respond desirably. Someone possessing a high amount of a desirable attribute does not have the same need to engage in SDR to convey a good impression with respect to that behavior domain as does someone lacking the attribute (McFarland & Ryan, 2000).

SDR is considered a general stylistic tendency of the person, meaning that its expression is not limited to just his or her responses on a personality questionnaire (Jackson, 1971). A person's SDR tendencies can be elicited by all manner of typical performance measures, such as attitude surveys, values questionnaires, occupational interest forms, and the like. It can be manifested in true–false item responses, Likert-type scale ratings, adjective checklist choices, and prose descriptions of self. Its expression is not limited to written self-descriptions but also can be seen in verbal descriptions (e.g., boasting about one's achievements) and in other nontest behaviors (e.g., displaying one's trophies and prizes).

Desirability Bias and Test Validity

Personality assessors have long known about the problem of desirability bias in respondents' item endorsements and its implications for test validity. Most proposed solutions to the problem have fallen into one of two not mutually exclusive camps (e.g., see Nederhof, 1985). The first involves addressing characteristics of the psychometric instrument itself. The basic idea here is very straightforward—minimize the response bias in the personality measure by using items neutral in desirability, for it is only on highly evaluative items (positive or negative) that a respondent has the opportunity of expressing this bias. This solution to SDR might prove difficult, however, if the trait being measured is inherently desirable (e.g., ambition) or undesirable (e.g., defensiveness). The second strategy for dealing with desirability bias involves addressing the context of the assessment. That notion is to minimize the motivation to engage in SDR by urging people to respond honestly for ethical reasons, for example, or by telling them that the test materials include validity scales that can detect faking (e.g., Montag & Comrey, 1982). The problem with this strategy is that it might not be effective in conditions where the motivation to misrepresent self is exceptionally strong (e.g., job-related assessments).[2] Fn2

If someone chooses to endorse the items of a personality measure untruthfully, then, by definition, that undertaking will negatively affect the construct validity of the assessment. That is, the person's estimated standing on the trait based on his or her obtained test score will be too high or too low vis-à-vis his or her true trait level. The question is, how will that response distortion, which

[1] Some researchers use the term *faking* synonymously with SDR (e.g., Ziegler & Buehner, 2009). We prefer to reserve the former term for a different type of misrepresentation of attributes, one that is not necessarily desirable in a universal sense. That type of misrepresentation has been called *role faking* by Kroger (1967; Kroger & Turnbull, 1970). It is where the direction of distortion is determined by the trait content within the personality scale rather than by its desirability, and any relation with general desirability is only incidental. Consider, for example, someone completing personality scales related to nurturance, activity level, and aggression. A person engaging in SDR would conceivably describe self as high in nurturance, neutral in activity level, and low in aggression. Someone faking the role of, say, combat soldier might characterize self as low, high, and high on those same three traits. SDR could be considered a special case of role faking, where the role in question in something akin to "good person." Furthermore, an average test respondent arguably possesses the ability to engage in SDR successfully by correctly choosing the most desirable item responses in some general sense. Yet the same respondent could easily fail at role faking if the role is unfamiliar to him or her or if it is one about which the person harbors invalid stereotypes.

[2] An issue in the personality literature on SDR deserves mention at this point, and it concerns whether there is substance to the response tendency. Some researchers have maintained that desirability is a rightful component of normal personality traits, and as such, eliminating SDR from measures of desirable and undesirable traits is tantamount to eliminating valid trait variance that properly belongs in the scales (McCrae & Costa, 1983). For example, a behavior domain like depression is inherently undesirable, so what would scores on a measure of depression mean if the items were free from any desirability component? Addressing this issue in a satisfactory manner is beyond the scope of this article. We do note, however, that failure to control for SDR in a personality measure means that one can never know whether a respondent's high (low) score on a desirable (undesirable) scale is due to the person's level of the trait, his or her general tendency to describe self in desirable terms, or some combination of the two (see also Holden & Passey, 2010).

is based on SDR and impacts the construct validity of the assessments, manifest itself in the observed psychometric properties of the measuring device? Specifically, to what extent do responses to the items of a personality questionnaire have to be biased in the desirable direction to be detected by the measure's empirical validity?

Psychological scales are often validated empirically by correlating their scores with some relevant criterion variable. The prediction criterion most often represents an independent measure of the same trait, but it could be a completely different variable entirely. If a personality trait is known to be a determinant of some significant behavior or outcome, then a valid measure of that trait should predict that criterion (Hong & Paunonen, 2009; O'Connor & Paunonen, 2007; Paunonen, 1998, 2003). Such criteria with known, or strongly suspected, personality determinants might include job performance, academic achievement, marital satisfaction, health-risk behaviors, and so on.

At first glance, one might suggest that anything that impacts on a measure's construct validity, such as SDR, should impact on its criterion (or predictive) validity. Thus, response distortion in a sample of test scores due to SDR should attenuate the ability of the measure to predict a validation criterion. However, consider that if everyone in the sample falsified their responses an equal amount toward the desirable end of the trait scale from their true responses, the addition of the constant to their scale scores would not change the correlation of that scale with the criterion of interest, or with any criterion for that matter. Any problem of desirability bias for criterion prediction, therefore, must materialize when respondents in a sample engage in SDR to varying degrees (Rosse, Stecher, Miller, & Levin, 1998).

Past Research on SDR and Validity

Some researchers have studied the effect of SDR on empirical validity estimation, sometimes arriving at inconsistent conclusions. Part of the problem resides in the fact that one can never know whether and to what extent any one person in a respondent sample has trait scores that are distorted by desirability bias. Because such information is a prerequisite to determining the effects of the bias on the psychometric properties of the measure, it must be inferred. There are two general approaches to inferring the presence of SDR in a sample of test data (but see also Zickar & Drasgow, 1996; Zickar & Robie, 1999; Ziegler & Buehner, 2009). The first is to administer a general desirability scale to respondents along with the personality trait scale of interest. People are assumed to have distorted trait scale scores to the extent that they have high desirability scale scores. This then allows one to determine, for example, whether desirability bias acts as a moderator or suppressor of test–criterion correlations. The second way to infer SDR is to run a study where some of the respondents are instructed (or have a strong incentive) to "fake good" on the test, or to present the best impression possible in their item endorsements. People in that group are assumed to represent greater response distortion than those in a control group who are administered the test under normal "straight-take" instructions (or who do not have any particular incentive to fake good). The test–criterion correlations of the fake-good and straight-take groups (or job applicant and job incumbent groups, for example) can then be compared.

In general, studies that have looked at SDR as estimated by standard desirability scales have found little evidence of deleterious effects of the response bias on predictive validity. Hough, Eaton, Dunnette, Kamp, and McCloy (1990), for example, reported mean correlations between personality trait measures and job-related criteria for soldiers who scored low versus high on a social desirability scale. The differences between the two groups were, on the whole, small. Among the biggest discrepancies observed was that for a self-report measure of work orientation predicting an observer-rated criterion related to effort and leadership, with predictive validities of .25 for soldiers low on the authors' desirability scale and .20 for soldiers high on desirability (see Hough et al., 1990, Table 7). Hough (1998) reported on three large-sample studies in which incumbents in different occupational sectors were evaluated on personality–job performance correlations before and after eliminating the most egregious 5% of biased test responders, as estimated by an independent desirability scale. In general, the predictor–criterion correlations changed very little with the elimination of the biased protocols, going up slightly in Studies 1 and 3 and down slightly in Study 2.

Barrick and Mount (1996) partialed desirability scale scores from workers' personality trait scores before correlating the latter with a criterion measure of job performance. Among the biggest effects they found was a decrease in the predictive validity of a conscientiousness measure, going from .27 to .22 when controlling for a self-deception index of desirability bias (see Barrick & Mount, 1996, Table 2). (Note that if response distortion were affecting the validity of a personality assessment in the negative direction, as is normally assumed, it would be acting as a suppressor variable, and controlling its effects should raise the index of predictive validity rather than, as Barrick & Mount, 1996, have reported, lower it.) In a meta-analysis of studies related to personality and job performance, Ones, Viswesvaran, and Reiss (1996) estimated that the criterion-related validity of conscientiousness in predicting overall job performance was about .23 with no control for desirability. Remarkably, according to those authors, that value was still .23 (to two decimal places) when desirability was controlled in the personality measure (see Ones et al., 1996, Table 7). Ones et al. concluded that social desirability acts neither as suppressor, moderator, nor predictor of job-related criteria (see also McGrath et al., 2010).

More recently, Holden (2007, Study 2) reported a linear decrease in personality inventory validity (self–peer correlations) as a function of increasing scores for respondents on Paulhus's (1991) Balanced Inventory of Desirable Responding Impression Management scale. That decrease, however, was small overall. The mean validity of measures of the Big Five personality factors was .62 at the lowest level of desirability and .50 at the highest level. Furthermore, the bias acted as a statistically significant moderator of validity for only one of the five personality measures (see also Holden, Wheeler, & Marjanovic, 2012). Similar null findings regarding the putative suppressor and moderator properties of desirability scales have been reported elsewhere (e.g., Borkenau & Ostendorf, 1992; Holden & Passey, 2010; Piedmont, McCrae, Riemann, & Angleitner, 2000).

Studies in which distinct groups or clusters of respondents are strongly expected to vary in their levels of SDR have often led to contrary results compared to those cited above in which standard

desirability scales are used to estimate the extent of the bias. Such respondent groups might represent subjects in fake-good versus straight-take experimental conditions or workers in job applicant settings versus job incumbent settings. When experimental participants are instructed to fake good on a personality questionnaire, it is clear that they can do so successfully because their mean trait scores change in predictable directions. In a meta-analysis of such studies, Viswesvaran and Ones (1999, Table 2) reported a range in effect sizes (*d*) of 0.47 to 0.93 for Big Five factor scales (mean $d = 0.72$). Other studies have also amply documented the fact that job applicants have personality trait scores that are closer to the desirable ends of the scales than do job incumbents (e.g., Hough, 1998; Paunonen, Lönnqvist, Verkasalo, Leikas, & Nissinen, 2006). In one such comparison, Rosse et al. (1998, Table 2) reported effect sizes ranging from 0.13 to 1.16 for several Big Five facet scales (mean $d = 0.65$). The question of interest, of course, is, are the predictive validities of those scale scores different in the different groups?

Two studies clearly reveal the dramatic impact on test validity that can follow from instructions to deliberately fake a personality questionnaire. Holden (2007, Study 3) asked students in a university residence to fake good on an extraversion measure and compared their results with an equivalent group given straight-take instructions. The correlation between the students' self-ratings and their roommates' ratings of them dropped from .54 in the straight-take condition to .11 in the fake-good condition. Jackson, Wroblewski, and Ashton (2000) used a dependability scale to predict self-reported counterproductive work behaviors, the personality scale being completed under a straight-take condition and under a "make a good impression" response set. The correlation of the dependability trait scores with workplace delinquency was −.48 in the straight-take group but only −.18 in the fake-good group (see Jackson et al., 2000, Table 1).

Overview of the Present Study

We sought in this study to provide some resolution to the debate about the extent to which desirability bias in responses to the items of self-report personality (or other) questionnaires affects or moderates the observed correlations of those measures with external criteria (i.e., criterion validity). Our approach was not to estimate or manipulate SDR in the item endorsements of real questionnaire respondents. Instead, we simulated such data with a Monte Carlo procedure that can be summarized in three steps. First, we randomly sampled data from a given population distribution representing respondents' scores on a bipolar personality trait having relatively desirable and undesirable poles (e.g., honesty–dishonesty) and having a known criterion validity. Next, we added desirability to the trait scores, in varying amounts, by elevating some of the respondents so that they would be closer to the desirable end of the personality dimension. Then, we correlated both the desirability-free and desirability-saturated scores with criterion scores to assess validity. Of interest was the degree to which desirability bias in the simulated personality test scores compromised the ability of that test to predict relevant criteria.

The present Monte Carlo validity comparison study can be construed as simulating an assessment context in which respondents complete a personality questionnaire under two conditions. In one condition, the motivation to respond desirably is low (e.g., anonymous responses), and in the other condition, that motivation is high (e.g., nonanonymous responses). The challenging part of designing such a simulation study was in deciding how to model desirability bias in the latter case. For example, should every respondent in the sample be applied an equal amount of desirability bias, or should some receive more than others? If the latter, which respondents should get more desirability, and which should get less? Our solution to this problem was to evaluate several plausible models of desirability responding.

We should mention at this point that there have been other studies in this area that have used Monte Carlo methods to evaluate SDR effects on criterion validity (Berry & Sackett, 2009; Converse, Peterson, & Griffith, 2009; Komar, Brown, Komar, & Robie, 2008; Marcus, 2006). There are important differences between our study and these other studies, however. First, those other authors selected subsets of respondents from their computer-generated samples, based on the respondents' ranked (and distorted) trait scores, and evaluated criterion validity for those people only, simulating a personnel selection situation. In contrast, we evaluated validity in the full sample of respondents in the present research. Thus, we were simulating a broader context in which, for instance, the predictive validity of a personality measure is estimated for a general population of respondents. Another important difference is that those other studies evaluated desirable responding under the assumption that everyone in a sample who is faking has the same expected level of SDR. As described below, we evaluated that model too, but more complicated models as well. Finally, in our study, we formally evaluated SDR as a statistical moderator (and, less formally, as a suppressor) of test–criterion correlations, something that has not been done before in a Monte Carlo context.

Method

Data Generation

The first step in our simulation study was to generate two columns of numbers of length *n*—one column representing *n* respondents' personality true scores, *X,* and the other representing their criterion scores, *Y*. These columns were constructed by randomly sampling *n* observations, from two independent populations of $N(0, 1)$ having a given correlation, p_{XY}, which is the population validity of test *X* for predicting criterion *Y*. The *X* and *Y* scores were then algebraically transformed to represent values on a 9-point rating scale (1.0–9.0) with a mean of 5.0 and a standard deviation of 2.0. We applied this transformation for the following reasons: (a) The resultant personality measurement scale mimics the well-known stanine scale (Angoff, 1971, p. 519) used by the U.S. military to measure human attributes; (b) our personality test scores are consistent with (averaged) Likert-type rating scale scores (9-point, 7-point, etc.) that are commonly used in personality assessment; and (c) as shown below, the 9-point scale gave us some basis for choosing the

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

amount of desirability bias we eventually introduced into our simulated personality scores.[3] Fn3

Modeling Strategy

Having generated our simulated personality test scores *X* and criterion scores *Y,* we correlated those two variables to determine the observed validity of *X* in predicting *Y*. That correlation, r_{XY}, which should approximate the population value we fixed as one of our simulation parameters, p_{XY}, is the validity of a personality measure that is free of desirability bias. Next, we added a measured amount of desirability *D* into the *X* scores, creating *X**, and computed the correlation r_{X^*Y}, noting any change in the test's criterion validity. A prerequisite for this procedure, of course, is some model of SDR to be used as the basis for deciding on desirability amounts to use in transforming *X* into *X**.

We derived several linear and nonlinear models of desirability responding that were simulated in our study. Because all of them led us to essentially the same conclusions, we present only two linear models in the results that follow. (Other models are reported on briefly in the Discussion section.) We start by listing some fundamental assumptions about SDR that guided us in designing our models of desirable responding. We then describe the development of a baseline model, one excluding desirability bias that was used for comparison purposes, followed by the two SDR models.

Modeling Assumptions

In developing the models of SDR used as a basis for generating our simulated data, we made some assumptions about how the bias is manifested in personality questionnaire responses. These assumptions are founded on the generally accepted conceptualizations of SDR we outlined in the introduction to this article, conceptualizations that have been adequately supported by empirical data. For the sake of convenience, we present our assumptions below with reference to an arbitrary personality dimension that has a more desirable positive pole and a less desirable negative pole.[4] Fn4

Assumption 1. SDR is a bias to endorse the most favorable response option available for any particular personality item. This means that it is a systematic bias, in the sense that such item endorsements are generally predictable rather than random (e.g., Edwards, 1957; Jackson & Messick, 1962; Paulhus, 1984).

Assumption 2. SDR is an individual-difference variable that represents a continuum, whereby some people will exhibit more of the bias than others. Moreover, the SDR continuum is bipolar, meaning that some respondents can show a bias for choosing the least favorable response option for an item (Edwards, 1970; Marcus, 2006; Winder et al., 1975).

Assumption 3. The motive, drive, and opportunity to respond desirably are inversely related to the person's true standing on the trait (McFarland & Ryan, 2000). This means that people who are low on a desirable trait will be more inclined and able to engage in SDR than those who are high on the trait, in general.

Assumption 4. The difference between a person's obtained score on a desirable measure and his or her true score will generally be less at higher levels of true score. This is because (a) people already high on the positive trait will be less motivated to misrepresent their true standing on the dimension (Assumption 3) and (b) personality trait measures normally have a maximum obtained score, resulting in an operational ceiling on the size of the response bias (see Footnote 3).

Assumption 5. A moderate, but nevertheless noteworthy, effect size for SDR is a change in personality scores of 0.5 standard deviation units (Schmitt & Oswald, 2006). A large, if not extreme, effect size is a change of 1.0 standard deviation units (Berry & Sackett, 2009; Viswesvaran & Ones, 1999).

Baseline Model—No Desirability Bias

Simulation of test responses. For each set of parameters in our simulation (see Procedure section below), we ran reference conditions that did not include desirability bias. We did this by first generating our personality (*X*) and criterion (*Y*) data and evaluating their validity (r_{XY}). We then added some random error (but not desirability) to the personality score (resulting in *X**) and reevaluated criterion validity (r_{X^*Y}). We call this our Baseline Model, and it is depicted in Figure 1. F1

Figure 1 is a plot of an example simulated data set (with $n = 500$ respondents) showing the original 9-point personality scores (*X*) on the abscissa against the transformed personality scores (*X**) on the ordinate. The line at 45° represents the starting point for what we call the baseline case. It is the set of points where the error component added to the original personality scores equals zero, that is, where *X** equals *X*. Doing validity comparisons on a set of

[3] Our algebraic transformation of the generated variables *X* and *Y* onto the 9-point scale was a linear one, so it in no way affected of the results of our validity comparisons. However, we also forced any transformed values that were out of range of the 9-point scale to be within its boundaries (see also Berry & Sackett, 2009; Komar et al., 2008). Thus, numbers less than 1 or greater than 9 were changed to the respective scale endpoints. We decided on this procedure to simulate data that would best characterize real personality assessments, where respondents' total scores (or mean scores) cannot exceed the numerical limits imposed by the measurement scale. Such data truncation, of course, will generally alter the correlation of the scores with external criteria (i.e., their validity). Yet the proportion of numbers so changed in any data set was invariably small in our study (less than 5% of randomly sampled unit normal *Z* scores yielded 9-point scores that were out of range), so the effects of this aspect of our data manipulation on our results were trivial and are not reported.

[4] Throughout our study, we consider higher numbers on the generated 9-point scales to represent more of the desirable end of the bipolar trait continuum (e.g., honest), toward which there is a response bias, with lower numbers representing more of the undesirable end (e.g., dishonest). Reversing the scale, so that higher numbers are less desirable and response distortion is in the negative direction, would be a simple linear transformation and, as such, would in no way affect the results and conclusions we report in this article. Moreover, our choice of the end of the bipolar trait scale toward which there is a motivation to fake (i.e., the desirable pole) is arbitrary as far as our analyses are concerned. This means that our simulation results would apply equally to a fake-bad situation where the motivation is to respond toward the undesirable pole. Indeed, our results would even apply to a role-faking scenario (see Footnote 1) where there is a motivation to respond toward one pole (no matter which pole) of an evaluatively neutral dimension. These different situations would simply require a minor reconceptualization of our assumptions, whereby the term *desirable trait* would refer to any trait for which there is a motivation to have a high standing instead of a low standing (or vice versa), regardless of the trait's intrinsic virtues.

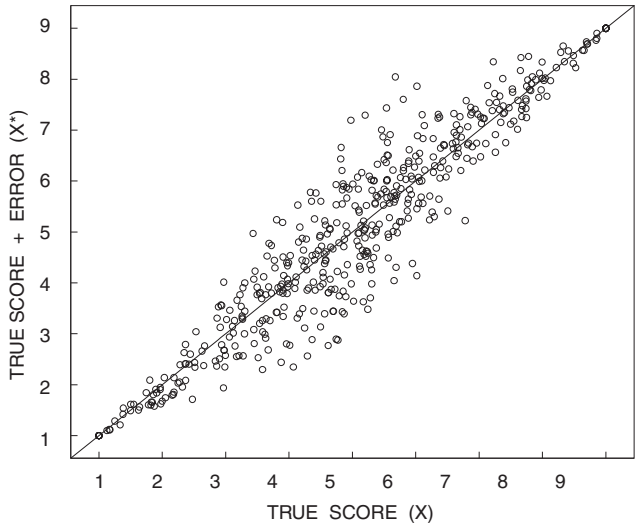

Figure 1. Example data set showing 500 respondents' simulated personality scores before (*X*) and after (*X**) adding random error (Baseline Model).

data where the modified personality test data *X** are identical to the original data *X* would, of course, be uninformative—the validity coefficients r_{XY} and r_{X^*Y} would be identical. So, to make our baseline simulation more realistic, we added a component of random error *E* in creating the modified scores *X**. The random errors we added to our simulated respondents' scores in our baseline conditions are represented by the points scattered about the line in Figure 1 (see Appendix A). As is apparent, we added the most error to trait scores at the midpoint of the 9-point scale (5), with decreasing error added as those scores approached the scale endpoints (1 or 9). (We also included truncation with transformed values that fell out of range; see Footnote 3.)

Parallels to real test data. The data we generated for our Baseline Model can be considered to represent a typical test–retest assessment situation. For example, a situation where (a) respondents complete the same personality inventory twice, (b) any motivation to respond desirably is the same on both occasions, and (c) normal random measurement error causes test scores to differ across assessments. A change in test validity under such cross-validation conditions, specifically when comparing r_{XY} to r_{X^*Y}, would then be due to test unreliability rather than to desirability bias, which is why we refer to it as the Baseline Model in this study.

We believe that our decision to add more error to simulated retest scores near the middle of the trait distribution, with less error at the two extremes, is realistic for two reasons. First, it models the finding that respondents at the extremes of a bipolar trait dimension are more consistent in their trait behaviors and in their scores on corresponding trait measures than are respondents in the middle of the dimension (Bem & Allen, 1974; Kenrick & Stringfield, 1980; Paunonen, 1988; Paunonen & Jackson, 1985). Second, less error variation observed at the extremes of the trait continuum conforms to the floor and ceiling effects that would necessarily characterize personality data representing a bounded 9-point scale.

Model 1—Moderate Desirability Bias

Simulation of test responses. Figure 2 represents the first model of desirability responding that we applied to our simulation data, called Model 1. The graph shows the original personality scores (*X*) on the abscissa plotted against the personality scores with desirability bias *D* added (*X**) on the ordinate. As noted with regard to the Baseline Model shown in Figure 1, the reference line at 45° represents the case in which each person's score with desirability is the same as his or her score without desirability (i.e., $D = 0$). The line above the reference line in Figure 2 shows the model of desirability responding we applied in the case of Model 1. As is apparent, and based on our modeling Assumption 3, more desirability was added to those respondents at lower levels of the trait than at higher levels. F2

The data points in Figure 2 indicate that, rather than modeling desirability bias by simply adding a constant to each person's original personality score at any given level of trait *X,* we added an average amount with some variation (Assumptions 1 and 2). For instance, respondents with an original personality test score of 1 (i.e., $X = 1$) were raised, on average, by 2 points toward the positive (desirable) end of the trait scale (i.e., $X^* = 3$). However, some people were raised more and some less, as indicated by the leftmost scatter of points shown in the figure. As the original personality score *X* increased, less desirability was added to each respondent, on average, and the variation in desirability values was smaller (see Appendix B). (As already mentioned in Footnote 3, any transformed values *X** beyond the 9-point scale maximum were truncated.) Subsequent testing confirmed that this level of desirability bias manipulation corresponded to an effect size of approximately 0.5, which we postulated in Assumption 5 to be of moderate size.

Parallels to real test data. Our simulation of desirability bias as depicted in Figure 2 was thought to be a plausible representation of real assessment situations. First, as stated earlier in this section,

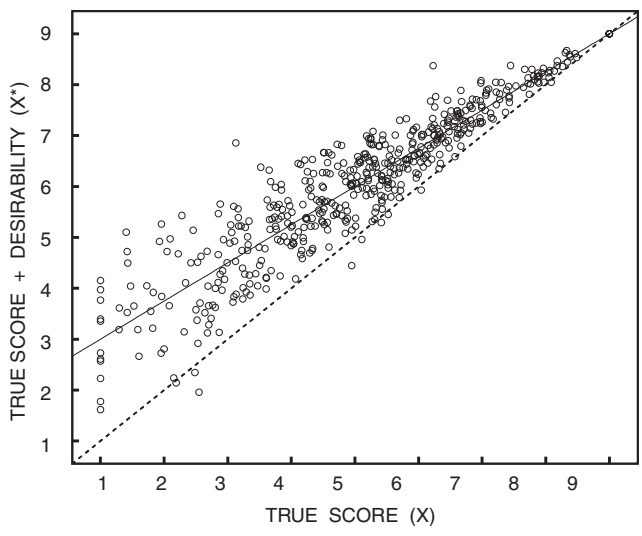

Figure 2. Example data set showing 500 respondents' simulated personality scores before (*X*) and after (*X**) adding moderate desirability bias (Model 1).

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

it is those who are at the undesirable end of a bipolar trait dimension who are most compelled (consciously or not) to make a good impression and who have the strongest need to engage in ego maintenance (Assumption 3). Hence, lower scores on the test should demonstrate a larger desirability component than should higher scores, on average. Second, SDR is an individual-difference variable (Assumption 2), so not everyone at the same level of trait will engage in the bias to the same degree. It is this diversity that causes the changes in rank ordering of the respondents when comparing their original trait scores X to their transformed trait scores X^* and any commensurate changes to test validity when those scores are correlated with a criterion Y. Third, the decreasing variability in desirability scores as trait level increases and the truncation in the transformed scores at the highest trait levels are consistent with what one would expect in real data on bounded assessment scales (Assumption 4). That is, people already high on a desirable trait do not need to embellish their self-descriptions so much to obtain the most favorable personality score, and in any case, operational constraints mean that numbers greater than the measurement scale's endpoint (9 in this case) are not possible, no matter how much the respondent might wish to elevate him- or herself on the personality dimension.[5] Fn5

Model 2—High Desirability Bias

Simulation of test responses. We also ran a series of simulations in which, compared to our moderate desirability conditions of Model 1, we added significantly more bias to our personality test scores. Our model of high desirability bias in this case, called Model 2, is represented in Figure 3 (see also Appendix B). As in the previous figures, the graph shows the original personality test scores on the abscissa (X) plotted against the personality scores with desirability D added on the ordinate (X^*). F3

Comparing our moderate desirability Model 1, shown in Figure 2, with our high desirability Model 2, shown in Figure 3, we see a much more dramatic effect of the response bias on the personality scores in the latter case. For example, respondents with true scores of 1 on the trait (i.e., $X = 1$) had, on average, biased scores of 5 on the trait (i.e., $X^* = 5$), an increase of four units on a 9-point scale.

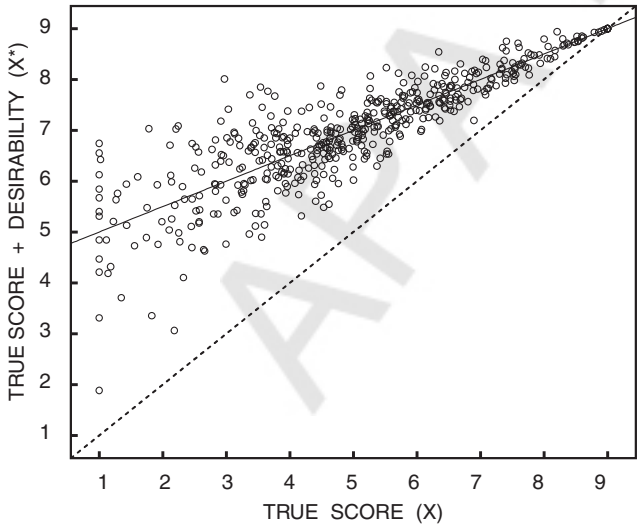

Figure 3. Example data set showing 500 respondents' simulated personality scores before (X) and after (X^*) adding high desirability bias (Model 2).

Parallels to real test data. Our Model 2 simulations were functionally identical to those of Model 1, described in the previous section, so the same correspondences between the simulation data and real data are presumed to apply. However, in Model 2, much more desirability bias was added to the simulated test scores. Subsequent testing confirmed that our high desirability bias manipulation in Model 2 corresponded to an effect size of approximately 1.0, consistent with Assumption 5. Such a huge change in test scores should seriously compromise the construct validity of any assessment instrument. The question our study sought to answer was, to what extent would that change compromise criterion validity compared to Model 1 and, especially, to the Baseline Model?

Procedure

Using IMSL (1987) routine RNMVN, we randomly sampled n personality true scores X and n criterion scores Y from independent populations having a predetermined correlation p_{XY} (i.e., criterion validity). We decided on three sample sizes for our evaluations, representing small, medium, and large validation studies: $n = 90$, 180, and 270. We also decided on three personality–criterion correlations representing small, medium, and large levels of criterion validity for the personality measure in the population: $p_{XY} =$ .20, .40, and .60.[6] This latter range of values might represent, for example, the estimated validity of conscientiousness in predicting job performance ($r_{XY} = .23$; Ones et al., 1996) to the observed validity of self-reports against peer reports on Big Five personality measures ($r_{XY} = .62$; Holden, 2007). Fn6

After generating X and Y columns of scores and correlating them to determine the observed validity of the personality true scores in the sample (r_{XY}), an amount of desirability bias D (according to Model 1 or Model 2) or just random error E (according to the Baseline Model) was added to the trait scores X. The validity of the modified X^* scores (r_{X^*Y}) was then determined and compared with the validity of the original X scores (r_{XY}). This procedure of generating X, Y, and X^* variables and computing validity coefficients was repeated 5,000 times for each simulation condition.

[5] One should not confuse the negative correlation between SDR and trait level, referred to throughout this article, with the positive correlation normally found between a desirability scale and a desirable trait scale. A score on an independent desirability scale indicates a general tendency to engage in SDR, regardless of the characteristic being assessed. In our case, however, a desirability score represents the amount of bias in an observed score on a particular personality trait measure, where less bias is likely to be manifested by someone who is already at the higher trait levels. Thus, someone who is near the ceiling of a (desirable) trait would necessarily have a low index of SDR in this study, but he or she could easily have a high index of SDR on an independent desirability scale.

[6] We calculated the power to detect significant validities for each of our simulation conditions (Cohen, 1992). With three exceptions, power was estimated to be .99 or greater (at $\alpha = .05$, one-tailed). The exceptions were those involving the smallest population validity of $p_{XY} = .20$, where power was computed as .95 for $n = 270$, .86 for $n = 180$, and .64 for $n = 90$. AQ: 11

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

Results

We present the results of our Baseline Model simulations first, in which we estimate the effects of random error on a personality test's coefficient of criterion validity. We then describe the effects on validity of adding moderate amounts of desirability bias to the simulated personality test scores, followed by the results of adding high amounts of desirability bias. This is followed by an evaluation of the effects on validity due to desirability bias relative to the effects due to random error.

Baseline Model: Effect of Random Error on Criterion Validity

T1,AQ:2 Table 1 shows the results of our analysis of the Baseline Model conditions, where personality test validity was determined before and after adding random error (vs. systematic error associated with desirability bias) to the simulated personality test scores. First, note that the Monte Carlo data-generation procedure did an excellent job in every case of producing personality test scores (X) and criterion scores (Y) that approximated the value p_{XY} fixed for the theoretical population validity. As seen in the upper half of the table, for the p_{XY} values of .20, .40, and .60, the generated data produced r_{XY} values of .199, .395, and .594, respectively, averaged over the different sample-size conditions. The slight underestimates in the observed validities (mean difference = .004) are probably due to the truncations we applied to a small number of extreme values in the generated data to fit them within the boundaries of the 9-point rating scale (see Footnote 3).

What happened to the r_{XY} validity correlations we reported in the paragraph above when random errors were added to the personality test scores X, producing X^*? The relevant r_{X^*Y} correlations are shown in the lower half of Table 1. Inspection of those correlations indicates that the amount of error we added to the personality scores did not much change test validity in any of the conditions. The average decrease over all conditions is only about .022. Note, however, that the two parameters we manipulated in our Monte Carlo design seemed to have different (albeit minor) effects. Averaged over the three population validity conditions p_{XY}, the validity decrease observed as a function of n was essentially constant, being about .022 for all three sample sizes. For the individual p_{XY} conditions, however, the effect of random error on observed validity was greater to the extent that the population validity was large—the decreases, averaged over the three sample-size conditions, for p_{XY} = .20, .40, and .60 were .011, .022, and .033, respectively. (As percentages, these values were about the same, corresponding to decreases in validity of 5.53%, 5.57%, and 5.56%, respectively.)

Table 1
Mean Validities for Baseline Conditions Before and After Adding Random Errors to Test Scores, as a Function of Population Validity (p_{XY}) and Sample Size (n)

		p_{XY}			
	n	.20	.40	.60	M
Before adding error	90	.200	.394	.593	.396
	180	.200	.396	.594	.397
	270	.197	.395	.595	.396
	M	.199	.395	.594	.396
After adding random error	90	.189	.372	.560	.374
	180	.189	.374	.561	.375
	270	.187	.372	.562	.374
	M	.188	.373	.561	.374

Note. Each mean is based on 5,000 simulated data sets.

Model 1: Effect of Moderate Desirability Bias on Criterion Validity

The results of our Model 1 simulations, in which we added a moderate amount of desirability (effect size d = 0.5) to the personality test scores (see Figure 2), are shown in Table 2. The T2,AQ:3 top half of the table lists the observed test validities r_{XY} when no bias (or additional error of any kind) was added to the scores and, as such, should replicate the values already reported in the top half of Table 1. As we expected from those earlier results, these r_{XY} validities turned out to be very close to the population p_{XY} values, testifying to the accuracy of the data-generation algorithm. The numbers are also very close to those of the independent runs already reported in the top half of Table 1, testifying to the algorithm's reliability.

The bottom half of Table 2 lists our validity results when the simulated personality test scores were infused with a moderate amount of desirability bias. The inclusion of that systematic bias caused the validities of the transformed scores (r_{X^*Y}) to be lower in each condition than the corresponding validities of the original scores without the bias (r_{XY}). However, the validity decrements were small, in general, being only .027 on average. There was a complete lack of effect for sample size, where the validity decrement due to desirability bias was .027 at each of the three levels of n. The population validity of the test, however, showed some effect, with more validity decrements occurring at the higher values of that parameter—for p_{XY} = .20, .40, and .60, the respective decreases in validity were .013, .026, and .040.

Model 2: Effect of High Desirability Bias on Criterion Validity

Our next set of analyses were based on Model 2, where we added a substantial amount of desirability bias (effect size d = 1.0) to our personality test scores (see Figure 3). The results of those analyses are shown in Table 3. Compared to the moderate desir- T3,AQ:4 ability simulation results shown in Table 2, adding more desirability bias to the data had more (negative) effect on the observed validity. Nevertheless, considering the relatively large level of desirability bias added in the Model 2 conditions, the drop in validity was not particularly striking overall. The simulated test's average r_{XY} validity of .395 across all nine conditions with no desirability bias (upper half of Table 3) dropped to an average r_{X^*Y} validity of .343 with high desirability bias (lower half of Table 3).

As already seen with the results of Model 1 (see Table 2), sample size (n) had absolutely no effect on the magnitude of the decreases in validity due to the inclusion of desirability in the test scores in Model 2. In the present high desirability context, the mean decrement across the population validity conditions was .052 for each of the three sample sizes (n = 90, 180, and 270). However, as also noticed with the previous results, the validity of

Table 2
Mean Validities for Model 1 Conditions Before and After Adding Moderate Desirability Bias to Test Scores, as a Function of Population Validity (p_{XY}) and Sample Size (n)

		p_{XY}			
	n	.20	.40	.60	*M*
Before adding desirability	90	.196	.392	.595	.394
	180	.198	.396	.594	.396
	270	.197	.396	.596	.396
	M	.197	.395	.595	.396
After adding moderate desirability	90	.184	.366	.554	.368
	180	.185	.370	.554	.370
	270	.184	.370	.556	.370
	M	.184	.369	.555	.369

Note. Each mean is based on 5,000 simulated data sets.

the test in the population (p_{XY}) did have some effect on validity decreases in these data. For p_{XY} of .20, .40, and .60, the decreases in validity due to the introduction of high desirability bias into the test scores were .027, .052, and .080, respectively. However, even the largest validity decrease in these analyses, from .594 to .514 (for $p_{XY} = .60$, averaged over sample-size conditions), probably would not be large enough to alter one's conclusions about a test's criterion predictiveness.

Desirability Effects Relative to Random Error

In the two preceding sections, we calculated each decrease in test validity by comparing criterion prediction before and after adding desirability bias to the test scores, that is, as the difference between r_{XY} and r_{X^*Y}. However, one would normally expect some amount of shrinkage in predictive validity in going from a derivation sample to a cross-validation sample even without desirability bias as a component of the latter set of test scores. We therefore decided to calculate the decreases in validity due to desirability bias, represented in Tables 2 and 3, relative to the decreases due to random error, represented in Table 1.

We recomputed the validity decrease in each desirability condition relative to the corresponding baseline condition by subtracting the latter from the former. These adjusted values, which are slightly smaller than their unadjusted counterparts, are summarized in Figure 4. The upper panel of the figure shows the mean validity decreases as a function of sample size *n,* averaged across the test's population validity. Those values reveal that the validity decrements in our simulation conditions (a) were roughly the same regardless of sample size, (b) were noticeably greater for our high desirability conditions of Model 2 than for our moderate desirability conditions of Model 1, but (c) were nonetheless not large in an absolute sense. With regard to the last point, the mean decrease across all conditions was only .018. F4

The lower panel of Figure 4 illustrates the average amount (over sample size) by which validity decreased when adding desirability bias to the test scores as a function of population test validity p_{XY}, relative to the decrease expected from the addition of simple random error. It is clear that validity decreases became more pronounced, in almost a linear fashion, with increases in population validity. Furthermore, as expected, the biggest effects occurred with Model 2, where the amount of desirability bias added to the test scores was substantially more than that added in the case of Model 1 (i.e., effect size $d = 1.0$ vs. 0.5). Despite these effects,

Table 3
Mean Validities for Model 2 Conditions Before and After Adding High Desirability Bias to Test Scores, as a Function of Population Validity (p_{XY}) and Sample Size (n)

		p_{XY}			
	n	.20	.40	.60	*M*
Before adding desirability	90	.196	.393	.594	.394
	180	.198	.396	.593	.396
	270	.197	.397	.595	.396
	M	.197	.395	.594	.395
After adding high desirability	90	.169	.342	.514	.342
	180	.171	.342	.514	.342
	270	.171	.345	.515	.344
	M	.170	.343	.514	.343

Note. Each value is based on 5,000 simulated data sets.

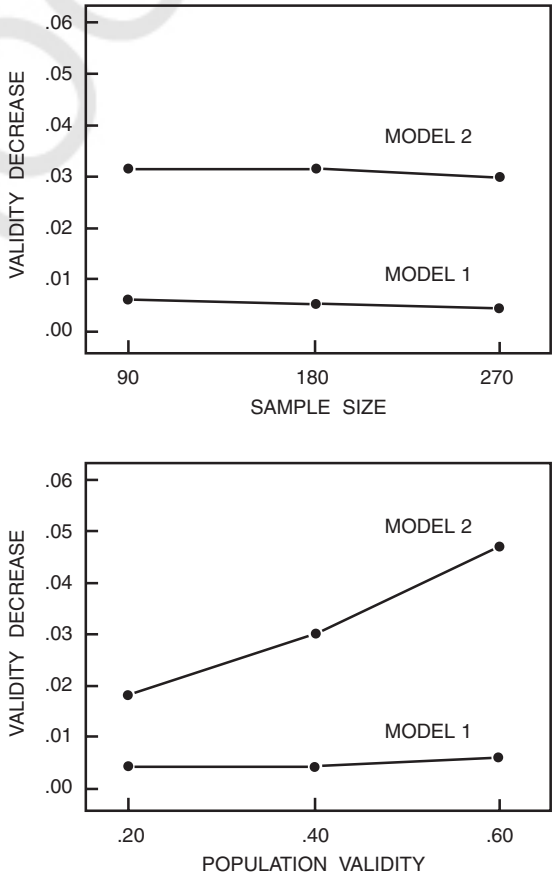

Figure 4. Mean decrease in validity for Model 1 (moderate desirability bias) and Model 2 (high desirability bias) simulations relative to the corresponding Baseline Model (random error) simulation. Results are illustrated as a function of sample size (upper figure) and population validity (lower figure), and each point is the average of 5,000 simulations in each of three conditions.

the mean decrease in validity even in the worst case scenario was not large in practical terms, amounting to about .047 for Model 2 when $p_{XY} = .60$.

Moderating Effect of Desirability on Criterion Validity

SDR has been proposed to moderate the relations between personality measures and relevant criteria. The direction of that effect, of course, is that the criterion validity of a personality measure is low to the extent that SDR is high, which seems to be a reasonable hypothesis. Such moderator effects can be tested in two ways. First, one can divide a respondent sample into two or more subgroups based on, say, a median or tertile split on the putative moderator variable, SDR in this case. Validity coefficients are then computed for each subgroup and compared. The expectation here is that validity (i.e., personality–criterion correlation) will be higher for the groups that are lower in SDR (e.g., see Holden, 2007, Study 2).

A better way to evaluate moderator effects is to use moderated multiple regression (Paunonen & Jackson, 1985), whereby criterion scores are predicted in a regression equation by the personality predictor, the moderator, and a personality by moderator product term. The product term carries the interaction of, in the present case, personality by SDR in their effects on the criterion (Cohen, 1978), and that term can be tested for statistical significance. The advantage of this multivariate approach to evaluating moderator effects over the subgroups analysis described in the previous paragraph is that the moderator variable is rightfully considered a continuous variable with the former method. Also, all subjects' data are included in the evaluation of moderator effects with the regression approach (rather than subsets of data), resulting in greater power for the relevant statistical tests. The empirical Type I error rate of moderated multiple regression analysis has been evaluated in a Monte Carlo study and found to be close to the corresponding nominal rate at $\alpha = .05$ (Paunonen & Jackson, 1988).

AQ: 5

We applied the technique of moderated multiple regression to each of the desirability-added simulations reported in the present study. Specifically, we evaluated SDR as a moderator of the relations between our respondents' personality scores and their criterion scores. The corresponding results are shown in Table 4. As is evident from that table, essentially null moderator effects were found when we added moderate desirability to the personality scores (see the top half of Table 4). The number of significant moderator effects for SDR in those simulations approximated the number expected by chance under the null hypothesis. The average number of moderator effects was 4.7% at $\alpha = .05$ (range = 4.1% to 5.3%) and 0.9% at $\alpha = .01$ (range = 0.6% to 1.1%). Surprisingly, the story was still the same when we added high desirability to the personality scores (see the bottom half of Table 4). The number of significant moderator effects averaged 4.5% at $\alpha = .05$ (range = 3.0% to 5.1%) and 0.8% (range = 0.5% to 1.1%) at $\alpha = .01$.

T4,AQ:6

Discussion

In this study, we simulated SDR to the items of a personality inventory by adding specified amounts of desirability bias to the computer-generated scores of hypothetical test respondents. We then computed the validity of the personality scores both with and without contamination by the bias. In comparing those validities, one finding became salient. Despite large components of desirability bias in some scores, the response distortion might not be obvious in the changes it produces in the measure's ability to predict relevant criteria. Consider our worst case scenario, where even those conditions might suggest that little is amiss with the personality measure. Referring back to Tables 1–3, one might expect (in round numbers) a test with an observed criterion validity of roughly .60 to produce a cross-validated coefficient of about .56 with no desirability bias in the test's scores (see Table 1), .55 with moderate desirability bias in the test's scores (see Table 2), or .51 under conditions of extreme social desirability responding (see Table 3).

Our results are consistent with other studies that have found relatively minor effects of SDR and corrections for SDR (Hough, 1998) on coefficients of empirical validity. Unlike some researchers, however, we do not automatically conclude that desirability bias is therefore not an issue for personality inventories or other measures of typical performance (e.g., Ones et al., 1996). We reiterate that response distortion due to SDR can profoundly compromise the construct validity of the assessment because the obtained scores for some of our respondents on the simulated measure departed substantially from their true scores. In Figure 3, for example, SDR caused dramatic changes in certain scores, amounting to as much as 5.7 points on a 9-point measurement scale. This

Table 4
Proportion of Simulations Showing Significant ($\alpha = .05/.01$) Moderator Effects of Desirability on Test Validity

		p_{XY}			
	n	.20	.40	.60	*M*
Model 1—Moderate desirability	90	.047/.008	.049/.009	.042/.007	.046/.008
	180	.053/.011	.047/.008	.046/.008	.049/.009
	270	.048/.010	.047/.009	.041/.006	.045/.008
	M	.049/.010	.048/.009	.043/.007	.047/.009
Model 2—High desirability	90	.049/.010	.050/.011	.042/.006	.047/.009
	180	.049/.011	.050/.007	.039/.006	.046/.008
	270	.051/.010	.042/.008	.030/.005	.041/.008
	M	.050/.010	.047/.009	.037/.006	.045/.008

Note. Each value is based on 5,000 simulated data sets.

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

reminds us of Ben-Porath and Waller's (1992) distinction between scale validity and protocol validity. The latter term refers to the trustworthiness of a particular person's test protocol, or scale scores. As those authors have observed in a clinical assessment context,

> finding that a measure of a psychological construct remains valid even when some of the subjects used in establishing validity responded invalidly speaks to the strength of the scale as a measure of that particular construct, but does not change the fact that the same scale may provide invalid information for any given client. (Ben-Porath & Waller, 1992, p. 16)

Notice the problem that can arise in the present context for expectations about criterion performance for different individuals. If that criterion were, say, worker productivity, one could make serious misidentifications of employees who are expected to do well at their jobs (see also Hough, 1998; Marcus, 2006; Rosse et al., 1998). To illustrate, the leftmost column of data points in Figure 3 reveals a respondent with a true trait score of 1.0 who received an obtained score of almost 7.0 by virtue of endorsing the most socially desirable response options for the test items. Furthermore, the high positive correlation observed between the trait and the criterion (r_{X^*Y} = .54 for those data) implies that high scorers on the personality measure should be more productive, statistically speaking, than low scorers. Yet the respondent in question only apparently belongs to the high trait and high productivity group. That person, in fact, is at the lowest level of trait and so should engender the lowest level of expected performance. (We ignore for the purposes of this argument those situations in which high SDR might be associated with better job performance; see Hogan, Hogan, & Roberts, 1996.) We do acknowledge that, at the level of the group, mean performance of workers might not be substantially impacted by retaining some individuals who are high in SDR, as demonstrated in a simulation study by Schmitt and Oswald (2006). Nevertheless, even one grossly invalid test protocol could have nontrivial consequences in some work settings.

We might ask at this point, why the meager effect of SDR on criterion validity coefficients, despite large changes to some people's test scores due to the response bias? Perhaps, as some have noted, certain statistical methods simply do not adequately reflect the degree or type of response distortion represented by SDR. Rosse et al. (1998), for example, concluded that "correlational analysis may be insensitive to changes in the rank ordering" of respondents due to the bias (p. 636). Alliger and Dwight (2000) stated that "the overall criterion-related validity coefficient is not an appropriate index of the effect of faking" (p. 61). We address this issue of test score changes and predictor–criterion correlations in a later section, when we discuss the lack of moderator effects we found for SDR on test validity.

Other Models of SDR

In the introduction to this article, we stated that we evaluated models of SDR other than those we presented formally in our Method and Results sections. We now describe some of those models briefly. We remark at this point that we found, again, surprisingly small effects on predictive validity even after adding to our simulated personality scores what we consider to be large components of response bias.

Figure 5 shows four additional models of SDR that we evaluated with simulated data. Those models are, perhaps, more appropriately viewed as two distinct models, having distinct assumptions, and each represented by two levels of added bias. Consider first the model represented in the top two panels of Figure 5 (see also Appendix B). Our Assumption 3, as outlined in the Method section, has been somewhat modified here. That original assumption, which might be questioned by some, was that there is a negative relation between trait level and SDR. Our revised assumption represents essentially a constant drive to respond desirably across the trait continuum. The top two panels of Figure 5 depict two levels of this drive factored into the personality scores X (with the upper boundary of the 9-point scale causing a necessary ceiling effect in the resultant scores X^*). F5

When we ran the simulations illustrated in the top two panels of Figure 5, we found the following results. (Because sample size had basically no effect on any of our supplementary results, as was the case with our main analyses, we do not report them.) With the smaller amounts of desirability bias (see the top-left panel of Figure 5), the mean decrease in criterion validity averaged over all the simulated conditions was .03. As with our primary analyses, there was some effect of population validity on the observed validity decrease, with mean decrements of .02, .03, and .05 at p_{XY} = .20, .40, and .60, respectively. When we added more desirability to the simulated personality data (see the top-right panel of Figure 5), the decremental effects were larger overall, but not by much—the mean decrease was .08, with individual values of .04, .08, and .11 at p_{XY} = .20, .40, and .60, respectively.

In the two bottom panels of Figure 5, we illustrate another SDR model we simulated in our study, having its own variation on Assumption 3 (see the Method section). Here, we simulated a situation where people with low levels of the trait X respond desirably to the test items. That desirability bias, however, decreases as trait level increases up to a point where the bias reaches zero and actually becomes negative (recall Assumption 2). Our simulation here (see Appendix B) is meant to represent a scenario where respondents who are at high levels of a (desirable) trait are somewhat modest in describing themselves (e.g., Tice, Butler, Muraven, & Stillwell, 1995) and, thus, tend to choose the less-than-desirable options for some test items.

Our simulations of SDR that included the presence of some respondents with a modesty bias produced results largely consistent with the results of our other analyses. When adding relatively small amounts of SDR and small amounts of modesty to the data (see the bottom-left panel of Figure 5), the overall mean decrease in predictive validity was .03. A slight effect was observed again for population validity, with decreases of .02, .03, and .05 for p_{XY} = .20, .40, and .60, respectively. Adding more bias to the personality scores (see the bottom-right panel of Figure 5), of course, exacerbated the decrements in predictive validity, with an overall decrease in validity of .11. The mean decreases for the individual levels of criterion validity in the population were .05, .11, and .16 at p_{XY} = .20, .40, and .60, respectively. As a group, these were the largest validity decreases we recorded for the many simulations of SDR we conducted in this study.

In summary, the supplementary models of bias we evaluated in this section showed decreases in predictive validity that became noticeably large only after substantial misrepresentation by respondents to items of the simulated personality trait measure (as

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

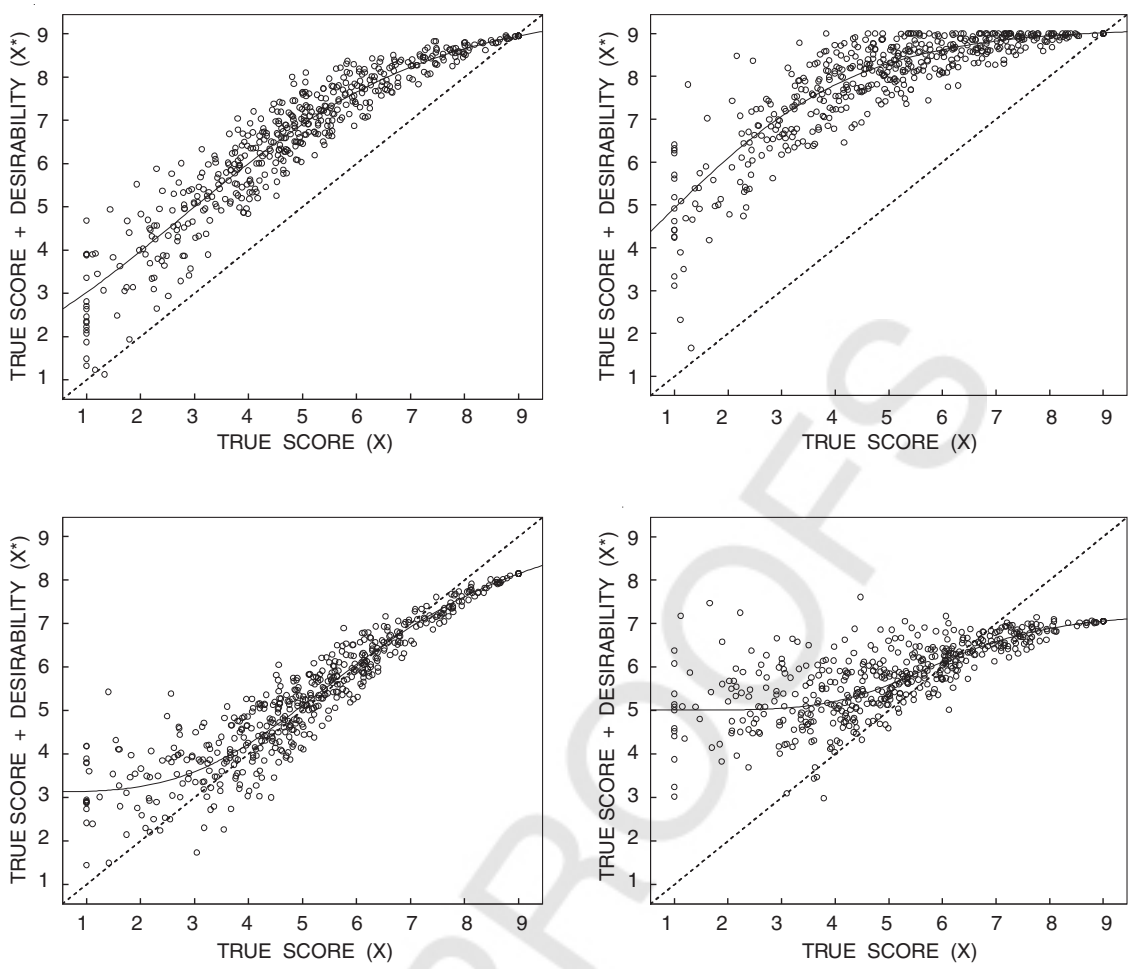

Figure 5. Example data sets for four supplementary models each showing 500 respondents' simulated personality scores before (*X*) and after (*X**) adding desirability bias (see text).

depicted in the bottom-right panel of Figure 5). We add that these decrements in predictive validity are probably somewhat overestimated because they have not been corrected for the shrinkage one would expect following normal cross-validation.

SDR as a Moderator of Test Validity

The notion has been expressed that SDR functions as a moderator variable affecting coefficients of predictive validity. The idea is that the correlation between a personality predictor and a criterion would be expected to be moderated by SDR, such that the correlation is lower when SDR is higher. As recently revealed in a meta-analysis by McGrath et al. (2010), however, empirical evaluations have shown very little support for a general validity-moderating effect for the desirability response bias (see also Ones et al., 1996). Our moderated multiple regression results shown in Table 4 concur with that negative finding.

Consistent null moderator effects. Although we did find some examples of desirability moderating the correlations between our personality predictors and their respective criteria, the rates of moderator effects in our data did not exceed those expected to arise by chance alone under the null hypothesis (i.e., no moderator effects in the population). The proportion of significant moderator effects was roughly 5% at $\alpha = .05$ and 1% at $\alpha = .01$, regardless of the number of test respondents in the sample (90, 180, or 270), the population validity of the test (.20, .40, or .60), or the amount of desirability we added to our personality scores (moderate or high). We add that the dearth of moderator effects for SDR characterized even our supplementary analyses illustrated in Figure 5, in which we changed some of the assumptions about desirability responding. When rounded to one decimal place, the mean numbers of moderator effects found for each of the four SDR models represented in that figure were identical: 4.5% at $\alpha = .05$ and 0.8% at $\alpha = .01$.

It appears that SDR biases as variously modeled in this study will generally not be detected with regression-based moderator analyses, which agrees with most other researchers' evaluations of the response bias as a moderator of test validity (see Ones et al., 1996). Furthermore, this observation generalized to subgroups analyses of moderator effects. As just one example, consider our high desirability Model 2. We reran that model with $N = 180$ respondents and at $p_{XY} = .60$ and examined the personality–criterion correlations (r_{X^*Y}) at nine distinct levels of SDR (the 180

respondents were ranked on their SDR scores and split into nine groups of 20). Those mean validity correlations, each averaged over 5,000 runs, were 0.32, 0.30, 0.31, 0.32, 0.34, 0.34, 0.35, 0.34, and 0.30, across Desirability Groups 1 through 9, respectively. This invariance in subgroup validities across the distinct levels of SDR confirms a lack of moderating effect for that variable in these data.

The absence of moderator effects in our simulated test scores, whether evaluated using moderated multiple regression analysis or subgroups analysis, might seem to be curious at first glance. Should not the groups who are low on a desirable trait and therefore high on SDR produce lower coefficients of test validity compared to those who are higher on the trait and engage in less SDR? Should not this effect be systematic, such that increasing trait scores are associated with decreasing response bias, which is then manifested as increasing trait-criterion validity (i.e., a moderator effect)? Does the lack of moderator effects in our data point to a problem with, perhaps, the application of the correlation coefficient in this context, as suggested by some (e.g., Rosse et al., 1998)?

Reason for null moderator effects. Careful deliberation about our simulation data suggests that the null moderator results we report might not, in fact, be unexpected. Let us consider Figure 3 again, which graphically depicts the effects of adding substantial desirability bias to the trait scores of 500 simulated test respondents. Compare those people lower in SDR, who are mostly higher on the trait and toward the right side of the scatterplot, with those people higher in SDR, who are mostly lower on the trait and toward the left side of the scatterplot. In looking at the original versus transformed test scores in the figure, one might suppose that the low SDR group on the right, but not the high SDR group on the left, would show roughly the same validity regardless of whether it is computed on the original trait scores without desirability bias (X) or the transformed trait scores that include the bias (X^*). We evaluated this conjecture by dividing those 500 respondents into nine roughly equal subsamples by level of SDR. The lowest SDR group had a validity of .34 for the original test scores X and a slightly weakened validity of .30 for the modified test scores X^*. For the highest desirability group, the corresponding test validities were .37 for the test scores without the bias and .33 for the test scores with the bias, essentially the same level of degradation we saw in the lowest desirability subgroup. Similar minor differences in validities were found across most of the other desirability subgroups. (For the nine SDR groups, ranging from low to high, the respective X and X^* validities were .34, .30; .29, .30; .37, .37; .19, .17; .46, .46; .24, .24; .40, .36; .42, .45; and .37, .33. Moreover, a moderated multiple regression analysis on the data of Figure 3 showed positively no SDR moderator effect, $t = -0.109$, $p > .90$.)

Is the near total absence of moderator effects in our data (and in other SDR data) surprising? We think not, with the following explanation. We have seen that adding individual differences in desirability bias to our simulated respondents' test scores changed their rank ordering somewhat on the personality measure overall (i.e., comparing their scores on X to X^*), causing some small drop in criterion correlation (i.e., comparing r_{XY} to r_{X^*Y}) when computed across the entire respondent sample, as illustrated in Tables 2 and 3. Yet, in the preceding paragraph, we have also seen that this effect of our data transformations on changes in ranks of the simulated respondents' personality scores was essentially uniform across the SDR continuum. Put another way, there was nothing in our data transformations that would be expected to differentially and systematically change the ranks of the simulated respondents' personality scores according to their levels of desirability responding, which is the hallmark of the SDR moderator effect in question.

To simulate individual differences in desirability responding, where respondents at the same level of trait demonstrated different levels of SDR, we included some randomness in transforming our test scores from X to X^* (see Appendix B). It was that aspect of our simulation that caused the decreases in test validities following the introduction of SDR into the personality data (see Tables 2 and 3). However, although the simulated individual differences in SDR produced changes in rank ordering of respondents on X versus X^*, with commensurate changes in the validity correlations r_{XY} versus r_{X^*Y}, that effect was constant across the different levels of desirability, as confirmed by our subgroups analysis described above. In fact, there is no a priori reason to expect that this random aspect of our data transformation should introduce any moderator effects into our personality–criterion scores.[7] (For a demonstration of how pure random responding on personality test items can generate significant moderator effects in validation studies, see Holden et al., 2012.) Fn7

Our analysis here suggests that that no moderator effects of SDR on test validity should be expected in our simulation data, where in fact none was found. Note that these results conform to much of the published literature on SDR, in which moderator effects in real-world personality data are also typically not found. We interpret this consistency in findings as further support for our belief that the manner in which we simulated SDR in this study is true to life. We add here that such a lack of statistical effects does not indicate a problem with the correlation coefficient as an index of test validity or moderation. The product–moment correlation is based on differences in the rank ordering of respondents on two variables and will summarize such differences when they occur. However, in agreement with Alliger and Dwight (2000), we believe that the statistic is seriously wanting as a summary of faking and protocol validity in a group of test respondents.

SDR as a Suppressor of Test Validity

It has been claimed that SDR can act as a suppressor variable in personality assessment contexts such as those we have simulated in this study. That is, the response bias represents error variance that has the effect of lowering or suppressing the validity of the personality measure in predicting the criterion of interest. Yet the recent meta-analysis by McGrath et al. (2010) of published em-

[7] We did, of course, differentially and systematically vary the desirability bias added to our simulated test scores in two ways (see equations in Appendix B). First, smaller desirability components, on average, were added as test scores increased. Second, the variability of those desirability components decreased as test scores increased. Neither of these two aspects of our data-generation procedures would be expected to have any differential effect on the rank ordering of respondents when comparing their test scores without desirability bias (X) to their test scores with desirability bias (X^*). Therefore, these (linear) transformations, by themselves, will neither alter overall test validity (i.e., comparing r_{XY} to r_{X^*Y}) nor alter conditional test validity (i.e., validity as a function of some putative moderator variable such as desirability).

pirical studies did not support the validity-suppressing influence of the desirability response bias in general (see also Borkenau & Ostendorf, 1992; Ones et al., 1996). As we demonstrate below, however, data from the present study suggest that SDR suppressor effects in personality scores of the type we have simulated can probably be dismissed a priori.

A suppressor variable has three notable properties: It is correlated with the predictor, it is not correlated with the criterion, yet it (somewhat paradoxically) enhances criterion prediction. Despite the zero criterion correlation, a suppressor variable is able to increment prediction by controlling for unwanted criterion-irrelevant variance in the primary (personality) predictor (Wiggins, 1973, p. 31). One way to evaluate a putative suppressor variable is to use it as a covariate in a partial correlation analysis. One simply computes the partial (or part) correlation between predictor and criterion after removing the effects of the variable in question (e.g., see McCrae & Costa, 1983). If the covariate is acting as a suppressor, the partial correlation will be larger in size than the original unpartialed value. (Suppressor effects can also be evaluated using multiple regression analysis; see Wiggins, 1973, p. 32.)

From a rational point of view, it seems reasonable to expect that a source of error variance such as SDR should serve to lower the criterion validity of a personality measure. This means that statistically controlling for its effects should yield higher coefficients of validity. From an equally rational point of view, however, SDR should generally fail as a suppressor. The reason is that the response bias will very likely lack one important characteristic of suppressor variables—zero criterion correlation.

To demonstrate our point here, consider once more the example of 500 simulated respondents' personality scores shown in Figure 3. As stated earlier, the correlation of the criterion scores with the personality scores containing SDR is .54 in that group of respondents (for $p_{XY} = .60$). Not mentioned, however, was the correlation of the personality scores with the SDR scores, which is −.50 for those data. Notice that these two correlations, the criterion–personality correlation of .54 and the personality–SDR correlation of −.50, considered together, suggest that there should also be some (negative) criterion–SDR correlation in this triumvirate. Indeed, this was the case. As illustrated in Figure 6, the correlation between the criterion and SDR is −.58 for the 500 respondents.

The nonzero correlation of −.58 between the criterion and SDR in Figure 6 means that SDR cannot function as a statistical suppressor of the personality–criterion relation, by definition. Therefore, SDR cannot improve the predictor–criterion correlation if statistically controlled. To verify our claim, the predictor–criterion correlation for those data is .54 with SDR in the personality scores, but it is .36 when SDR is partialed from the predictor and the criterion (or .29 if partialed from the predictor alone). The .18 drop in correlation after partialing SDR precludes any validity-suppressing function for that variable in this data set. We found similar decreases in correlations, rather than increases, after partialing SDR in our other simulation conditions as well.

Our conclusion here is that suppressor variables are unlikely to be found in assessment contexts like those we have modeled in this study because of the likely pattern of correlations that will characterize the relevant variables. (a) People who are low on the (desirable) trait will generally have higher components of SDR in their personality scores, causing a negative correlation between SDR and the personality predictor. (b) Those same people are likely to be low on the criterion, assuming a positive validity for the personality measure predicting that criterion. (c) This scenario implies a negative correlation between SDR and the criterion. (d) That nonzero correlation then precludes any suppressor effect for SDR on the personality–criterion correlation. In such situations, statistically removing the effects of SDR from the predictor and the criterion will lower the coefficient of validity rather than raise it, which is exactly what we found in our simulation data.

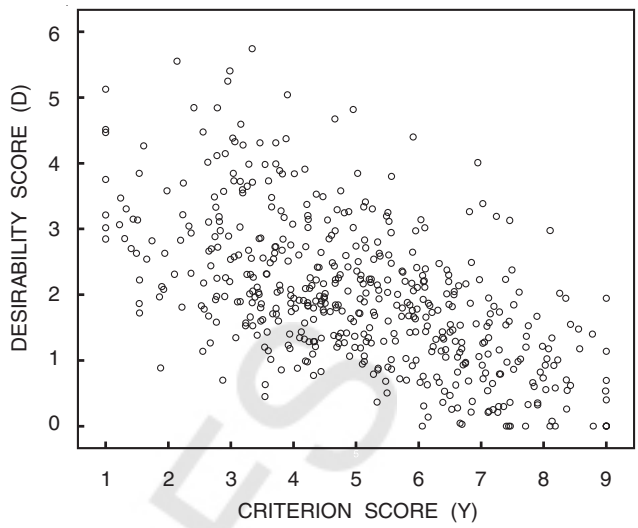

Figure 6. Scattergram illustrating the correlation between desirability scores (*D*) and criterion scores (*Y*) for the example data set of Figure 3.

Conclusions

Our results indicate that SDR's effects on personality test validity can be elusive. Following normal respondent-sampling procedures, only under the most extreme and unusual levels of distorted self-reports will the observed criterion validity of a personality measure be dramatically affected by SDR. Moreover, SDR will normally fail to show itself statistically as either a moderator of test validity or a suppressor of test validity. A careful review of relevant behavioral and methodological processes, however, led us to what some might consider a surprising assertion—that SDR's relative lack of effects on typical coefficients of validity, moderation, and suppression should not be unexpected. We reasoned this from the way in which the response bias is likely to operate to change a person's test scores, as we simulated in this study, and the way in which such changes in test scores can influence relevant validity statistics, as we observed in our analyses.

Our results and conclusions concerning the largely null effects of SDR on validity coefficients are at variance with those of some (but not all) researchers. A point that is relevant to explaining some of those discrepancies concerns the accuracy of one's indicator of SDR. In this study, we knew exactly the amount of desirability in any respondent's personality score because that was a variable we manipulated experimentally and individually. In many other studies of the response bias, however, SDR is inferred from an independent measure of desirability, one that might be only remotely

related to respondents' levels of misrepresentation on a separate personality measure (see Footnote 5). It would be easy to imagine such a detached measure of bias having statistical properties different than those we simulated in our assessments, leading to different conclusions about SDR's effects on test validity.

Despite the trivial effects we and others have reported for SDR on criterion prediction, it is important to remember that how well personality test scores are able to predict criterion scores is only one aspect of a test's validity. Another, more critical aspect is the accuracy of the obtained scores vis-à-vis respondents' true scores on the trait, which is the essence of construct validity (Messick, 1989). Perhaps a personality measure replete with desirability response bias can still function as an effective predictor of a relevant criterion. Nevertheless, the distorted scores on such a measure can seriously misrepresent some respondents' trait levels, leading to erroneous expectations about individual levels of criterion performance. This is a strong argument for minimizing desirability bias in personality inventories and other typical performance measures.

References

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60,* 59–72. doi:10.1177/00131640021970367

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personal constructs. *Journal of Applied Psychology, 81,* 261–272. doi:10.1037/0021-9010.81.3.261

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506–520. doi:10.1037/h0037130

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4,* 14–19. doi:10.1037/1040-3590.4.1.14

Bernreuter, R. G. (1933). Validity of the Personality Inventory. *Personnel Journal, 11,* 383–386.

Berry, C. M., & Sackett, P. R. (2009). Faking in personnel selection: Tradeoffs in performance versus fairness resulting from two cut-score strategies. *Personnel Psychology, 62,* 833–863. doi:10.1111/j.1744-6570.2009.01159.x

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6,* 199–214. doi:10.1002/per.2410060303

Converse, P. D., Peterson, M. H., & Griffith, R. L. (2009). Faking on personality measures: Implications for selection involving multiple predictors. *International Journal of Selection and Assessment, 17,* 47–60. doi:10.1111/j.1468-2389.2009.00450.x

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Oxford, England: Harper.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York, NY: Dryden.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories.* New York, NY: Holt, Rinehart, & Winston.

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

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportment, 39,* 184–201. doi:10.1037/cjbs2007015

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49,* 446–450. doi:10.1016/j.paid.2010.04.015

Holden, R. R., Wheeler, S., & Marjanovic, Z. (2012). When does random responding distort self-report personality assessment? An example with the NEO PI-R. *Personality and Individual Differences, 52,* 15–20. doi:10.1016/j.paid.2011.08.021

Hong, R. Y., & Paunonen, S. V. (2009). Personality traits and health-risk behaviors in university students. *European Journal of Personality, 23,* 675–696. doi:10.1002/per.736

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11,* 209–244.

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. doi:10.1037/0021-9010.75.5.581

IMSL. (1987). *International mathematical and statistical libraries* (10th ed.). Houston, TX: Author.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78,* 229–248. doi:10.1037/h0030852

Jackson, D. N., & Messick, S. (1962). Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition* (pp. 129–155). New York, NY: Wiley.

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. doi:10.1207/S15327043HUP1304_3

Kenrick, D. T., & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87,* 88–104. doi:10.1037/0033-295X.87.1.88

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

Kroger, R. O. (1967). Effects of role demands and test-cue properties upon personality test performance. *Journal of Consulting Psychology, 31,* 304–312. doi:10.1037/h0024657

Kroger, R. O., & Turnbull, W. (1970). Effects of role demands and test-cue properties upon personality test performance: Replication and extension. *Journal of Consulting and Clinical Psychology, 35,* 381–387. doi:10.1037/h0030263

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

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888. doi:10.1037/0022-006X.51.6.882

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812–821. doi:10.1037/0021-9010.85.5.812

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136,* 450–470. doi:10.1037/a0019216

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. doi:10.1037/h0053634

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Montag, I., & Comrey, A. L. (1982). Personality construct similarity in Israel and the United States. *Applied Psychological Measurement, 6,* 61–67. doi:10.1177/014662168200600107

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15,* 263–280. doi:10.1002/ejsp.2420150303

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43,* 971–990. doi:10.1016/j.paid.2007.03.017

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

AQ: 9 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. doi:10.1037/0021-9010.78.4.679

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609. doi:10.1037/0022-3514.46.3.598

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). New York, NY: Academic Press.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic bias in self-perceptions: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66,* 1025–1060. doi:10.1111/1467-6494.00041

Paunonen, S. V. (1988). Trait relevance and the differential predictability of behavior. *Journal of Personality, 56,* 599–619. doi:10.1111/j.1467-6494.1988.tb00904.x

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74,* 538–556. doi:10.1037/0022-3514.74.2.538

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84,* 411–422. doi:10.1037/0022-3514.84.2.411

Paunonen, S. V., & Hong, R. Y. (in press). In defense of personality traits. In P. R. Shaver & M. Mikulincer (Eds.), *Handbook of personality and social psychology.* Washington, DC: American Psychological Association.

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92,* 486–511. doi:10.1037/0033-295X.92.4.486

Paunonen, S. V., & Jackson, D. N. (1988). Type I error rates for moderated multiple regression analysis. *Journal of Applied Psychology, 73,* 569–573. doi:10.1037/0021-9010.73.3.569

Paunonen, S. V., Lönnqvist, J.-E., Verkasalo, M., Leikas, S., & Nissinen, V. (2006). Narcissism and emergent leadership in military cadets. *Leadership Quarterly, 17,* 475–486. doi:10.1016/j.leaqua.2006.06.003

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593. doi:10.1037/0022-3514.78.3.582

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

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. doi:10.1037/0021-9010.91.3.613

Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology, 69,* 1120–1138. doi:10.1037/0022-3514.69.6.1120

Vernon, P. E. (1934). The attitude of the subject in personality testing. *Journal of Applied Psychology, 18,* 165–177. doi:10.1037/h0074033

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

Winder, P., O'Dell, J. W., & Karson, S. (1975). New motivational distortion scales for the 16 PF. *Journal of Personality Assessment, 39,* 532–537.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–87. doi:10.1177/014662169602000107

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

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69,* 548–565. doi:10.1177/0013164408324469

(*Appendices follow*)

tapraid5/z2g-perpsy/z2g-perpsy/z2g00612/z2g3927d12z | xppws | S=1 | 4/6/12 | 10:50 | Art: P-2011-0727

Appendix A

Equations Underlying the Baseline Model Illustrated in Figure 1

All simulated personality trait scores for respondents ranged from 1.0 to 9.0, with higher numbers representing more desirable descriptions. To each respondent's true score in the baseline condition was added some amount of random error, but no desirability bias. The amount of error added, however, was not constant across the trait continuum. Instead, it was highest near the middle of the 9-point trait scale, decreasing linearly as trait level approached either extreme (1 or 9). These aspects of our simulation procedure are evident in the scatter of points about the line shown in Figure 1 in the main text.

An error score (E) for a respondent was determined as a random deviate from the unit normal distribution (Z), which was then scaled by trait level (X) according to the following function:

$$E = Z * (1 - 0.25 * |X - 5|).$$

Final Transformed Trait Score

The final baseline trait score for each respondent (X^*) was computed as the true score (X) on the trait plus the random error component (E). Because the difference between the true score and the final trait score was entirely due to random error, the desirability component (D) in the latter would therefore be

$$D = 0.$$

Appendix B

Equations Underlying the SDR Models Illustrated in Figures 2, 3, and 5

All simulated personality trait scores for respondents ranged from 1.0 to 9.0, with higher numbers representing more desirable descriptions. To each respondent's true score in each socially desirable responding (SDR) condition was added a constant of desirability bias, representing the modeled normative tendency to engage in SDR at that person's level of trait, and some amount of random error, representing that individual's personal deviation from the norm in SDR. In describing the derivation of these scores below, we use the following notation: X represents the original true scores on the trait, X' represents the true scores with the desirability constant added, and X^* represents the true scores with the desirability constant plus the random individual-difference component added.

In the paragraphs that follow, we make reference to the relevant figures contained in the main text of this article where appropriate.

Figure 2 (Model 1)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 2 points in the desirable direction, on average. As true scores increased, less desirability bias was added, such that SDR was zero at the highest level of trait (9.0). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the straight line in Figure 2, was linear in form:

$$X' = 2.25 + 0.75X.$$

Figure 3 (Model 2)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 4 points in the desirable direction, on average. As true scores increased, less desirability bias was added, such that SDR was zero at the highest level of trait (9.0). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the straight line in Figure 3, was linear in form:

$$X' = 4.50 + 0.50X.$$

Figure 5 (Top-Left Panel)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 2 points in the desirable direction, on average. As true scores increased, the desirability bias added was fairly constant across trait levels, but declined rapidly to zero as trait scores reached their ceiling (see the main text). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the curved line in Figure 5 (top-left panel), was logistic in form:

$$X' = 9.563/(1 + 3.336 * e^{-0.432X}).$$

(*Appendices continue*)

Figure 5 (Top-Right Panel)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 4 points in the desirable direction, on average. As true scores increased, the desirability bias added was fairly constant across trait levels, but declined rapidly to zero as trait scores reached their ceiling (see the main text). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the curved line in Figure 5 (top-right panel), was logistic in form:

$$X' = 9.119/(1 + 1.459 * e^{-0.539X}).$$

Figure 5 (Bottom-Left Panel)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 2 points in the desirable direction, on average. As true scores increased, less desirability bias was added until it passed through zero and became slightly negative at very high levels of trait (see the main text). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the curved line in Figure 5 (bottom-left panel), was sigmoidal in form:

$$X' = (3.131 * 555.1 + 9.728 * X^{3.402})/(555.1 + X^{3.402}).$$

Figure 5 (Bottom-Right Panel)

Respondents with the lowest true scores on the trait (1.0) were raised by roughly 4 points in the desirable direction, on average. As true scores increased, less desirability bias was added until it passed through zero and became moderately negative at very high levels of trait (see the main text). The function relating the true trait score (X) to the trait score with desirability bias added (X'), as illustrated by the curved line in Figure 5 (bottom-right panel), was sigmoidal in form:

$$X' = (5.011 * 24{,}344.6 + 7.280 * X^{5.606})/(24{,}344.6 + X^{5.606}).$$

Individual Differences in SDR

For each SDR model, random error was added to each respondent's desirability-modified true score so that not everyone at the same level of trait would show the same amount of desirability bias. This was intended to simulate real-world conditions where there are individual differences in the tendency to respond desirably at any given trait level. The amount of error added, however, was not constant across the trait continuum, being highest at the lowest level of trait and decreasing linearly as trait level increased. These aspects of our simulation procedure are evident in the scatter of points about the functions plotted in Figures 2, 3, and 5.

The error score (E) for each respondent was determined as a random deviate from the unit normal distribution (Z), which was then scaled by true trait level (X) according to the following function:

$$E = Z * (1.125 - 0.125X).$$

Final Transformed Trait Score

The final desirability-transformed trait score for each respondent (X^*) was computed as the person's true score (X) plus the modeled desirability constant for his or her level of trait derived from the equations for X' above (i.e., $X' - X$) plus the person's individual deviation from the desirability norm (E) at that trait level. The total component of desirability bias in a person's transformed trait score (D) would therefore be

$$D = X^* - X$$

Received April 29, 2011
Revision received January 10, 2012
Accepted March 14, 2012 ■