Psychological Bulletin
2010, Vol. 136, No. 3, 450–470


0033-2909/10/$12.00 DOI: 10.1037/a0019216

# Evidence for Response Bias as a Source of Error Variance in Applied Assessment

Robert E. McGrath and Matthew Mitchell
Fairleigh Dickinson University

Brian H. Kim
Occidental College

Leaetta Hough
The Dunnette Group, Ltd.

After 100 years of discussion, response bias remains a controversial topic in psychological measurement. The use of bias indicators in applied assessment is predicated on the assumptions that (a) response bias suppresses or moderates the criterion-related validity of substantive psychological indicators and (b) bias indicators are capable of detecting the presence of response bias. To test these assumptions, we reviewed literature comprising investigations in which bias indicators were evaluated as suppressors or moderators of the validity of other indicators. This review yielded only 41 studies across the contexts of personality assessment, workplace variables, emotional disorders, eligibility for disability, and forensic populations. In the first two contexts, there were enough studies to conclude that support for the use of bias indicators was weak. Evidence suggesting that random or careless responding may represent a biasing influence was noted, but this conclusion was based on a small set of studies. Several possible causes for failure to support the overall hypothesis were suggested, including poor validity of bias indicators, the extreme base rate of bias, and the adequacy of the criteria. In the other settings, the yield was too small to afford viable conclusions. Although the absence of a consensus could be used to justify continued use of bias indicators in such settings, false positives have their costs, including wasted effort and adverse impact. Despite many years of research, a sufficient justification for the use of bias indicators in applied settings remains elusive.

*Keywords:* response bias, suppressor variables, moderating variables, employee selection, disability evaluation

Psychologists have been trying to develop methods for the identification of inaccurate self-presentation for more than 90 years (Marston, 1917). At least since Hartshorne and May (1928) demonstrated that many of the children for whom there was strong evidence of cheating denied having done so, psychologists have been particularly concerned about the potential for responding in an invalid manner to standardized psychological measures. Initial research on this topic was largely pragmatic, focusing on practical strategies for minimizing the impact of inaccurate responding. Cronbach's (1946) analysis of response sets initiated a more conceptual discussion of biased test performance that continues to this day (e.g., Holden, 2008; McGrath, 2008). In the intervening years, dozens of psychological measures have been developed that are intended to detect inaccurate responding, and thousands of studies have been conducted on inaccurate responding and the minimization of its effects. In fact, inaccurate responding may well be the most extensively studied topic in the field of applied psychological measurement.

Despite psychologists' concern over inaccurate responding, a number of articles have been published over the years questioning the importance of response bias in psychological assessment (e.g., Block, 1965; Ones & Viswesvaran, 1998; Piedmont, McCrae, Riemann, & Angleitner, 2000; Rorer, 1965). These critiques cumulatively suggest the value of a comprehensive investigation into the degree to which psychological measures are effective at detecting inaccurate responding in real-world settings.

This article focuses on a very specific subset of the research literature that has been published concerning inaccurate responding to psychological measures, having to do with the real-world criterion-related validity of indicators intended to detect inaccurate responding. The next section provides an introduction to the topic of response bias as background to the empirical review that follows.

Robert E. McGrath and Matthew Mitchell, School of Psychology, Fairleigh Dickinson University; Brian H. Kim, Department of Psychology, Occidental College; Leaetta Hough, The Dunnette Group, Ltd., St. Paul, Minnesota.

We are grateful to Yossef Ben-Porath, Lewis Goldberg, Paul Green, Kevin Greve, Paul Lees-Haley, Jennifer Lyne, David Nichols, Martin Rohling, and Neal Schmitt for their comments on drafts of this article and/or their help identifying manuscripts for possible inclusion. The positions expressed in this article are those of the authors and should not be taken as representing those of our colleagues who provided input.

Correspondence concerning this article should be addressed to Robert E. McGrath, School of Psychology T-WH1-01, Fairleigh Dickinson University, Teaneck, NJ 07666. E-mail: mcgrath@fdu.edu

## Response Bias and its Detection

### Definitional and Measurement Issues

In this article, a substantive indicator is defined as a psychological instrument that is of interest because of its anticipated relevance to the prediction of some criterion. A response bias is defined as a consistent tendency to respond inaccurately to a substantive indicator, resulting in systematic error in prediction. A response bias indicator is an instrument developed specifically to detect the operation of a response bias.

**Self-report.** A variety of response biases have been identified in the context of self-report measurement. *Inconsistent responding,* also referred to as random or careless responding, occurs when the respondent varies his or her responses across items in an unsystematic manner. *Acquiescence,* or yea-saying, refers to a tendency to endorse the most positive response alternative (*true* or *strongly agree*) without consideration of its accuracy, whereas *negativism,* or nay-saying, reflects the opposite tendency. In the case of scales consisting of polytomous items, two additional possible biases emerge. *Extreme responders* tend to endorse alternatives near the endpoints of the item scale (Hamilton, 1968), whereas *neutral bias,* or moderacy, is manifested in a tendency to choose the middle option of the scale (Hamamura, Heine, & Paulhus, 2008; Schmitt & Allik, 2005). Research on these last two concepts has focused on cultural differences in the tendency to use extreme responses rather than on their relevance to applied psychological measurement, so they will not appear further in this review.

The forms of bias described thus far do not consider the content of the items as comprising the substantive indicator. The two most extensively studied response biases, in contrast, involve responding to item content in a manner that portrays the respondent inaccurately. *Positive impression management* (PIM), the failure to report aberrant tendencies, goes by various other names, including "socially desirable responding"; "impression management"; "underreporting," when used in conjunction with substantive indicators of negative attributes such as psychopathology; or "faking good." *Negative impression management* (NIM) involves responding in an excessively aberrant manner and is also referred to as "faking bad"; "overreporting," when used in conjunction with substantive indicators of negative attributes; or "malingering." Some of these terms, such as "faking bad," are popular but can be problematic because they imply a specific motivation for the misrepresentation; others are more neutral. *Inaccurate responding* refers to a lack of self-knowledge without a consistent tendency toward the underestimation or overestimation of positive features (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990), although this concept has been studied far less extensively than PIM or NIM.

One strategy that has been suggested for the control of response bias in self-report involves designing items with bias in mind, for example, by eliminating items that correlate too highly with a measure of social desirability or by balancing positively and negatively keyed items to compensate for acquiescent or negativistic responding (Jackson, 1970). However, this strategy can result in elimination of some of the most criterion-valid items (e.g., Johnson, 2004). Furthermore, in applied settings the goals of a psychological evaluation often include gathering information about the respondent's honesty and test-taking attitude. For these reasons, the more popular option has been the use of response bias indicators in combination with substantive indicators.

A number of such indicators are now available. Some were developed as part of a larger inventory such as the so-called validity scales of the Minnesota Multiphasic Personality Inventory (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Hathaway & McKinley, 1967) and the Personality Assessment Inventory (Morey, 1991). Others were developed as free-standing instruments, among the most popular of which are the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1998) and the Marlowe–Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960).

**Performance measures.** The character of response biases is somewhat different for substantive indicators based on the respondent's performance rather than for those based on self-report, although there are parallels. The concepts of NIM and PIM generally remain relevant. A useful additional concept when considering response bias in performance-based substantive indicators is that of insufficient or submaximal effort. In the context of neuropsychological assessment, insufficient effort tends to be associated in the literature with malingering or NIM (Sweet et al., 2000), and insufficient effort on measures of psychopathology such as the Rorschach can be considered indicative of PIM (Meyer, 1997). To the extent that non–content-based biases can occur on performance measures (e.g., random responding on a multiple-choice knowledge test), such biases tend to produce an overly negative valuation of the individual that is consistent in outcome with NIM.

A particularly important topic of research in the area of performance-based measurement has to do with the development and validation of bias indicators intended to detect the overreporting of dysfunction in neuropsychological assessment (e.g., Allen, Conder, Green, & Cox, 1997; Tombaugh, 1996). Even the Rorschach inkblot method, which became popular, in part, for its putative resistance to attempts at dissimulating, includes several scores intended to detect underreporting (Exner, 2002).

**Motivation to distort.** In an extremely influential article, Paulhus (1984) proposed that response bias can result from two motivations. Impression management occurs when biased responding is motivated by a desire to mislead the test administrator. This might occur if the assessment is being used for purposes of hiring, if the respondent is seeking disability on psychological grounds, or if the respondent is making a claim of not guilty by reason of insanity. Alternatively, self-deception occurs when the respondent is unaware of the truth. The BIDR was originally intended to distinguish between the two motivations. However, Paulhus and John (1998) subsequently concluded that no scale is effective at discriminating between the two motivations. They also stated that the two scales of the BIDR were instead specific to moralistic and egoistic elements of PIM. Despite reinterpretation of the BIDR scales by its author 10 years ago, articles are still being published asserting that the two BIDR scales are sensitive to the motivation for PIM (e.g., Zaldívar, Molina, López Ríos, & García Montes, 2009). This paradox highlights the possibility that certain generally accepted beliefs about response bias are a greater reflection of strong intuitive presuppositions than empirical evidence.

## Validation Issues

The literature evaluating the validity of response bias indicators generally involves one of three research strategies.[1] Perhaps the bulk of validation research evaluates whether scores on response bias indicators are higher when individuals are instructed to distort the results than under normal instructions. Simulation studies consistently find that means on response bias indicators for groups instructed to fake and means for groups receiving standard instructions can differ by an amount that exceeds the size of the within-group standard deviation (e.g., Baer & Miller, 2002; Dunnette, McCartney, Carlson, & Kichner, 1962; Hough et al., 1990; Nies & Sweet, 1994; Rogers, Sewell, Martin, & Vitacco, 2003; Viswesvaran & Ones, 1999; Zickar, Gibby, & Robie, 2004). A related literature similarly finds that indicators of random or careless responding are sensitive to computer-generated random response data (e.g., Pinsoneault, 2007).

A second line of research has evaluated whether bias indicators can significantly identify individuals with a motivation to distort their current status or individuals who are suspected of distorting. For example, such studies consistently find evidence of elevated scores on indicators of overreporting among individuals suspected of malingering (e.g., Flaro, Green, & Robertson, 2007; Nelson, Sweet, & Demakis, 2006) and elevated scores on indicators of PIM among job applicants (Dunnette et al., 1962; Hough, 1998; Rosse, Stecher, Miller, & Levin, 1998; Stokes, Hogan, & Snell, 1993). Other studies find that specialized instructions warning the individual about the potential for detecting distorted responding can result in lower scores on bias indicators (e.g., Butcher, Morfitt, Rouse, & Holden, 1997; Dwight & Donovan, 2003), although it is uncertain whether such instructions also enhance the validity of substantive predictors (Olson, Fazio, & Hermann, 2007; Robson, Jones, & Abraham, 2008).

Both of these research strategies have significant limitations as a sufficient basis for the use of bias indicators in applied assessment. The first strategy depends on a simulation of uncertain generalizability to situations in which there is an actual motivation to distort. Furthermore, finding that individuals instructed to distort produce elevated scores on a bias indicator does not ensure that most people with elevated scores on a bias indicator were actually distorting. The second strategy assumes that the group differences cannot be attributed to other factors such as recognition that one is under suspicion as a malingerer.

A more direct approach to evaluating the validity of response bias indicators is based on the hypothesis that a valid bias indicator should be able to enhance the predictive accuracy of a valid substantive indicator. This hypothesis, which is subsequently referred to as the *response bias hypothesis,* has unusual implications for the demonstration of validity. Specifically, the criterion-related validity of a response bias indicator is not reflected in its correlation with a criterion but in the degree to which its use enhances the criterion-related validity of a second substantive indicator. Statistically, this hypothesis can be evaluated by treating the bias indicator as a suppressor or as a moderator of the relationship between substantive indicator and criterion.

This proposition can be exemplified by the manner in which bias indicators are used in applied settings. In some cases they are used as suppressors of substantive indicator validity, when the score on the bias indicator is used in additive combination with the score on the substantive indicator to generate what is believed to be a better predictor of the desired criterion. From a statistical perspective, if the underreporting bias tends to depress scores on the substantive indicator, then the *y*-intercept that results when a criterion is regressed onto the substantive indicator will be greater for underreporters than for individuals who respond accurately (see Figure 1a). The best-known example of this approach is the use of the K scale on the MMPI to correct scores on some of the substantive scales of this instrument (McKinley, Hathaway, & Meehl, 1948), though other examples exist (e.g., Millon, Davis, & Millon, 1997).

Alternatively, an elevated score on a bias indicator may be taken as evidence that the results of the substantive indicators should be rejected completely. For example, an elevated score on the MMPI Variable Response Inconsistency scale, which is thought to be an indicator of inconsistent or random responding, is typically taken as evidence that the entire MMPI cannot be interpreted (Greene, 2000). Similarly, an elevated score on a bias indicator may be used as sufficient grounds for rejecting a job applicant (Burns & Christiansen, 2006; Dudley, McFarland, Goodman, Hunt, & Sydell, 2005; Goffin & Christiansen, 2003). In statistical terms, this practice implies that response biases moderate the validity of substantive indicators. For example, the slope that results from regressing the criterion onto the substantive indicator should be greater for those individuals answering honestly than for those answering randomly (see Figure 1b).

To summarize, a response bias can change the score on the substantive indicator. For example, a person who is underreporting should have a lower score on a measure of negative attributes than a person who is answering with perfect accuracy. If that is the case, then an indicator of that response bias should be treated as a suppressor of the validity of the substantive indicator. Alternatively, a response bias can attenuate or even eliminate the criterion-related validity of a substantive indicator. If so, then an indicator of that response bias should be treated as a moderator of the substantive indicator's validity. Of the two, evidence of a moderating bias would be a more serious problem, because it implies the need to reject the results of the substantive indicator completely.

## Statistical Issues

**Suppression.** A variety of statistical strategies has been used to gauge the presence of a suppressor effect in research on bias indicators. A common strategy for detecting suppression effects

[1] This statement is not intended to characterize the literature on response bias but rather to reflect just the literature on the validity of bias indicators. A substantial portion of the literature on response bias is not validational in intent. A number of studies have examined the correlates of scores on response bias indicators in an attempt to identify the determinants of biased responding (e.g., DiStefano & Motl, 2009; Lalwani, Shrum, & Chiu, 2009). Another literature has examined the level of overlap between response bias indicators and substantive indicators, with a high level of overlap interpreted as evidence that substantive indicators are, in fact, largely determined by response bias (e.g., Stevens, Friedel, Mehren, & Merten, 2008). This literature contributes to the understanding of response bias only if the response bias indicator involved, in fact, acts primarily as an indicator of response bias rather than of some other construct. Accordingly, the results of the current review have implications for these other bodies of literature as well as for applied testing practice.

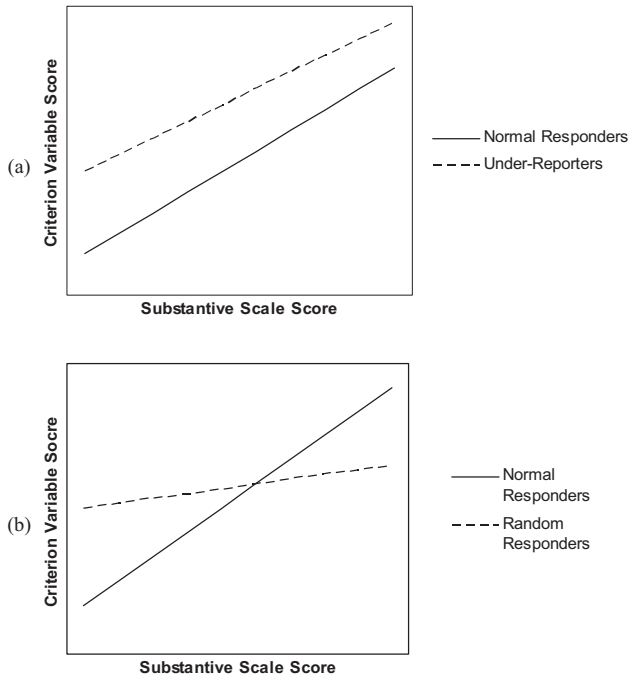


*Figure 1.* Each hypothetical graph involves a substantive predictor and a suitable criterion. (a) Underreporting is used to demonstrate how a response bias can serve as a suppressor. At any level of the predictor, underreporters, on average, demonstrate a higher level of the criterion. Including an indicator of underreporting as a second predictor produces an additive increment in validity. (b) Random responding is used to demonstrate how a response bias can serve as a moderator. Among those who answer normally, higher scores on the predictor, on average, are associated with higher levels of the criterion. The slope is much flatter among those who respond randomly. In this case, the interaction term between the predictor and an indicator of random responding enhances prediction. In practice, individuals with an elevated score on the indicator are often simply eliminated from consideration.

involves examining the improvement in fit resulting from combining the response bias indicator with the substantive indicator in some way. This approach is often implemented with regression analysis. For example, Anderson, Warner, and Spencer (1984) asked job applicants to rate their mastery of various job-relevant tasks as well as of similarly worded but nonexistent tasks (e.g., matrixing solvency files). Score on the first set of tasks was considered the substantive indicator, whereas score on the second set was considered the bias indicator. The authors suggested regressing the substantive indicator onto the bias indicator and using the residual as an indicator of job competence corrected for bias (see equation on p. 576; for a similar recommendation, see also the self-criterion residual described by Paulhus & John, 1998). Anderson et al. also examined the incremental validity of their bias indicator over their substantive indicator.

A related strategy involves simply adding or subtracting the bias indicator from the substantive indicator without any sort of regression-based weighting (e.g., Colvin, Block, & Funder, 1995). The best-known example of this strategy is the MMPI K correction. K is thought to be an indicator of PIM, and the K correction was developed on the basis of the assumption that raw scores on certain MMPI substantive psychopathology indicators tend to be reduced by efforts to present oneself in a positive light. A portion of the raw score on the K scale is added to those substantive indicators: the higher the raw score is on K (which is thought to reflect the degree to which the respondent is engaging in PIM), the greater this correction.

Demonstrating an increment in fit from adding a bias indicator to a substantive indicator is not an adequate test of the response bias hypothesis, however. Some researchers have concluded that response bias indicators often encompass substantive variance beyond that found in the substantive indicator (e.g., Borkenau & Amelang, 1985; Morey et al., 2002), in which case adding the bias indicator to the substantive indicator could improve fit because of an additive effect rather than a suppressor effect. If the raw score on K is in part a function of some element of psychopathology, then adding K to the substantive indicator will improve fit even if K has nothing to do with PIM. Similarly, if Anderson et al.'s (1984) score on nonexistent tasks reveals something substantive about the person's depth of knowledge of the job, then any improvement that comes from adding it to the score on real tasks could have nothing to do with response bias.

The best strategy for testing whether a response bias is suppressing the validity of a substantive indicator involves demonstrating that (a) the substantive indicator and the bias indicator are correlated, and (b) the substantive indicator correlates more highly with the criterion after partialing the bias indicator. That is, a semipartial correlation between the substantive indicator and the criterion that is greater than their zero-order correlation coefficient would be the strongest evidence that the bias indicator serves as a suppressor (Cohen, Cohen, West, & Aiken, 2003).

**Moderation.** Response bias as a moderator also has been evaluated in several suboptimal ways. Weinberger, Schwartz, and Davidson (1979) conducted a widely cited study in which individuals with high scores on the MCSDS who reported low trait anxiety (repressors) demonstrated stronger physiological and behavioral stress reactions than either (a) individuals who were low on both social desirability and trait anxiety or (b) low-defensive, high-anxious participants. However, the absence of a defensive high-anxious group made it impossible to evaluate whether the relationship between self-reported anxiety and physiological reactions varied across levels of socially desirable responding, as a moderator effect would require. Some studies that expanded on this design by including a defensive high-anxious group and then analyzed the groups as four levels of a single variable so that the interaction could not be evaluated separately (e.g., Derakshan & Eysenck, 1998).

A better strategy involves correlating the substantive indicator and the criterion separately within subsets of the sample that have elevated or nonelevated scores on a bias indicator. If a response bias is operating, this correlation should be smaller in the group with elevated bias scores; failure to find a difference, or finding higher correlations among individuals thought to be demonstrating a response bias, would be inconsistent with the response bias hypothesis. Some studies have modified this procedure, comparing the entire sample with the sample excluding respondents with elevated bias scores. Although conceptually similar to the previous approach, this method produces partial overlap between the groups and is thus inappropriate when implemented with statistics designed either for independent groups or dependent groups.

A second acceptable strategy is moderated multiple regression or moderated logistic regression. These approaches involve regressing the criterion on the substantive and bias indicators, then evaluating the increment in fit that results from adding a multiplicative term as a third predictor. The multiplicative term reflects the degree to which the bias indicator moderates the relationship between the substantive indicator and the criterion.

Practical and statistical considerations lead to different conclusions about whether the group comparison or moderated regression is the better choice for detecting a moderator effect (McGrath, 2001). On the one hand, the comparison of correlations between groups is more consistent with the applied use of response bias indicators when a moderating effect is believed to be present, where an elevated score on a bias indicator is often used to reject the scores generated by that respondent. On the other hand, moderated regression is statistically superior because it avoids the complications introduced by having to choose a cut score for a dimensional bias indicator (Dwyer, 1996; MacCallum, Zhang, Preacher, & Rucker, 2002).

Although moderated regression is the recommended statistical analysis in this situation (Frazier, Tix, & Barron, 2004), the successful use of the analysis requires some familiarity with its peculiarities. The effect-size statistics commonly used in connection with moderated regression—the partial and semipartial correlations and Cohen's (1988) $f^2$—represent alternate metrics for expressing the increment in the proportion of overlap between predictors and criterion due to the addition of the interaction term. These increments are often so small that they can be misinterpreted as trivial. Aguinis, Beaty, Boik, and Pierce (2005) found, in a 30-year review of studies in which moderated regression was used with one categorical and one dimensional predictor, that the mean $f^2$ value was only .002 (see also Chaplin, 1991). However, several studies have concluded that even very small moderator effects can be important (e.g., Evans, 1985; McClelland & Judd, 1993) and demonstrate reasonable power for significance tests (Aguinis et al., 2005; McGrath, 2008). This is an important issue to keep in mind when evaluating the results of moderated regression.

Another issue in the use of this analysis has to do with the nondirectional character of a multiplicative term. This term enhances prediction of the criterion whether the relationship between the substantive indicator and the criterion becomes weaker as score on the bias indicator increases (which one would expect if the response bias hypothesis is correct) or the relationship becomes stronger (the opposite of what would be expected). An example of the latter possibility is offered by O'Connor (2006). He hypothesized that high scores on indicators of PIM are actually associated with being honest and forthright, and high scorers may therefore be more rather than less likely to present themselves in an accurate manner in comparison with the general population. Providing evidence for the response bias hypothesis in the case of a moderator effect therefore requires examining the relationship between the substantive and criterion variables within levels of the bias variable (Aiken & West, 1991). The optimal strategy for detection of a response bias moderating effect via regression involves demonstrating that (a) the multiplicative term based on the substantive and bias indicators produces an increment in fit and (b) this increment occurs because the relationship between the substantive indicator and the criterion weakens as score on the bias indicator increases.

## Review of the Research

On the basis of the issues raised in the preceding section, we conducted a review of the response bias literature with the goal of identifying studies in which bias indicators were used as moderators or suppressors of the criterion-related validity of substantive indicators. This review is distinctive among reviews of the response bias literature in that (a) it focuses exclusively on studies that evaluated whether response bias indicators suppress or moderate the validity of substantive indicators and (b) it compares the evidence across testing contexts. The context of an assessment, particularly the extent to which there is a motivation to distort, is a potentially important moderator of the degree to which efforts to distort occur (Ben-Porath & Waller, 1992; Schmit & Ryan, 1993). For example, the general assessment of personality often has little effect on the respondent's subsequent functioning except in work-related situations, and thus there is relatively little incentive to present in an inaccurate manner. Similarly, emotional distress and psychopathology are usually assessed as part of the treatment planning process. It can be expected that the respondent will usually answer honestly unless the respondent believes some other goal can be achieved by distorting the results. In contrast, work-setting, forensic, and disability evaluations often involve an adversarial element, and the motivation to distort can be substantial. In short, the context of the assessment is likely to be an important moderator of the base rate for bias.

We began by searching PsycINFO for any study that included a common variant of one of the following terms in the title: *feign*, *fake*, *malinger*, *random responding*, *social desirability*, *dissimulate*, *response bias*, *response set*, *response style*, or *impression management*. The initial search generated approximately 4,000 hits, most of which were unrelated to the topic of response bias. Over 600 of the original hits represented dissertations. These were ultimately excluded from consideration primarily because of retrieval problems and concerns about the quality of some of the research.

Abstracts of articles relevant to the topic of response bias were reviewed for any evidence that the researchers evaluated the presence of a suppressor or moderator effect in an applied setting. We excluded studies or analyses in which both the substantive indicator and the criterion were self-report measures, because such results were potentially confounded by the presence of the same bias in both instruments. Other contexts besides the five mentioned above were considered for inclusion—in particular, child custody evaluation and assessment for criminal court proceedings—but were omitted because no studies meeting the criteria for inclusion were found.

This process identified surprisingly few studies. These, in turn, were reviewed for leads to other studies, as were a number of recent publications concerning response bias. Finally, colleagues who were thought likely to be knowledgeable about similar studies were asked to submit additional possibilities. Despite extensive searching, the final pool consisted of only 41 studies.

### General Personality Assessment

The search generated 22 studies that evaluated the validity of response bias indicators as moderators or suppressors of relationships between substantive scales of personality and criteria col-

lected by the use of other measurement models in a general normal sample. For unknown reasons, earlier studies in this set tended to focus more on bias as a suppressor, whereas later studies typically evaluated bias as a moderator of scale validity. It is unclear whether this pattern reflects increasing recognition that bias indicators are more commonly used to reject rather than to correct substantive scales (i.e., as moderators rather than as suppressors), increasing comfort with more sophisticated statistical concepts, such as moderated regression or random error.

Table 1 provides a summary of each study. The column at the far right of the table summarizes all outcomes reported in the studies, organized according to the type of relationship (suppression or moderation) and response bias. The percentage of statistical outcomes that were in the direction expected if the response bias hypothesis were correct provided a useful benchmark; this percentage is presented when it could be computed. Higher percentages suggest more consistent evidence for the successful detection of response biases in the sample. Overall, out of 44 sets of outcomes described in the table, only 12 provided evidence supportive of the response bias hypothesis.

Results from four types of analyses occurred with enough frequency to allow aggregation across studies for purposes of drawing summative conclusions. To reduce the impact of any one study on the results, we computed a mean effect size for each study weighted by the sample sizes for the analyses. We then computed an average of these mean effect sizes after weighting by the mean sample size across analyses from the study. No correction was made for unreliability because the focus of this review is on the operational effectiveness of bias indicators and because the reliability of a number of the criteria was uncertain. This issue of unreliability will be addressed further in the discussion of the results.

The most common test of suppression was the optimal strategy described above, in which the correlation between a substantive indicator and the criterion was compared with the semipartial correlation controlling for the response bias indicator. Authors reported results from 183 uncorrected and 357 semipartial correlations. The average bivariate correlation between the substantive indicator and the criterion was .36. After partialing the response bias indicator, the average correlation dropped slightly, to .33. Not only was there no evidence of a suppression effect but partialing the bias indicator, on average, slightly reduced correlations between substantive indicators and criteria. The most likely explanation for such an outcome would be that the bias indicators primarily represented additional substantive variance.

Three statistical approaches to evaluating moderator effects occurred frequently enough to allow a report of aggregate findings. McCrae, Stone, Fagan, and Costa (1998) described a unique approach that involved computing a profile agreement statistic between the substantive indicator and the criterion. This statistic was then correlated with scores on bias indicators. The size of these correlations is indicative of the degree to which bias moderates the relationship between self-report and observer report. Over 60 comparisons, the mean correlation was approximately 0.

More common were the two strategies listed above as optimal. The first involved using some cut score to identify potentially invalid cases. The mean correlation between substantive indicator and criterion was then computed separately within the putatively valid and putatively invalid cases.[2] As noted previously, this analytic approach offers a reasonable parallel to the practical use of bias indicators to exclude cases from consideration. Across 350 such comparisons, the mean correlation among putatively valid cases was .27. Among putatively invalid cases, this mean was .29. Again, the difference in the mean effects was nearly indistinguishable and inconsistent with the response bias hypothesis.

Finally, the increment in fit for an interaction term was available for 35 comparisons.[3] The mean increment was again almost 0. Though the number of analyses reflected in this average is small, the conclusion is consistent with information about interactions provided in Table 1. Of 638 significance tests for interactions, only 49 (<8%) were significant. To evaluate whether the lack of significant results could be attributable to low statistical power, the proportion of significant interactions in a set was correlated with the sample size for that set. This correlation proved to be −.02, which is inconsistent with the hypothesis that the lack of significance is primarily due to insufficient power. However, this estimate should be interpreted cautiously as it was based on only 10 data points, each of 10 studies serving as a single data point. Consistent with this conclusion, the two studies that involved the most extensive analysis of interaction terms (Borkenau & Ostendorf, 1992; Piedmont et al., 2000) found no consistency in the directionality of the interaction terms.

These results provide little evidence to suggest that members of the general population, when asked to describe their personality in circumstances where there is no motivation to misrepresent themselves, actually do so. The failure to find response bias when there is no motivation to distort in itself may not seem particularly momentous. However, it is important to remember the longstanding belief that certain individuals provide distorted results not because of any external incentive to do so but because they are characteristically incapable of perceiving themselves accurately (Paulhus, 1984). The failure to find evidence that bias influences the validity of substantive scales when there is no external motivation to distort raises the possibility that the role of self-deception, at least in the general population, has been overestimated.

## Emotional Disorders Assessment

Most of the studies examining the response bias hypothesis in emotional disorders have focused on two issues. The first is the validity of bias indicators under instructions to simulate, a set of studies already excluded from consideration. The second is the validity of the MMPI K correction. The evidence is quite consistent that K correction either has no effect on the validity of MMPI substantive indicators or actually tends to reduce their validity (Alperin, Archer, & Coates, 1995; Archer, Fontaine, & McCrae, 1998; Barthlow, Graham, Ben-Porath, Tellegen, & McNulty, 2002; Colby, 1989; Heilbrun, 1963; Wooten, 1984). Because this evidence has been reviewed previously (e.g., see Barthlow et al.,

[2] Kurtz and Parrish (2001) trichotomized cases, whereas Holden (2007) provided predictor–criterion correlations across 12 values of the bias indicator. For these analyses, only the most extreme groups from these two studies were considered.

[3] Of these 35 analyses, 10 involved the increment in $R^2$, whereas 25 examined the $f^2$ value associated with the moderator term. The two statistics are closely related, but $f^2$ tends to be slightly larger than $R^2$.

Table 1
*Summary of Articles: General Personality Assessment*

| Study | Sample | Substantive indicators | Bias indicators | Response styles | Criteria | Results |
|---|---|---|---|---|---|---|
| Dicken (1963) | 410 from various populations | CPI | CPI | Acquiescence, PIM | Informant ratings | Suppression: 7 of 12 (58%) correlations larger after partialing acquiescence[a]<br>Suppression: 30 of 66 (45%) correlations larger after partialing PIM |
| McCrae & Costa (1983) | 215 community residents | NEO | EPI Lie, MCSDS | PIM | Spouse ratings | Suppression: 2 of 42 (5%) correlations larger after partialing PIM |
| McCrae (1986) | 62 community residents | Well-being indicators | ESDS | PIM | Informant ratings | Suppression: 0 of 3 (0%) correlations larger after partialing PIM |
| McCrae et al. (1989) | 112 community residents, 82 angiography patients, 113 law students | MMPI HO | K scale | PIM | Informant ratings, interview-based ratings of hostility potential, mortality risk | Suppression: 0 of 20 (0%) correlations larger after controlling for K scale scores |
| Weed et al. (1990) | 1,681 spouses | MMPI-2 | MMPI-2 O-S index | NIM, PIM | Spouse ratings | Moderation: Mean of 43 correlations .05 higher in NIM group<br>Moderation: Mean of 43 correlations .02 lower in PIM group[a] |
| Borkenau & Ostendorf (1992) | 300 community residents | EPI, NEO | ESDS, MCSDS, SFS, Factor scores from PIM indicators | PIM | Informant ratings | Suppression: 12 of 60 (20%) correlations larger after partialing PIM<br>Moderation: 37 of 60 (62%) correlations smaller in PIM group[a]<br>Moderation: 30 of 60 (50%) interactions in expected direction |
| Tomaka et al. (1992) | 64 male college students | Perceived Stress Scale | MCSDS | PIM | Tension physiological measures | Moderation: 0 of 3 (0%) interactions significant |
| Shapiro et al. (1995) | 209 college students | AES, MMPI HO | MCSDS | PIM | Tension blood pressure and heart rate | Moderation: 2 of 6 (33%) interactions significant, no consistent direction |
| Melamed (1996) | 82 workers | Emotional Reactivity Scale | MCSDS | PIM | Tension blood pressure and heart rate | Moderation: For 3 interactions the largest increment in validity was .001 |
| Barger et al. (1997) | 119 college students | TMAS | MCSDS | PIM | Heat rate and skin conductance | Moderation: 0 of 2 (0%) interactions significant |

(*table continues*)

Table 1 (*continued*)

| Study | Sample | Substantive indicators | Bias indicators | Response styles | Criteria | Results |
|---|---|---|---|---|---|---|
| McCrae et al. (1998) | 94 married community residents | NEO | NEO | Acquiescence, Extreme responding, Inconsistent responding, NIM, PIM | Spouse ratings | Moderation: 7 of 12 (58%) correlations between acquiescence and profile similarity in expected direction[a] |
| | | | | | | Moderation: 3 of 12 (25%) correlations between extreme responding and profile similarity in expected direction |
| | | | | | | Moderation: 6 of 12 (50%) correlations between inconsistent responding and profile similarity in expected direction |
| | | | | | | Moderation: 4 of 12 (33%) correlations between over-reporting and profile similarity in expected direction |
| | | | | | | Moderation: 10 of 12 (83%) correlations between under-reporting and profile similarity in expected direction[a] |
| | | | | | | Moderation: 2 of 80 (3%) interactions significant, and those pointed in opposite directions |
| Eysenck & Derakshan (1999) | 156 students and community residents | FSAQ | MCSDS | PIM | Informant ratings | Moderation: 1 of 1 (100%) correlations larger in PIM group[a] |
| | | | | | | Moderation: 0 of 1 (0%) interactions significant |
| al'Absi et al. (2000) | 46 male students and community residents | AES | MCSDS | PIM | Tension cardiac and endocrine measures | Moderation: 1 of 5 (20%) interactions significant at .05 |
| Piedmont et al. (2000) | 178 students | MPQ, NEO | MPQ | Acquiescence | Informant ratings | Suppression: 18 of 82 (22%) correlations larger after partialing bias |
| | 1,728 community residents | | NEO | Inconsistent responding | | Suppression: Median of 36 correlations was .06 lower after partialing one PIM measure; .01 lower after partialing another; no different after partialing NIM |
| | | | | Negativism | | Moderation: 20 of 348 (6%) interactions significant, no consistent direction |
| | | | | NIM | | Moderation: 1 of 2 (50%) correlations lower in acquiescent group |
| | | | | PIM | | Moderation: 4 of 14 (29%) correlations lower in inconsistent group |
| | | | | | | Moderation: 2 of 3 (67%) correlations lower in PIM/NIM group[a] |
| | | | | | | Moderation: 6 of 19 (32%) correlations lower in mixed bias group |
| | | | | | | Moderation: 22 of 180 (12%) interactions significant |
| | | | | | | Moderation: Median of 90 correlations .02 lower in biased group for one sample[a]; no different in another |

(*table continues*)

Table 1 (*continued*)

| Study | Sample | Substantive indicators | Bias indicators | Response styles | Criteria | Results |
|---|---|---|---|---|---|---|
| Kurtz & Parrish (2001) | 109 college students | NEO | NEO | Inconsistent responding | Retest, Informant ratings | Moderation: 9 of 12 (65%) test-retest statistics lower in inconsistent group[a] |
| | | | | | | Moderation: 6 of 11 (55%) correlations lower in inconsistent group[a] |
| | | | | | | Moderation: 1 of 7 (14%) interactions significant |
| Lee & Klein (2002) | 134 undergraduate business students | NEO | BIDR | PIM | Learning | 23 of 36 (64%) correlations larger after partialing bias[a] |
| Egloff & Schmukle (2003)[b] | 207 college students | STAI-Trait | BIDR, Updated MCSDS | PIM | IAT Anxiety | Moderation: 0 of 5 (0%) interactions significant; the largest increment in overlapping variance was .008 |
| Pauls & Stemmler (2003) | 78 female students and community residents | STAI | MCSDS | PIM | Tension physiological measures | Moderation: 1 of 3 (33%) interactions significant, not in correct direction |
| Hofmann et al. (2005) | 93 college students | Attitude scales | MCSDS | PIM | IAT Attitudes | Moderation: 0 of 2 (0%) interactions in expected direction |
| O'Connor (2006) | 223 college students | NEO | BIDR, MCMI Desirability and Debasement, NEO NPM | PIM, NIM | Informant ratings | Moderation: 5 of 39 (13%) interactions in expected direction for PIM |
| | | | | | | Moderation: 24 of 29 (83%) interactions in expected direction for NIM |
| Holden (2007) | 420 college students | FFI | IM | PIM | Informant ratings | Moderation: 5 of 5 (100%) interactions in expected direction[a] |
| Kurtz et al. (2008) | 183 college students | FFI | MCSDS | PIM | Informant ratings | Suppression: 0 of 10 (0%) correlations larger after partialing bias |

*Note.* CPI = California Psychological Inventory; PIM = -positive impression management; NEO = NEO Personality Inventory; EPI = Eysenck Personality Inventory; MCSDS = Marlowe–Crowne Social Desirability Scale; ESDS = Edwards Social Desirability Scale; MMPI = Minnesota Multiphasic Personality Inventory; HO = Hostility Scale; K scale = validity scale tapping PIM; O-S = Obvious–Subtle Index; NIM = negative impression management; SFS = Sets of Four Scale; AES = Anger Expression Scale; TMAS = Taylor Manifest Anxiety Scale (short form); FSAQ = Four Systems Anxiety Questionnaire; MPQ = Multidimensional Personality Questionnaire; BIDR = Balanced Inventory of Desirable Responding; STAI = State–Trait Anxiety Inventory; IAT = Implicit Association Test; MCMI = Millon Clinical Multiaxial Inventory; NPM = Negative Presentation Management; FFI = NEO Five Factor Inventory; IM = Impression Management Scale. When possible, results are provided in terms of the number and percentage of comparisons in the direction more consistent with the response bias hypothesis. For studies that evaluated multiple response biases, these are grouped by response bias when possible. Certain studies examined higher order interactions. In such studies, any significant interaction involving substantive predictors and bias scales was counted as a significant outcome for purposes of tallying.

[a] This outcome is consistent with the response bias hypothesis (e.g., more than half of analyses are in the expected direction). [b] This study reversed typical practice by using a performance measure as the predictor of outcome on a self-report measure.

2002), with the general conclusion that the K correction is a questionable method of protecting against response bias, including this literature here would have skewed the results against response bias indicators.[4]

Once this literature was excluded, however, we could find only three studies having to do with the evaluation of emotional disorders that met criteria for inclusion in the current review (see Table 2). This finding was unexpected, raising concerns about the sufficiency of the evidence base for using bias indicators in psychiatric settings.

Holden, Mendonca, and Serin (1989) found that three moderator terms were associated with a significant increment in the proportion of overlapping variance with clinician ratings of suicidality. However, these authors did not evaluate the direction of the moderator effect, so the findings remain equivocal.

Archer et al. (1998) provided evidence that eliminating cases on the basis of the MMPI-2 Variable Response Inconsistency (VRIN) scale, which employs an innovative approach to the detection of inconsistent responding suggested by Tellegen (1988), improved correlations with clinician ratings, on average, by .16. In contrast, McGrath, Rashid, Hayman, and Pogge (2002) found that correlations between MMPI substantive scales and clinician data declined, on average, by .02 after excluding cases because of elevated scores on response bias indicators. Exclusion in this study was based on a combination of response bias indicators, so it is uncertain whether all of the indicators they examined, or only some of them, reduced validity. However, the modal reason for an invalid response in their sample was an elevated score on the VRIN scale.

Clearly, there is insufficient evidence to justify drawing conclusions about the validity of any response bias indicator commonly used in the assessment of emotional disorders other than the K correction of the MMPI. There is some intriguing evidence to suggest that inconsistent responding as indicated by the VRIN scale of the MMPI may reduce the validity of substantive indicators. This has been demonstrated in only one study, however, and the results reported by McGrath et al. (2002) raised some concerns about whether this finding can be replicated. It is noteworthy, in light of this discussion, that Kurtz and Parrish's (2001) study of inconsistent responding generated some of the most consistent evidence for the operation of response biases in general personality functioning and Piedmont et al. (2000) remarked that they considered the VRIN scale the most promising of the response bias indicators they examined.

## Work-Setting Assessment

A substantial literature exists on the use of psychological instruments in corporate settings, and bias indicators are used widely in employee selection and evaluation (e.g., Goffin & Christiansen, 2003). Most studies looking at response bias in connection with work settings do not pursue the key question of whether putative bias affects the relationships between evaluation methods and subsequent job performance, but we were able to find 11 studies that evaluated bias indicators as suppressors or moderators in real-world settings (see Table 3).

Three of these studies focused exclusively on job applicants, six on incumbents, and two on a combination of the two. In the studies that examined applicants, the bias indicator does not seem to have been considered during the hiring process, a strategy that enhances the potential for finding an effect. Incumbents were typically informed that the results of the testing would have no impact on their employment, though it is reasonable to suspect that employees would vary in response to such assurances from indifference to suspicion. The criterion usually consisted of a binary variable indicating subsequent job tenure or some sort of performance evaluation. Not surprisingly, these studies focused almost exclusively on the issue of PIM, which was usually referred to as social desirability or impression management in this literature.

Only four of 18 sets of analyses listed in the table were supportive of the response bias hypothesis. Aggregation was possible for four different types of statistical output, though the number of analyses contributing to each aggregate was substantially smaller than was the case for the assessment of personality in general. The mean of 32 bivariate correlations was .15, whereas the mean for 32 semipartial correlations was .12. The mean correlations were substantially smaller than in the case of general personality assessment, even though the types of substantive scales used were similar, suggesting that (a) substantive personality indicators are less effective for the prediction of job performance than for the types of criteria used in general personality research or (b) the validity of these indicators is generally attenuated in evaluation settings with potential consequences. Whatever the explanation, the conclusion is the same: Partialing response bias tended to reduce the size of correlations with criteria rather than enhance them.

It was possible to compare 32 correlations and semipartial correlations, of which 21 were based on applicants and nine on incumbents (two involved both). For applicants, the mean correlation was .19 and the mean semipartial correlation was .17. For incumbents these values were .10 and .06, respectively. Job applicants generated higher correlations with criteria, but in all cases the results were inconsistent with the response bias hypothesis.

Stokes et al. (1993) correlated individual items with criteria, then computed correlations between those correlations and item social desirability ratings. This analytic strategy bears some resemblance to McCrae et al.'s (1998) computation of correlations with profile similarity statistics. Across 22 analyses, the mean of these correlations was .22. That is, items more susceptible to socially desirable responding were actually better predictors of criteria. This finding does not directly address whether socially desirable responding reduces validity, but it is inconsistent with the argument. It also supports prior conclusions that indicators of response bias, at least social desirability, may incorporate important substantive variance that may not be adequately addressed by the substantive scale.

A total of 99 correlations between substantive scales and criteria were computed separately for putatively valid and invalid cases. Contrary to the finding for personality assessment in general, there was some evidence here of a slight benefit from eliminating potentially invalid cases. The mean correlation among individuals with elevated scores on bias indicators was .14, whereas the mean corre-

[4] One study in the section on normal personality (McCrae et al., 1989) focused on the K correction. Given the concerns raised here, the mean semipartial correlation was recomputed omitting this study. The mean increased to .35, but the general conclusion was the same: partialing the bias indicator still reduced the mean correlation with the criterion slightly.

Table 2
*Summary of Articles: Emotional Disorders Assessment*

| Study | Sample | Substantive indicator | Bias indicator | Response styles | Criterion | Results |
|---|---|---|---|---|---|---|
| Holden et al. (1989) | 97 inpatients | Hopelessness scale | PRF Desirability scale | PIM | Clinician ratings | Moderation: 3 of 3 (100%) interactions significant[a] |
| Archer et al. (1998) | 692 inpatients | MMPI-2 | MMPI-2 | Inconsistency | Clinician ratings | Moderation: 28 of 30 (93%) correlations lower in inconsistent group[b] |
| McGrath et al. (2002) | 752 inpatients | MMPI-2 | MMPI-2 | Acquiescence, inconsistency, negativism, NIM | Clinician ratings | Moderation: Eliminating protocols based on validity scales reduced validity coefficients an average of .012–.023 |

*Note.* PRF = Personality Research Form; MMPI-2 = Minnesota Multiphasic Personality Inventory–II; NIM = negative impression management.
[a] This outcome is consistent with the response bias hypothesis. However, change in $R^2$ is not a sufficient basis for concluding moderation is in the expected direction. [b] This outcome is consistent with the response bias hypothesis (more than half of analyses are in the expected direction).

lation for those individuals classified as valid was .16. All of these correlations came from a single study involving incumbents who were soldiers, informed that the testing was not relevant to their future career (Hough et al., 1990). Clearly, the finding bears replication.

Results from this study were particularly supportive for analyses involving the evaluation of inconsistent responding. As Table 3 indicates, 85% of the correlations between substantive predictors and criteria were larger in the consistent group than in the inconsistent group. Judging from the number of studies devoted to each form of bias, however, it is evident that PIM is considered a far more serious concern in work settings than is inconsistent responding. In particular, none of the studies cited considered whether the same pattern would emerge for inconsistency among job candidates.

Finally, three studies provided information about the increase in the proportion of overlapping variance that resulted from adding a moderator term. Across six analyses the mean increment was .00. The results, as a whole, continue to support the conclusion that the use of bias indicators may not enhance the effectiveness, and may even reduce the effectiveness, of substantive predictors, though inconsistent responding may represent an exception in certain settings.

It is worth noting that researchers in the field of employee selection have introduced two other innovative research methods for evaluating the value of response bias indicators. Thus far, these methods have been used only in work settings and therefore cannot be compared with results in the other contexts included in this review. Even so, the results provide further support for questioning the degree to which response bias plays a role in the responding of job applicants.

One method uses meta-analysis to estimate the size of each of the three bivariate correlations between bias indicator, substantive indicator, and criterion. This strategy allows the researcher to estimate semipartial correlations, controlling for bias, from a substantially larger body of research than would meet the criteria for the present review. Using this methodology, Ones, Viswesvaran, and Reiss (1996) found for all five substantive dimensions they examined that the mean semipartial correlation equaled the mean correlation. More recently, Li and Bagger (2006) used a similar approach to evaluate Paulhus's (1984) two facets of social desirability, self-deception and impression management. The results were essentially equivalent, with the mean semipartial correlation never exceeding the mean correlation by more than .04.

The second strategy involves the use of factor analysis. When the factor structure for substantive indicators are compared across respondents instructed to answer honestly and respondents instructed to fake, the factor structure for the latter group tends to be simpler and to collapse into a single factor (Ellingson, Sackett, & Hough, 1999). In contrast, factor structures are usually comparable across groups with different levels of naturally occurring motivation to distort (e.g., Ellingson, Smith, & Sackett, 2001; Fan, Wong, Carroll, & Lopez, 2008; Hogan, Barrett, & Hogan, 2007; Marshall, De Fruyt, Rolland, & Bagby, 2005; Michaelis & Eysenck, 1971; Smith & Ellingson, 2002; but see Schmit & Ryan, 1993, for an exception). Invariance in the factor structure argues against the hypothesis that individuals with a motivation to distort are responding to substantive indicators differently than those without this motivation, at least in work-related settings.

The failure to find support for the response bias hypothesis in the context of work settings has an important implication for the assessment of response bias in the context of disability and forensic evaluation. Clearly, job applicants—and to a lesser extent job incumbents—can benefit from self-misrepresentation, at least to the extent that they do not meet the criteria the assessment is attempting to detect. Motivation to mislead is not a sufficient basis for assuming that purposeful and successful deception is occurring.

## Disability Assessment

There is a substantial literature devoted to disability claimants who ostensibly fake physical or emotional distress. The most popular topic in this literature seems to be differences on response bias indicators between individuals with and without a motivation to distort. The number of studies that met criteria for inclusion in the current review was surprisingly small: We were only able to identify four (see Table 4).[5] Two of those evaluated pain patients (Fishbain, Cutler, Rosomoff, & Steele-Rosomoff, 2002; Logan, Claar, & Scharff, 2008), and neither of those provided evidence of moderation or suppression effects.

Most studies on disability malingering focus on the misrepresentation of cognitive abilities during neuropsychological assessment. A particularly important form of bias indicator in this

[5] We are particularly grateful to Paul Lees-Haley and Paul Green for the assistance they provided in identifying potentially relevant literature on neuropsychological malingering.

Table 3
*Summary of Articles: Workplace Assessment*

| Study | Sample | Substantive indicator | Bias indicator | Response styles | Criterion | Results |
|---|---|---|---|---|---|---|
| Kriedt & Dawson (1961)[a] | 41 clerical workers | Gordon Personality Inventory | Total score | PIM | Supervisor evaluation | Suppression: 0 of 4 (0%) correlations larger after partialing bias |
| Anderson et al. (1984) | 66 clerical applicants | Self-rated experience level | Bogus experience level items | PIM | Typing performance test | Suppression: Bias indicator increased $R^2$ by .16[b] |
| | | | | | | Moderation: 0 of 1 (0%) interaction significant |
| Arnold et al. (1985) | 415 accountants, | Job satisfaction ratings | MCSDS | PIM | Turnover | Moderation: 0 of 4 (0%) logistic interactions significant |
| George & Smith (1990) | 47 government applicants | Assessment center self-ratings | MCSDS | PIM | Supervisor ratings, promotion review | Moderation: 0 of 2 (0%) correlations lower in PIM group |
| Hough et al. (1990)[a] | 9,359 military personnel | ABLE temperament scales | ABLE | Inaccurate responding, Inconsistent responding, NIM, PIM | Supervisor and peer evaluations, performance records | Suppression: 0 of 33 (0%) $R$ values significantly increased by inaccurate responding variable |
| | | | | | | Moderation: 0 of 33 (0%) interactions significant for inaccurate responding |
| | | | | | | Moderation: 28 of 33 (85%) correlations lower in inconsistent group[c] |
| | | | | | | Moderation: 22 of 33 (67%) correlations lower in NIM group[c] |
| | | | | | | Moderation: 19 of 33 (58%) correlations lower in PIM group[c] |
| Jamner et al. (1991) | 26 paramedics | MMPI HO | MCSDS | PIM | Tension physiological measures | Moderation: 1 of 3 (33%) interactions significant, not in correct direction |
| Stokes et al. (1993)[a] | 810 incumbents, 555 applicants | Constructed biographical data scale | Constructed scale | PIM | Job tenure | Suppression: 0 of 2 (0%) correlations larger after partialing PIM |
| | | | | | | Moderation: 2 of 22 (9%) correlations between PIM and predictor–criterion correlations in expected direction |
| | | | | | | Moderation: 0 of 2 (0%) interaction terms significant |
| Christiansen et al. (1994) | 84 forestry applicants and incumbents | 16PF | Constructed 16PF scales | NIM, PIM | Supervisor ratings | Suppression: 1 of 4 (25%) correlations larger after controlling for bias |
| Barrick & Mount (1996) | 286 truck driver applicants | PCI | BIDR | PIM | Job tenure, supervisor ratings | Suppression: 5 of 20 (25%) correlations larger after partialing PIM |
| Martocchio & Judge (1997)[a] | 97 employees | NEO Conscientiousness | BIDR | PIM | Learning | Suppression: 0 of 4 (0%) correlations larger after partialing bias |
| Reid–Seiser & Fritzsche (2001)[a] | 90 customer service representatives | NEO | NEO, BIDR | PIM | Supervisor ratings | Moderation: 0 of 15 (0%) interactions significant |

*Note.* PIM = positive impression management; MCSDS = Marlowe–Crowne Social Desirability Scale; ABLE = Assessment of Background and Life Experiences; NIM = negative impression management; 16PF = Sixteen Personality Factor Test; PCI = Personal Characteristics Inventory; BIDR = Balanced Inventory of Desirable Responding; NEO = NEO Personality Inventory; MMPI = Minnesota Multiphasic Personality Inventory; HO = Hostility scale. When possible, results are provided in terms of the number and percentage of comparisons in the expected direction separated by response bias.

[a] This study permitted separate examination of incumbents. [b] This outcome is consistent with the response bias hypothesis. However, change in $R^2$ is not a sufficient basis for concluding that suppression is occurring. [c] This outcome is consistent with the response bias hypothesis.

Table 4
*Summary of Articles: Disability Assessment*

| Study | Sample | Substantive indicators | Bias indicator | Response style | Criterion | Results |
|---|---|---|---|---|---|---|
| Chronic pain | | | | | | |
| Fishbain et al. (2002) | 96 chronic pain sufferers | STAI–State | Conscious Exaggeration Scale | NIM | Return to work | Suppression: 0 of 1 (0%) regression analysis found a significant bias effect |
| Logan et al. (2008) | 414 adolescent chronic pain patients | CDI, RCMAS | RCMAS Lie scale | PIM | Clinician ratings | Moderation: 0 of 2 (0%) of comparisons in expected direction |
| Cognitive impairment | | | | | | |
| Bowden et al. (2006)[a] | 86 neuropsychological referrals | Delayed memory score, FSIQ, PIQ | WMT | NIM | Injury severity | Moderation: 0 of 4 (0%) interactions significant |
| Rohling & Demakis (2010)[a] | 477 neuropsychological referrals | GMI | WMT | NIM | Injury severity | Moderation: 0 of 4 (0%) interactions significant |

*Note.* STAI = State–Trait Anxiety Inventory; NIM = negative impression management; CDI = Children's Depression Inventory; RCMAS = Revised Children's Manifest Anxiety Scale; PIM = positive impression management; FSIQ = Full-Scale Intelligence Quotient; PIQ = Performance Intelligence Quotient; WMT = Word Memory Test; GMI = General Memory Index.
[a] This study reversed typical practice by using a performance measure as the predictor of outcome on substantive scales.

context is the forced-choice symptom validity test (SVT), though other options are available (see Bender & Rogers, 2004). SVTs generally involve a relatively easy recognition task. If each item is accompanied by four possible choices, correct outcomes on 25% of the items can be expected simply by chance. Originally, it was thought that malingerers should perform below the chance level. This hypothesis has not generally been supported; in fact, below-chance performance is quite rare (e.g., Gervais, Rohling, Green, & Ford, 2004). Even so, groups with a pre-existing incentive to present themselves negatively, such as individuals who are seeking compensation for brain trauma, consistently generate lower scores on SVTs than groups without similar motivation (e.g., Green, Lees-Haley, & Allen, 2002; Greve et al., 2006).

There is strong intuitive appeal associated with the use of SVTs to detect malingering, although that appeal is attenuated when the basis for detection is below-average performance rather than below-chance performance. Unfortunately, despite extensive searching we found only two studies meeting the criteria for inclusion that used an SVT to evaluate cognitive malingering. In both cases, the SVT used was the Word Memory Test of Green, Allen, and Astner (1996).

Bowden, Shores, and Mathias (2006) examined interactions using the duration of posttraumatic amnesia as an objective indicator of injury severity and three different measures of neuropsychological functioning as predictors. None of four interactions with the Word Memory Test were significant, and the proportion of variance accounted for by the interaction was consistently less than .01. Because the results were not significant, the authors did not explore whether the direction of the multiplicative relationship was consistent with the response bias hypothesis.

Bowden et al. (2006) were responding to an earlier study by Green, Rohling, Lees-Haley, and Allen (2001), which had concluded that effort is a more important determinant of scores on neuropsychological indicators than severity of injury. Yet Green et al. did not test the interactions directly relevant to the response bias hypothesis. Bowden et al. challenged Green et al.'s conclusion that effort should be controlled for in neuropsychological evaluations, on the grounds that the case is incomplete without demonstration of a significant moderator effect associated with effort.

In an effort to synthesize the two sets of findings, Rohling and Demakis (in press) re-analyzed both studies. After attempting to control for certain demographic differences between the samples by selection, they recomputed two of the interactions reported by Bowden et al. (2006) and computed four previously unreported interactions on the basis of the sample of Green et al. (2001). None of these interactions proved to be significant.

It can be argued that the results may have been attenuated by reliance on the Word Memory Test, which seems to be associated with a higher positive rate than other popular SVTs (Gervais et al., 2004; Greve, Binder, & Bianchini, 2008). This finding suggests the possibility of a higher false positive rate (but see Greiffenstein, Greve, Bianchini, & Baker, 2008), although it could also mean a higher valid positive rate instead (or in addition). Unfortunately, the relative hit rate of different SVTs remains an indeterminate issue in the absence of a clear criterion.

A second issue is the relatively restricted set of substantive indicators evaluated, focusing exclusively on memory and intellectual ability, and of criteria, which were restricted to injury severity. Despite evidence that injury severity is related to a broad spectrum of cognitive abilities (e.g., Draper & Ponsford, 2008), external validity concerns warrant replication with a broader variety of both substantive indicators and nontest criteria.

## Forensic Assessment

Surprisingly, we found only one study meeting criteria for inclusion in this review that specifically targeted a forensic pop-

ulation (see Table 5).[6] Edens and Ruiz (2006) found that two of four interactions were significant when predictors were treated either as dimensional or as dichotomized variables. Unfortunately, the authors did not evaluate the direction of the multiplicative terms.

## Possible Reasons for the Failure to Confirm

Across all five contexts evaluated, we found a lack of evidence for the response bias hypothesis. However, the implications of the phrase "lack of evidence" differ across contexts. In three contexts—emotional disorders, disability evaluation, and forensic evaluation—the evidence was simply insufficient to draw firm conclusions, and that outcome will be the topic of the next section. In those contexts where a sufficient body of evidence was found to draw conclusions about the hypothesis—general personality and work-related assessment—the evidence generally failed to corroborate the hypothesis, although careless responding represents a possible exception. We identified three possible explanations for the failure to support the latter finding: popular response bias indicators may not be particularly effective indicators of self-misrepresentation, the base rate of self-misrepresentation has been seriously misestimated, or criteria are too coarse to provide a sufficient basis for the detection of bias. The remainder of this section will be devoted to consideration of each of these possibilities.

### The Validity of Bias Indicators

Serious concerns have been raised about the validity of two of the most thoroughly studied response bias indicators, the MMPI K correction and Obvious-Subtle Index (Barthlow et al., 2002; Weed et al., 1990). Both the BIDR and the MCSDS, the most popular stand-alone indicators of PIM, have been criticized as well (e.g., Barger, 2002; Burns & Christiansen, 2006). Finding that the validity of the most thoroughly studied indicators is questionable raises concerns about the effectiveness of this class of instruments as a whole.

A related possibility is that biased responding may be a more complex and subtle phenomenon than most bias indicators are capable of gauging. PIM has been hypothesized to encompass several facets, including claims of extreme virtue, denial of negative qualities, and perhaps claims of personal superiority. Many indicators of PIM fail to consider one and sometimes two of these facets (Lanyon, 2004; Paulhus & John, 1998). In contrast, carelessness can be seen as the most grossly evident of the identified biases, because it requires no attention to the items or even the response alternatives and thus may be particularly amenable to detection.

Another possible reason for invalidity is the extent to which some bias indicators reflect some aspect of substantive variation rather than bias. To cite just several examples of a large literature revealing that bias indicators are sensitive to much more than bias, studies have suggested that PIM measures can be elevated by religiosity, emotional stability, or conscientiousness (Francis, Fulljames, & Kay, 1992; Martocchio & Judge, 1997; Ones et al. 1996), self-report bias indicators can be closely associated with general level of functioning (Morey et al., 2002), and neuropsychological measures of effort are sensitive to true cognitive impairment (Merten, Bossink, & Schmand, 2007). This problem of alternative substantive explanations for elevations on bias indicators is likely to be particularly acute when invalid responding is identified with a norm-referenced rather than an absolute standard (Medoff, 1999). It is difficult to account for a below-chance score on a SVT by any other means than purposeful misrepresentation, but the same is not the case when the score is just relatively low.

Another possibility is that moderators may be suppressing the validity of response bias indicators in certain situations. For example, Holden (2008) provided an example of a situation in which a floor effect on a measure completed under standard instructions interfered with further reduction of scores under instructions to underreport. Birkeland, Manson, Kisamore, Brannick, and Smith (2006) concluded that the degree to which job applicants attempted to distort their responses varied as a function of personality dimension, type of job, and type of test. The possibility that moderator effects account for some of the present results may merit further investigation.

### The Prevalence of Bias

It is commonly assumed among users of assessment instruments that biased responding is a common phenomenon, at least in certain applied settings. One survey of neuropsychologists suggested that, on average, 29% of personal injury cases, 30% of disability cases, 31% of chronic pain cases, and 19% of criminal cases referred for assessment involved malingering (Mittenberg, Patton, Canyock, & Condit, 2002), although Rogers and Correa (2008) concluded that these estimates were inflated by the inclusion of cases of suspected malingering and simple exaggeration. Other surveys have suggested suspected or definite malingering in over 20% of insanity evaluations (Rogers, 1986) and in more than 25% of head injury cases (Reynolds, 1998). Gouvier, Lees-Haley, and Hammer (2003) estimated the annual medical and legal costs associated with malingering at $5 billion.

The hypothesis that individuals often self-perceive or self-report inaccurately is popular even outside the applied assessment literature. Amador and colleagues (e.g., Amador & David, 1998) have described poor insight about their symptomatology as a common feature among individuals with psychotic disorders. Another body of literature has been devoted to inaccuracy in self-knowledge as a trait variable (e.g., Colvin et al., 1995; Taylor & Brown, 1994; Vogt & Colvin, 2005). Clearly, there is widespread faith among psychologists in the phenomenon of response bias. If this faith is misplaced and biased responding is relatively unusual in most settings, correlational results would be attenuated by low base rate.

Responding to test items is increasingly considered to be the product of a fairly complex cognitive process (Belli, Schwarz, Singer, & Talarico, 2000; Schwarz & Oyserman, 2001), and measurement error may occur for a variety of reasons besides purposeful or unintended bias. McCrae et al. (1998) found that out of 345 reliably classified instances of inconsistency between self-ratings

[6] One article cited under emotional disorders (Holden et al. 1989) also described a replication involving suicidality among prisoners. This study did not meet criteria for inclusion because the criterion was a self-report indicator. Also, the replication suffered the same flaw as the analysis described previously, specifically, failure to analyze the direction of the multiplicative term.

Table 5
*Summary of Articles: Forensic Populations*

| Study | Sample | Substantive indicators | Bias indicator | Response style | Criterion | Results |
|---|---|---|---|---|---|---|
| Edens & Ruiz (2006) | 349 inmates | PAI Antisocial Features scale | PAI PIM scale | PIM | Infractions | Moderation: 2 of 4 (50%) interactions significant[a] |

*Note.* PAI = Personality Assessment Inventory; PIM = Positive impression management.
[a] This outcome is consistent with the response bias hypothesis. However, change in $R^2$ is not a sufficient basis for concluding moderation is in the expected direction.

and spouse ratings, only 27 were attributed to factors that could be considered response bias. Unfortunately, this was the only study we found that attempted to identify reasons for inconsistency between substantive indicators and criteria.

One might note that the McCrae et al. (1998) sample was drawn from a normal population, and the rate of biased responding may be higher in settings where there is a motivation to malinger. However, there is some evidence to suggest that at least some individuals who are thought to be malingering in part because of context may, in fact, not be malingering. For example, Sayer, Spoont, and Nelson (2004) found that veterans seeking disability for posttraumatic stress disorder (a context in which the rate of malingering is thought to be very high) utilized more mental health services after obtaining disability benefits, a phenomenon that suggests the successful claimants were actually in greater need of mental health services (although the sample was biased in that it excluded individuals denied benefits). Similarly, Fishbain, Cutler, Rosomoff, and Rosomoff (1999) found that although physician estimates of the rate of malingering among chronic pain patients ranged as high as 75%, a review of 12 studies provided little evidence that malingering is a widespread phenomenon in this population. Research cited earlier that found factor structure invariance across respondents with and without a motivation to distort (e.g., Ellingson et al., 2001) is also consistent with the hypothesis that the rate of malingering has been overestimated, at least in the case of job applicants. However, research finding that individuals with a motivation to distort generate higher scores in disability evaluations (e.g., Green et al., 2001) is potentially inconsistent with this hypothesis unless another reasonable cause for these differences can be identified.

If applied assessors misestimate the prevalence of biased responding, this may reflect difficulties detecting true cases of malingering. Schacter (1986) found that clinicians instructed to identify malingerers often perform no better than chance. Various factors could potentially contribute to professionals' overestimating the rate of motivated distortion, including unintended blaming of the victim and suspicious behaviors by an individual with a motivation to distort, particularly if the individual suspects that the evaluator distrusts the claimant's motivations (Glenton, 2003). The potential for covariance misestimation is particularly strong if individuals who are clearly misrepresenting themselves tend to generate positive scores on bias indicators, even if they are only a small proportion of the population of positive responders (Arkes, 1981).

Even if the respondent is consciously attempting to manipulate the results, the respondent may be inconsistent about, or ineffective at, faking (Birkeland et al., 2006). For example, Amelang, Schäfer, and Yousfi (2002) found that self-reports completed under instructions to fake bad still correlated about .40 with peer ratings. It has been suggested that the ability to distort successfully might vary across individuals (e.g., Buller & Burgoon, 1994; Dunnigan & Nofi, 1995; Handel, 1989; LaFrenière, 1988; McFarland & Ryan, 2000; Snell, Sydell, & Lueke, 1999). The intention to fake may therefore be more common than successful faking, and it may be that the rarity of biased responding is not the issue so much as the rarity of successful biased responding.

Alternatively, the base rate of faking can be so high that the same outcome occurs. For example, Rosse et al. (1998) compared job applicants with incumbents and concluded that socially desirable responding is widespread among the former. Dunning, Heath, and Suls (2004) presented evidence that overly optimistic self-evaluation is widespread. Either a very high or a very low base rate will substantially reduce incremental validity attributable to moderator or suppressor effects in the types of analyses used to detect response bias in applied settings.

## Criterion Coarseness

Another possibility is that at least some of the criteria used in the studies described (e.g., spouse ratings, supervisor ratings, clinician ratings, injury severity) are too unreliable, too gross, too amenable to distortion themselves (e.g., clinician ratings based on patient statements), or too indirectly related to the constructs represented by the substantive indicators to facilitate validation of response bias indicators. An article by Shedler, Mayman, and Manis (1993) raised the possibility that physiological criteria would be particularly sensitive to defensive responding (see also Weinberger et al., 1979). We note, however, that this proposition leaves uncertain the relevance of bias to other types of criteria.

Although there is good reason to question the adequacy of criteria in psychology (e.g., Hough & Oswald, 2005), three responses to this explanation are appropriate. First, the same types of criteria are typically used to validate substantive indicators and do so quite well. If they are insufficient to serve the same purpose for bias indicators, then this suggests that bias effects must be of a much subtler nature, raising concerns over whether they can be detected with sufficient accuracy. Second, unless better criteria become available, this argument renders the validity of response bias indicators impervious to disproof. Third, this explanation is not sufficient for effects that are, on average, opposite in direction from expectation.

## The Case Against Measuring Bias

An important and unexpected finding from this review of the literature was how few studies in applied settings where bias

indicators are widely used, such as clinical settings where individuals with emotional disorders are treated or in neuropsychological assessment, have actually evaluated whether response bias indicators improve the criterion-related validity of substantive indicators. Admittedly the paucity of studies may partly reflect deficiencies in the search process. The literature on response bias is massive and often not readily identifiable as related to the topic. In particular, the exclusion of dissertations primarily on the basis of retrieval issues could have skewed the findings. Even so, given the thoroughness of the review, we believe it is unlikely that there exists a substantial corpus of studies involving assessment for emotional disorders, disability, or forensic purposes that were omitted despite meeting criteria for inclusion in this review and that, on balance, provide a strong case for the response bias hypothesis evaluated here. Furthermore, it is reasonable to expect that a publication bias exists in favor of studies that substantiate the response bias hypothesis, given (a) the well-known bias toward publishing significant effects (Dickersin, 2005) and (b) strong faith in the prevalence of response bias among psychologists conducting assessments in applied settings. Finally, a hypothesis supported largely by unpublished dissertations cannot be considered to have been supported sufficiently, particularly if the practical implementation of that hypothesis can have life-altering implications for test takers.

In the absence of sufficient evidence for or against the measurement of response bias, a reasonable argument can be made for continuing to use bias indicators. For example, consistent evidence that measures of effort such as SVTs account for a substantial proportion of variability in neuropsychological tests, combined with the absence of evidence that these results represent anything besides motivated self-misrepresentation, would seem to provide a strong circumstantial case that poor performance on SVTs must represent response bias. In addition, a false negative resulting from response bias in high-stakes testing can have potentially dangerous social consequences if, for example, a police officer or airline pilot is successfully able to disguise extremely suicidal or aggressive proclivities. It is also important to note, in support of this contention, that to our knowledge, no psychological measure has ever proved immune from distortion by motivated misrepresentation. These are important arguments to keep in mind when evaluating the utility of looking for bias in applied testing situations.

At the same time, it is important to consider that the cost of a false positive on a bias indicator is not necessarily negligible. A finding reported by Christiansen et al. (1994) highlights the degree to which the use of bias indicators to correct scores for biased responding can influence outcomes from an assessment process. They estimated that correcting for response bias changed the rank ordering of 85% of job applicants, although the actual number of applicants affected depended on the selection ratio (percentage of applicants hired). They correctly cautioned practitioners about the difficulty involved in defending the procedure in a court of law when the use of bias indicators does not affect criterion-related validity but does affect individual hiring decisions.

A more specific concern has to do with the extent to which bias indicators can potentially contribute to the adverse impact of an assessment procedure. Research consistently demonstrates that individuals from certain cultural minority groups tend to generate higher scores on PIM measures than do White participants (e.g., Dudley et al., 2005; Hough, 1998; Hough, Oswald, & Ployhart, 2001). Dudley et al. found that although response bias indicators did not enhance the validity of substantive indicators, minority candidates were disproportionately eliminated from consideration if scores on substantive indicators were corrected for response bias. Again, it is difficult to defend this procedure in a court of law when corrected scores that have an adverse impact are used for hiring decisions even though criterion-related validity is not changed by the procedure.

A third consideration is that of the financial and effort costs associated with a false positive on a bias indicator, especially when the substantive indicator is a particularly demanding one, such as the MMPI or an entire neuropsychological battery. McGrath and Ingersoll (1999) found across several MMPI studies that 14%–16% of potential participants were excluded because of elevated scores on bias indicators, a substantial waste of time and information if the majority of those cases represented false positives.

In contrast, the potential benefits are questionable. Simulations conducted in relation to work settings suggest that even if bias indicators are reasonably valid methods for identifying distorted responding, exclusion of cases on the basis of "invalid" responding would improve mean performance outcomes by only about 0.1 standard deviations (Schmitt & Oswald, 2006). From a utility perspective, it is difficult to justify the use of response bias indicators in applied personnel selection settings, except perhaps in work situations with significant public safety implications (police, airline pilots, etc.).

## Conclusions

At one time, researchers hypothesized that self-report indicators were hopelessly compromised by response bias (e.g., Edwards, 1957). Since then, attitudes have changed substantially as evidence has accrued questioning the role of response bias in respondent behavior. The present review raises concerns about the validity of bias indicators in those settings where sufficient research exists to draw a conclusion and the justification for their use in those settings where the research is insufficient. The strongest case can be made for the measurement of inconsistency, but the assessor must consider the degree to which random or careless responding is likely to be a factor in a particular testing situation. In addition, even in the case of inconsistent responding, the justification is based on a relatively small set of studies. The research implications of this review are straightforward: Proponents of the evaluation of bias in applied settings have some obligation to demonstrate that their methods are justified, using optimal statistical techniques for that purpose.

What is troubling about the failure to find consistent support for bias indicators is the extent to which they are regularly used in high-stakes circumstances, such as employee selection or hearings to evaluate competence to stand trial and sanity. If the identification of bias is considered essential, perhaps the best strategy would be to require convergence across multiple methods of assessment before it is appropriate to conclude that faking is occurring (Bender & Rogers, 2004; Franklin, Repasky, Thompson, Shelton, & Uddo, 2002). This would be particularly practical in the case of neuropsychological assessment, where a variety of relatively distinct methods have been developed for detecting insufficient effort.

The ultimate solution to the question of response bias remains elusive. What this review has demonstrated is that regardless of all

the journal space devoted to the discussion of response bias, the case remains open whether bias indicators are of sufficient utility to justify their use in applied settings to detect misrepresentation.

## References

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


Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94–107. doi:10.1037/0021-9010.90.1.94

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park CA: Sage.

*al'Absi, M., Bongard, S., & Lovallo, W. R. (2000). Adrenocorticotropin responses to interpersonal stress: Effects of overt anger expression style and defensiveness. *International Journal of Psychophysiology, 37,* 257–265. doi:10.1016/S0167-8760(00)00108-2

Allen, L. M., Conder, R. L., Green, P., & Cox, D. R. (1997). *CARB '97: Manual for the Computerized Assessment of Response Bias.* Durham, NC: CogniSyst.

Alperin, J. J., Archer, R. P., & Coates, G. D. (1995). Developme8nt and effects of an MMPI-A K-correction procedure. *Journal of Personality Assessment, 67,* 155–168. doi:10.1207/s15327752jpa6701

Amador, X. F., & David, A. S. (1998). *Insight in psychosis.* New York: Oxford University Press.

Amelang, M., Schäfer, A., & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions. *Psychologische Beitrage, 44,* 24–41.

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

*Archer, R. P., Fontaine, J., & McCrae, R. R. (1998). Effects of two MMPI-2 validity scales on basic scale relations to external criteria. *Journal of Personality Assessment, 70,* 87–102. doi:10.1207/s15327752jpa7001

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323–330. doi:10.1037/0022-006X.49.3.323

*Arnold, H., Feldman, D., & Purbhoo, M. (1985). The role of social-desirability response bias in turnover research. *Academy of Management Journal, 28,* 955–966. doi:10.2307/256249

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment, 14,* 16–26. doi:10.1037/1040-3590.14.1.16

Barger, S. (2002). The Marlowe–Crowne affair: Short forms, psychometric structure and social desirability. *Journal of Personality Assessment, 79,* 286–305. doi:10.1207/S15327752JPA7902

*Barger, S. D., Kircher, J. C., & Croyle, R. T. (1997). The effects of social context and defensiveness on the physiological responses of repressive copers. *Journal of Personality and Social Psychology, 73,* 1118–1128. doi:10.1037/0022-3514.73.5.1118

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

Barthlow, D. L., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & McNulty, J. L. (2002). The appropriateness of the MMPI-2 K correction. *Assessment, 9,* 219–222. doi:10.1177/1073191102009003001

Belli, R. F., Schwarz, N., Singer, E., & Talarico, J. (2000). Decomposition can harm the accuracy of behavioural frequency reports. *Applied Cognitive Psychology, 14,* 295–308. doi:10.1002/1099-0720(200007/08)14:4<295::AID-ACP646>3.0.CO;2-1

Bender, S. D., & Rogers, R. (2004). Detection of neurocognitive feigning: Development of a multi-strategy assessment. *Archives of Clinical Neuropsychology, 19,* 49–60. doi:10.1016/S0887-6177(02)00165-8

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

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14,* 317–329.

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence and social desirability in the MMPI.* New York: Appleton–Century–Crofts.

Borkenau, P., & Amelang, M. (1985). The control of social desirability in personality inventories: A study using the principal-factor deletion technique. *Journal of Research in Personality, 19,* 44–53. doi:10.1016/0092-6566(85)90036-4

*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

*Bowden, S. C., Shores, E. A., & Mathias, J. L. (2006). Does effort suppress cognition after traumatic brain injury? A re-examination of the evidence for the Word Memory Test. *The Clinical Neuropsychologist, 20,* 858–872. doi:10.1080/13854040500246935

Buller, D. B., & Burgoon, J. K. (1994). Deception: Strategic and nonstrategic communication. In J. A. Daly & J. M. Wiemann (Eds.), *Strategic interpersonal communication* (pp. 191–223). Hillsdale NJ: Erlbaum.

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R. L. Griffith, & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113–148). Greenwich, CT: Information Age.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2 (Minnesota Multiphasic Personality Inventory–2): Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Morfitt, R. C., Rouse, S. V., & Holden, R. R. (1997). Reducing MMPI-2 defensiveness: The effect of specialized instructions on retest validity in a job applicant sample. *Journal of Personality Assessment, 68,* 385–401. doi:10.1207/s15327752jpa6802

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59,* 143–178. doi:10.1111/j.1467-6494.1991.tb00772.x

*Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847–860. doi:10.1111/j.1744-6570.1994.tb01581.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Colby, F. (1989). Usefulness of the K correction in MMPI profiles of patients and nonpatients. *Psychological Assessment, 1,* 142–145. doi:10.1037/1040-3590.1.2.142

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152–1162. doi:10.1037/0022-3514.68.6.1152

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475–494.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability

independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354. doi:10.1037/h0047358

Derakshan, N., & Eysenck, M. W. (1998). Working memory capacity in high trait-anxious and repressor groups. *Cognition & Emotion, 12,* 697–713. doi:10.1080/026999398379501

*Dicken, C. (1963). Good impression, social desirability, and acquiescence as suppressor variables. *Educational and Psychological Measurement, 23,* 699–721. doi:10.1177/001316446302300406

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–34). Chichester, United Kingdom: Wiley.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 46,* 309–313. doi:10.1016/j.paid.2008.10.020

Draper, K., & Ponsford, J. (2008). Cognitive functioning ten years following traumatic brain injury and rehabilitation. *Neuropsychology, 22,* 618–625. doi:10.1037/0894-4105.22.5.618

Dudley, N. M., McFarland, L. A., Goodman, S. A., Hunt, S. T., & Sydell, E. J. (2005). Racial differences in socially desirable responding in selection contexts: Magnitude and consequences. *Journal of Personality Assessment, 85,* 50–64. doi:10.1207/s15327752jpa8501

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

Dunnigan, J. F., & Nofi, A. A. (1995). *Victory and deceit: Dirty tricks at war.* New York: Morrow.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest,* 5, 69–106. doi:10.1111/j.1529-1006.2004.00018.x

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16,* 1–23. doi:10.1207/S15327043HUP1601

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8,* 360–362. doi:10.1037/1040-3590.8.4.360

*Edens, J., & Ruiz, M. (2006). On the validity of validity scales: The importance of defensive responding in the prediction of institutional misconduct. *Psychological Assessment, 18,* 220–224. doi:10.1037/1040-3590.18.2.220

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

*Egloff, B., & Schmukle, S. C. (2003). Does social desirability moderate the relationship between implicit and explicit anxiety measures? *Personality and Individual Differences, 35,* 1697–1706. doi:10.1016/S0191-8869(02)00391-4

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

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

Evans, M. G. (1985). A Monte Carlo study of the effects of correlated method variance in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes, 36,* 305–323. doi:10.1016/0749-5978(85)90002-0

Exner, J. E., Jr. (2002). *The Rorschach: A comprehensive system: I. Basic foundations and principles of interpretation* (4th ed.). New York: Wiley.

*Eysenck, M. W., & Derakshan, N. (1999). Self-reported and other-rated trait anxiety and defensiveness in repressor, low-anxious, high-anxious, and defensive high-anxious groups. *Anxiety, Stress & Coping: An International Journal, 12,* 127–144. doi:10.1080/10615809908248326

Fan, J., Wong, C. C., Carroll, S. A., & Lopez, F. J. (2008). An empirical investigation of the influence of social desirability on the factor structure of the Chinese 16PF. *Personality and Individual Differences, 45,* 790–795. doi:10.1016/j.paid.2008.08.008

Fishbain, D. A., Cutler, R. B., Ros63omoff, H. L., & Rosomoff, R. S. (1999). Chronic pain disability exaggeration/malingering and submaximal effort research. *Clinical Journal of Pain, 15,* 244–274. doi:10.1097/00002508-199912000-00002

*Fishbain, D. A., Cutler, R. B., Rosomoff, H. L., & Steele-Rosomoff, R. (2002). Does the Conscious Exaggeration Scale detect deception within patients with chronic pain alleged to have secondary gain? *Pain Medicine, 3,* 39–46. doi:10.1046/j.1526-4637.2002.02002.x

Flaro, L., Green, P., & Robertson, E. (2007). Word Memory Test failure 23 times higher in mild brain injury than in parents seeking custody: The power of external incentives. *Brain Injury, 21,* 373–383. doi:10.1080/02699050701311133

Francis, L. J., Fulljames, P., & Kay, W. K. (1992). The functioning of the EPQ Lie scale among religious subjects in England. *Journal of Psychology and Christianity, 11,* 255–261.

Franklin, C. L., Repasky, S. A., Thompson, K. E., Shelton, S. A., & Uddo, M. (2002). Differentiating overreporting and extreme distress: MMPI-2 use with compensation-seeking veterans with PTSD. *Journal of Personality Assessment, 79,* 274–285. doi:10.1207/S15327752JPA7902

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51,* 115–134. doi:10.1037/0022-0167.51.1.115

*George, D. I., & Smith, M. C. (1990). Organizational assessment in personnel selection. *Public Personnel Management, 19,* 175–190.

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology, 19,* 475–487. doi:10.1016/j.acn.2003.05.001

Glenton, C. (2003). Chronic back pain sufferers—Striving for the sick role. *Social Science & Medicine, 57,* 2243–2252. doi:10.1016/S0277-9536(03)00130-8

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11,* 340–344. doi:10.1111/j.0965-075X.2003.00256.x

Gouvier, W. D., Lees-Haley, P. R., & Hammer, J. H. (2003). The neuropsychological examination in the detection of malingering in the forensic arena: Costs and benefits. In G. P. Prigatano & N. H. Pliskin (Eds.), *Clinical neuropsychology and cost outcomes research: A beginning* (pp. 405–424). New York: Psychology Press.

Green, P., Allen, L., & Astner, K. (1996). *The Word Memory Test: A user's guide to the Oral and Computer-Administered Forms, US Version 1.1.* Durham, NC: CogniSyst.

Green, P., Lees-Haley, P. R., & Allen, L. M., III. (2002). The Word Memory Test and the validity of neuropsychological test scores. *Journal of Forensic Neuropsychology, 2,* 97–124.

Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen, L. M., III. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15,* 1045–1060. doi:10.1080/02699050110088254

Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of Memory Malingering and Word Memory Test: A new comparison of failure concordance rates. *Archives of Clinical Neuropsychology, 23,* 801–807. doi:10.1016/j.acn.2008.07.005

Greve, K. W., Bianchini, K. J., Black, F. W., Heinly, M. T., Love, J. M., Swift, D. A., & Ciota, M. (2006). Classification accuracy of the Test of

Memory Malingering in persons reporting exposure to environmental and industrial toxins: Results of a known-groups analysis. *Archives of Clinical Neuropsychology, 21,* 439–448. doi:10.1016/j.acn.2006.06.004

Greve, K. W., Binder, L. M., & Bianchini, K. J. (2008). Rates of below-chance performance in forced-choice symptom validity tests. *The Clinical Neuropsychologist, 23,* 534–544. doi:10.1080/13854040802232690

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44,* 932–942. doi:10.1016/j.paid.2007.10.034

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69,* 192–203. doi:10.1037/h0025606

Handel, M. I. (1989). *Military deception in peace and war.* Jerusalem, Israel: The Magnes Press.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: I. Studies in deceit.* New York: Macmillan.

Hathaway, S. R., & McKinley, H. C. (1967). *Minnesota Multiphasic Personality Inventory Manual* (Rev. ed.). New York: Psychological Corporation.

Heilbrun, A. B. (1963). Revision of the MMPI K correction procedure for improved detection of maladjustment in a normal college population. *Journal of Consulting Psychology, 27,* 161–165. doi:10.1037/h0042150

*Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit–explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality, 19,* 25–49. doi:10.1002/per.537

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270–1285. doi:10.1037/0021-9010.92.5.1270

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

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences, 44,* 311–321. doi:10.1016/j.paid.2007.08.012

*Holden, R. R., Mendonca, J., & Serin, R. (1989). Suicide, hopelessness, and social desirability: A test of an interactive model. *Journal of Consulting and Clinical Psychology, 57,* 500–504. doi:10.1037/0022-006X.57.4.500

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11,* 209–244. doi:10.1207/s15327043hup1102&3

*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

Hough, L. M., & Oswald, F. L. (2005). They're right, well. . . mostly right: Research evidence and an agenda to rescue personality testing from 1960s insights. *Human Performance, 18,* 373–387. doi:10.1207/s15327043hup1804

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194. doi:10.1111/1468-2389.00171

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 62–97). New York: Academic Press.

*Jamner, L. D., Shapiro, D., Goldstein, I. B., & Hug, R. (1991). Ambulatory blood pressure and heart rate in paramedics: Effects of cynical hostility and defensiveness. *Psychosomatic Medicine, 53,* 393–406.

Johnson, J. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research, 39,* 273–302. doi:10.1207/s15327906mbr3902

*Kriedt, P. H., & Dawson, R. I. (1961). Response set and the prediction of clerical job performance. *Journal of Applied Psychology, 45,* 175–178. doi:10.1037/h0041918

*Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76,* 315–332. doi:10.1207/S15327752JPA7602

*Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences, 45,* 22–27. doi:10.1016/j.paid.2008.02.012

LaFrenière, P. (1988). The ontogeny of tactical deception in humans. In R. W. Byrne & A. Whiten (Eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans* (pp. 238–252). Oxford, United Kingdom: Clarendon Press.

Lalwani, A. K., Shrum, L. J., & Chiu, C.-Y. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology, 96,* 870–882. doi:10.1037/a0014622

Lanyon, R. I. (2004). Favorable self-presentation on psychological inventories: An analysis. *American Journal of Forensic Psychology, 22,* 53–65.

*Lee, S., & Klein, H. J. (2002). Relationships between conscientiousness, self-efficacy, self-deception, and learning over time. *Journal of Applied Psychology, 87,* 1175–1182. doi:10.1037/0021-9010.87.6.1175

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14,* 131–141. doi:10.1111/j.1468-2389.2006.00339.x

*Logan, D. E., Claar, R. L., & Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain, 136,* 366–372. doi:10.1016/j.pain.2007.07.015

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. doi:10.1037/1082-989X.7.1.19

Marshall, M. B., De Fruyt, F., Rolland, J. P., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO PI-R. *Psychological Assessment, 17,* 379–310810784. doi:10.1037/1040-3590.17.3.379

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology, 2,* 117–163. doi:10.1037/h0073583

*Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences of self-deception and self-efficacy. *Journal of Applied Psychology, 82,* 764–773. doi:10.1037/0021-9010.82.5.764

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390. doi:10.1037/0033-2909.114.2.376

*McCrae, R. R. (1986). Well-being scales do not measure social desirability. *Journal of Gerontology, 41,* 390–392.

*McCrae, R. R., & Costa, P. T. (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

McCrae, R. R., Costa, P. T., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine, 51,* 58–65.

*McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66,* 285–313. doi:10.1111/1467-6494.00013

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. (2001). Toward more clinically relevant assessment research. *Journal of Personality Assessment, 77,* 307–322. doi:10.1207/S15327752JPA7702

McGrath, R. E. (2008). Not all effect sizes are the same: Comments on Holden (2008). *Personality and Individual Differences, 44,* 1819–1823. doi:10.1016/j.paid.2008.01.015

McGrath, R. E., & Ingersoll, J. (1999). Writing a good cookbook: I. A review of MMPI high-point code system studies. *Journal of Personality Assessment, 73,* 149–178. doi:10.1207/S15327752JPA7302

*McGrath, R. E., Rashid, T., Hayman, J., & Pogge, D. (2002). A comparison of MMPI-2 high-point coding strategies. *Journal of Personality Assessment, 79,* 243–256. doi:10.1207/S15327752JPA7902

McKinley, J. C., Hathaway, S. R., & Meehl, P. E. (1948). The Minnesota Multiphasic Personality Inventory: VI. The K scale. *Journal of Consulting Psychology, 12,* 20–31. doi:10.1037/h0061377

Medoff, D. (1999). MMPI-2 validity scales in child custody evaluations: Clinical versus statistical significance. *Behavioral Sciences and the Law, 17,* 409–411. doi:10.1002/(SICI)1099-0798(199910/12)17:4<409::AID-BSL357>3.0.CO;2-N

*Melamed, S. (1996). Emotional reactivity, defensiveness, and ambulatory cardiovascular response at work. *Psychosomatic Medicine, 58,* 500–507.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29,* 308–318. doi:10.1080/13803390600693607

Meyer, G. (1997). On the integration of personality assessment methods: The Rorschach and the MMPI. *Journal of Personality Assessment, 68,* 297–330. doi:10.1207/s15327752jpa6802

Michaelis, W., & Eysenck, H. J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in real-life motivation. *Journal of Genetic Psychology, 118,* 223–234.

Millon, T., Davis, R., & Millon, C. (1997). *MCMI-III manual* (2nd ed.). Minneapolis, MN: National Computer Systems.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24,* 1094–1102. doi:10.1076/jcen.24.8.1094.8379

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual.* Odessa, FL: Psychological Assessment Resources.

Morey, L. C., Quigley, B. D., Sanislow, C. A., Skodol, A. E., McGlashan, T. H., Shea, M. T., Stout, R. L., Zanarini, M. C., & Gunderson, J. G. (2002). Substance or style? An investigation of the NEO-PI-R validity scales. *Journal of Personality Assessment, 79,* 583–599. doi:10.1207/S15327752JPA7903

Nelson, N. W., Sweet, J. J., & Demakis, G. J. (2006). Meta-analysis of the MMPI-2 Fake Bad scale: Utility in forensic practice. *Clinical Neuropsychologist, 20,* 39–58. doi:10.1080/13854040500459322

Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology, 9,* 501–552. doi:10.1016/0887-6177(94)90041-8

*O'Connor, B. P. (2006). Social desirability measures and the validity of self-reports: A comprehensive search for moderated relationships in five-factor model space. In S. P. Shohov (Ed.), *Advances in psychology research* (Vol. 40, pp. 39–73). Hauppauge, NY: Nova Science.

Olson, M. A., Fazio, R. H., & Hermann, A. D. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science, 18,* 287–291. doi:10.1111/j.1467-9280.2007.01890.x

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11,* 245–269. doi:10.1207/s15327043hup1102&3

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

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. (1998). *Manual for the Balanced Inventory of Desirable Responding (BIDR-7).* Toronto, Ontario, Canada: Multi-Health Systems.

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

*Pauls, C. A., & Stemmler, G. (2003). Repressive and defensive coping during fear and anger. *Emotion, 3,* 284–302. doi:10.1037/1528-3542.3.3.284

*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

Pinsoneault, T. B. (2007). Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols. *Psychological Assessment, 19,* 159–164. doi:10.1037/1040-3590.19.1.159

*Reid-Seiser, H. L., & Fritzsche, B. A. (2001). The usefulness of the NEO PI-R Positive Presentation Management scale for detecting response distortion in employment contexts. *Personality and Individual Differences, 31,* 639–650. doi:10.1016/S0191-8869(00)00168-9

Reynolds, C. R. (1998). *Detection of malingering during head injury litigation.* New York: Plenum Press.

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

Rogers, R. (1986). *Conducting insanity evaluations.* New York: Van Nostrand Reinhold.

Rogers, R., & Correa, A. A. (2008). Determinations of malingering: Evolution from case-based methods to detection strategies. *Psychiatry, Psychology and Law, 15,* 213–223. doi:10.1080/13218710802014501

Rogers, R., Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10,* 160–177. doi:10.1177/1073191103010002007

*Rohling, M. L., & Demakis, G. J. (2010). Bowden, Shores, & Mathias (2006): Failure to replicate or just failure to notice: Does effort still account for more variance in neuropsychological test scores than TBI severity? *The Clinical Neuropsychologist, 24,* 119–136.

Rorer, L. G. (1965). The great response style myth. *Psychological Bulletin, 63,* 129–156. doi:10.1037/h0021888

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

Sayer, N. A., Spoont, M., & Nelson, D. B. (2004). Disability compensation for PTSD and use of VA mental health care. *Psychiatric Services, 55,* 589. doi:10.1176/appi.ps.55.5.589

Schacter, D. L. (1986). On the relation between genuine and simulated amnesia. *Behavioral Sciences & the Law, 4,* 47–64. doi:10.1002/bsl.2370040104

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

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology, 89,* 623–642. doi:10.1037/0022-3514.89.4.623

Schmitt, N., & Oswald, F. J. (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

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 2,* 127–160. doi:10.1016/S1098-2140(01)00133-3

*Shapiro, D., Goldstein, I. B., & Jamner, L. D. (1995). Effects of anger/hostility, defensiveness, gender, and family history of hypertension on cardiovascular reactivity. *Psychophysiology, 32,* 425–435. doi:10.1111/j.1469-8986.1995.tb02093.x

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist, 48,* 1117–1131. doi:10.1037/0003-066X.48.11.1117

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

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219–242. doi:10.1016/S1053-4822(99)00019-4

Stevens, A., Friedel, E., Mehren, G., & Merten, T. (2008). Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. *Psychiatry Research, 157,* 191–200.

*Stokes, G. S., Hogan, J. B., & Snell, A. F. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. *Personnel Psychology, 46,* 739–762. doi:10.1111/j.1744-6570.1993.tb01567.x

Sweet, J. J., Wolfe, P., Sattlberger, E., Numan, B., Rosenfeld, J. P., Clingerman, S., & Nies, K. J. (2000). Further investigation of traumatic brain injury versus insufficient effort with the California Verbal Learning Test. *Archives of Clinical Neuropsychology, 15,* 105–113. doi:10.1016/S0887-6177(98)00153-X

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116,* 21–27. doi:10.1037/0033-2909.116.1.21

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 621–663.

*Tomaka, J., Blascovich, J., & Kelsey, R. M. (1992). Effects of self-deception, social desirability, and repressive coping on psychophysiological reactivity to stress. *Personality and Social Psychology Bulletin, 18,* 616–624. doi:10.1177/0146167292185012

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM).* New York: Multi-Health Systems.

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

Vogt, D. S., & Colvin, C. R. (2005). Assessment of accurate self-knowledge. *Journal of Personality Assessment, 84,* 239–251. doi:10.1207/s15327752jpa8403

*Weed, N. C., Ben-Porath, Y. S., & Butcher, J. N. (1990). Failure of Wiener and Harmon Minnesota Multiphasic Personality Inventory (MMPI) subtle scales as personality descriptors and as validity indicators. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 281–285. doi:10.1037/1040-3590.2.3.281

Weinberger, D. A., Schwartz, G. E., & Davidson, R. J. (1979). Low-anxious, high-anxious, and repressive coping styles: Psychometric patterns and behavioral and physiological responses to stress. *Journal of Abnormal Psychology, 88,* 369–380. doi:10.1037/0021-843X.88.4.369

Wooten, A. J. (1984). Effectiveness of the K correction in the detection of psychopathology and its impact on profile height and configuration among young adult men. *Journal of Consulting and Clinical Psychology, 52,* 468–473. doi:10.1037/0022-006X.52.3.468

Zaldívar, F., Molina, A., López Ríos, F., & Garc´ia Montes, J. (2009). Evaluation of alcohol and other drug use and the influence of social desirability: Direct and camouflaged measures. *European Journal of Psychological Assessment, 25,* 244–251. doi:10.1027/1015-5759.25.4.244

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168–190. doi:10.1177/1094428104263674

Received July 5, 2008
Revision received February 8, 2010
Accepted February 11, 2010 ■