PERSONNEL PSYCHOLOGY
2007, 60, 683–729

RECONSIDERING THE USE OF PERSONALITY TESTS IN PERSONNEL SELECTION CONTEXTS

FREDERICK P. MORGESON
Michigan State University

MICHAEL A. CAMPION
Purdue University

ROBERT L. DIPBOYE
University of Central Florida

JOHN R. HOLLENBECK
Michigan State University

KEVIN MURPHY
Pennsylvania State University

NEAL SCHMITT
Michigan State University

Although long thought to be unrelated to job performance, research in the early 1990s provided evidence that personality can predict job performance. Accompanying this research was a resurgence of interest in the use of personality tests in high-stakes selection environments. Yet there are numerous potential problems associated with the current operational use of personality. As such, 5 former journal editors from *Personnel Psychology* and the *Journal of Applied Psychology* (2 primary outlets for such research), who have collectively reviewed over 7,000 manuscripts and who have no vested interest in personality testing, reconsider the research on the use of personality tests in environments where important selection decisions are made. Their comments are based on a panel discussion held at the 2004 SIOP conference. Collectively, they come to several conclusions. First, faking on self-report personality tests cannot be avoided and perhaps is not the issue; the issue is the very low validity of personality tests for predicting job performance. Second, as such, using published self-report personality tests in selection contexts should be reconsidered. Third, personality constructs may have value for employee selection, but future research should focus on finding alternatives to self-report personality measures.

At the 2004 Society for Industrial and Organizational Psychology conference in Chicago, a panel discussion was held in which current and former journal editors of *Personnel Psychology* and the *Journal of Applied*

Correspondence and requests for reprints should be addressed to Frederick P. Morgeson, The Eli Broad Graduate School of Management, Michigan State University, N475 N. Business Complex, East Lansing, MI, 48824-1122; morgeson@msu.edu.

Psychology were assembled to discuss the issue of faking in personality testing. The resulting discussion evolved from a focus on faking to a focus on the broader issue of whether the use of personality tests to make high-stakes employment decisions could be justified. With the exception of one of the panelists, there was consensus that there are a number of significant problems associated with the use of self-report personality tests in selection contexts and that perhaps a reappraisal of this literature might be in order. This paper briefly describes the history of the use of personality tests in industrial and organizational (I-O) psychology and specifically addresses the issue of fakability. The paper then summarizes and expands upon the points made during the panel discussion.

A Brief History on the Use of Personality Tests in I-O Psychology

In 1965, Guion and Gottier summarized 12 years (1952–1963) of research published in the *Journal of Applied Psychology* and *Personnel Psychology* on the use of personality tests in selection contexts. They concluded that "It is difficult in the face of this summary to advocate, with a clear conscience, the use of personality measures in most situations as a basis for making employment decisions about people" (Guion & Gottier, 1965, p. 160). This view prevailed for more than 25 years, until the 1991 publication of two meta-analyses on the validity of personality tests for personnel selection (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Although these two summaries found similar levels of validity as previous quantitative reviews (cf. Schmitt, Gooding, Noe, & Kirsch, 1984), they concluded that meta-analytically corrected estimates of validity were meaningful, and personality measures should once again be used in selection contexts. Based on this evidence, some have concluded, "From a practical standpoint, recent findings. . .suggest that the controversy over whether personality tests can be useful for prediction of employee performance is no longer pertinent" (Kanfer, Ackerman, Murtha, & Goff, 1995, p. 597).

As a consequence, interest in and research on personality testing in I-O psychology has dramatically increased in the past 10 years (Barrick & Ryan, 2003; Hough & Furnham, 2003; Schmitt, Cortina, Ingerick, & Wiechmann, 2003). Figure 1 demonstrates this increase by graphing the number of articles published in the *Journal of Applied Psychology* and *Personnel Psychology* and presentations made at the annual conferences of the Society for Industrial and Organizational Psychology on personality from 1985 to 2005. As can be seen, there was almost no research interest in personality in these outlets during the late 1980s, some emerging interest in the early 1990s, and then an explosive growth in interest starting a few years after the meta-analyses (in 1995). The interest continues to be

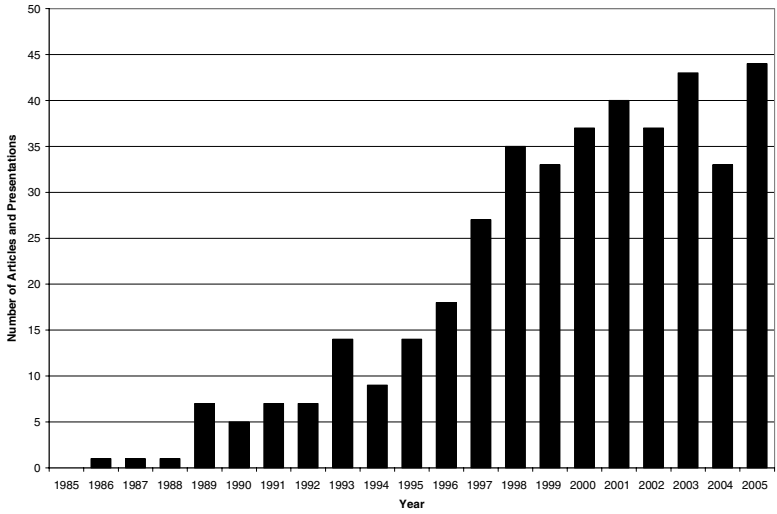

Figure 1: **Journal Articles and Presentations on Personality Research Related to Selection or Job Performance.**

strong at the end of this period in 2005. Thus, as one examines the history of research on personality tests in employment contexts, there has been a dramatic shift in research activity, suggesting changing opinions on the appropriateness of their use.

In addition to questions about their criterion-related validity, another criticism of the use of personality tests is that they are potentially open to being faked by motivated applicants. This is because the "right" or most positive answer may be apparent to the candidates. The only thing preventing candidates from providing the positive answer when it is not true is their own honesty or lack of self insight, neither of which can be assumed when there is a desirable outcome at stake (like getting a job). Furthermore, it is likely that some people fake more than others in particular situations, meaning that faking might result from an interaction between person and situation characteristics. This implies that faking might do more than simply elevate test scores; it might change the rank-order of examinees, and it might lead to different rank orders in different situations even if the pool of examinees is kept constant. In contrast, with cognitive ability tests, candidates cannot fake the right answer by simply wanting to do so. They must have the ability to recognize the right answer.

This concern has led researchers to provide warnings about personality tests. As Guion and Cranny (1982) noted, "We suspect that the influence of motivational variables is much greater with interest and personality inventories than with tests of cognitive abilities" (p. 242). For example,

validities based on current employees may differ from those based on candidates because the latter have more of an incentive to fake.

Some research conducted on the issue of faking in personality testing suggested that although faking may occur, it does not affect the validity of the test. For example, Hough, Eaton, Dunnette, Kamp, and McCloy (1990) found that less than a third of comparisons between the validity of "accurate" and "overly desirable" respondents were significantly different. Others have used different techniques, such as controlling for impression management and self-deceptive enhancement (Barrick & Mount, 1996; Cunningham, Wong, & Barbee, 1994), and examining the impact of social desirability on personality factor structures (Ellingson, Smith, & Sackett, 2001). All have concluded that faking has a minimal impact.

Other research, however, has been less optimistic and suggests that faking may have a pronounced impact on selection decisions. For example, some research has found that although no differences in validity are found, faking can significantly affect hiring decisions (Christiansen, Goffin, Johnston, & Rothstein, 1994; Rosse, Stecher, Miller, & Levin, 1998). This research has shown that faking will be more problematic as selection ratios decrease and if top-down selection is used. That is, different people will be hired due to faking. Other research has indicated that traditional covariate techniques are ineffective at partialling out intentional distortion (Ellingson, Sackett, & Hough, 1999), and different underlying constructs are measured across testing situations (i.e., applicant vs. nonapplicant samples; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001).

Given the current popularity of personality in selection contexts and the potential for dishonest responding, one might wonder, are we wrong in concluding personality tests are useful in selection contexts? A particularly valuable perspective on these issues might come from the editors of I-O psychology journals. There are several reasons why an editor's perspective is valuable. First, in their role as the final arbiters between authors and reviewers on article publication decisions, editors typically take a balanced perspective. Editors are more likely to consider the strengths and weaknesses of personality tests when compared with the wide range of other selection alternatives, whereas most authors in this area focus on one or two specific types of measures, leading to debates between advocates of particular types of selection measures. Second, editors will have reviewed great numbers of both published and unpublished research papers. As such, they are highly experienced in terms of judging the quality of scientific research and, therefore, have a broader and more comprehensive perspective than most. Third, editors have a fairly well developed understanding as to the strengths and weaknesses of both published and unpublished research. This understanding allows them to critically evaluate the current state of the literature on faking in personality testing.

The purpose of the this paper is to report on a panel discussion that was conducted on this topic at a recent (2004) conference of the Society for Industrial and Organizational Psychology (Morgeson et al., 2004). Five recent editors from *Personnel Psychology* and *Journal of Applied Psychology* participated on this panel. These two journals were selected because they are the primary outlets for research on this topic and they are the most subscribed journals by I-O psychologists. (A sixth editor was included on the panel, but she was a current editor of this journal and thus understandably decided not to participate on this paper.) The editors, their journals, and their terms are listed below:

Michael Campion, *Personnel Psychology*, 1990–1996
Robert Dipboye, (Associate Editor) *Journal of Applied Psychology*, 1996–2002
John Hollenbeck, *Personnel Psychology*, 1996–2002
Kevin Murphy, *Journal of Applied Psychology*, 1996–2002
Neal Schmitt, *Journal of Applied Psychology*, 1989–1994

The journal editors were used for this panel for two reasons. First, they have reviewed and made publication decisions on over 7,000 manuscripts, including a great many on the topic of personality testing; hence, they are in a good position to judge the cumulative personality research. Second, although these editors are among the most widely published authors in the field of I-O psychology, and some have published articles using personality measures, none of them have a primary research stream on personality testing or could be considered an advocate or a critic of personality testing in the published literature. Therefore, they are relatively impartial.

This paper has been reorganized from the panel discussion in order to more clearly explicate the main issues and questions addressed. Nine specific questions will be addressed:

(a) What is the current state of the literature on faking in personality tests?
(b) What is the criterion-related validity of personality measures?
(c) What effect does faking have in personality measurement?
(d) How can personality tests be used in a more effective way?
(e) What are some of the other concerns about faking?
(f) Can we detect faking in personality tests?
(g) What recommendations would you give about the use of personality tests in selection?
(h) Should personality research focus on an expanded criterion domain?
(i) What kind of future research should be done?

In addition, panelists were given the opportunity to make minor revisions and additions for the purposes of clarifying their original remarks.

The responses of each editor are indicated by name as appropriate under each topic.

One final point needs to be made about the nature of the comments made by the panelists. This paper (and panel session that the paper is drawn from) is designed to gather the opinions of a diverse set of former journal editors (whose opinions are anchored in their experience in the field of I-O psychology). Given their different perspectives, the panelists sometimes reach different conclusions on a particular issue. The goal of this paper is not necessarily to present a uniform or agreed-upon position and always come to common conclusions. Instead, the differences in interpretations and opinions reflect some of the diversity in opinion in this area of research and practice. We comment on some of these differences throughout, but it is not the intention of this paper to resolve potential areas of disagreement among the panelists.

Operational Issues

Issue 1: What Is the Current State of the Literature on Faking in Personality Tests?

Michael Campion. I would like to briefly summarize the research literature on faking in personality tests. I reviewed the literature in the normal fashion (e.g., computer search, manual search, cross-referencing, and so on) and I found 112 articles on the topic up to May of 2002. I summarized the literature in terms of 10 questions. This is shown in Table 1, but I will briefly restate the key findings here.

(a) Can people distort their responses (i.e., fake)? There were 39 studies on this topic with virtually all finding that scores could be distorted.
(b) Do people distort their responses? There were 14 studies, usually comparing applicants to incumbents, with 7 finding distortion and more finding distortion but not as much as expected. I should also note that directed faking studies where subjects are instructed to fake show much greater effects of faking than studies of real applicants, suggesting that real applicants do not fake as much as they could.
(c) Is there variance in the amount of distortion? There were three studies and all found that people differ in the amount they distort.
(d) Does the testing situation affect the amount of distortion? There were seven studies, with five finding that the situation matters (e.g., setting, method of administration).
(e) Is there a 2-factor model of faking (i.e., self-deception vs. impression management)? There were four studies, with three finding support for the model.

TABLE 1

Summary of Literature Review on Faking in Personality Tests

Topic	No. (%)	Findings
1. Can people distort their responses (i.e., fake)?	39 (34.8)	• Virtually all articles found that scores could be altered by response distortion, such as by instructing subjects to fake. (4, 5, 7, 8, 11, 13, 16, 18, 19, 21, 27, 30, 40, 47, 48, 53, 57, 59, 62, 72, 73, 75, 76, 77, 79, 81, 82, 84, 85, 86, 88, 89, 92, 94, 97, 100, 101, 105, 107)
2. Do people distort their responses (e.g., applicant versus incumbent)?	14 (12.5)	• 7 articles reported that applicants score higher than non applicants. (3, 11, 13, 23, 71, 102, 109) • 4 articles found that faking does occur, though not as much as previously thought or even minimally. (40, 47, 88, 89) • 3 articles found applicant and non applicant scores to be similar. (1, 48, 83)
3. Is there variance in the amount of distortion?	3 (2.7)	• All articles found variation in the amount of distortion by individuals. (6, 11, 79)
4. Does the testing situation (e.g., method of admin., familiarity of settings, clinical or field, etc.) affect the amount of distortion?	7 (6.3)	• 5 articles reported that the testing situation had some sort of impact. (27, 60, 66, 107, 110) • 2 articles reported no impact from testing situation. (1, 3)
5. Is there a two-factor model of faking (self-deception versus impression mgmt)?	4 (3.6)	• 3 of 4 articles found that a two-factor model was useful for classifying types of response distortion. (36, 56, 69)
6. Does faking affect criterion-related validity?	17 (15.2)	• 8 articles indicate that response distortion did affect the validity of the measures. (11, 12, 19, 64, 80, 93, 98, 101) • 10 articles indicate that faking did not appear to affect the validity of the measures. (10, 13, 20, 21, 26, 31, 40, 51, 70)
7. Does faking affect factor structure?	4 (3.6)	• 2 articles found an effect on factor structure. (3, 7) • 2 articles found no effect. (1, 2)

(Continued)

TABLE 1
(Continued)

Topic	No. (%)	Findings
8. Can response distortion be detected?	33 (29.5)	• 12 articles reported varying success with identifying distortion. (23, 25, 28, 34, 35, 46, 76, 77, 78, 94, 99, 111) • 7 articles reported that distortion could not be adequately detected. (59, 64, 72, 74, 75, 92, 106) • 14 articles proposed various methods that may be useful for detecting distortion. (7, 18, 29, 45, 52, 54, 68, 71, 82, 87, 91, 95, 105, 112)
9. Can response distortion be mitigated?	10 (8.9)	• 3 articles reported some amount of success (39, 55, 86) • 2 articles reported little success (9, 87) • 2 articles reported mixed results (7, 58) • 3 articles recommended methods for mitigating distortion (20, 57, 98)
10. Other questions	24 (21.4)	(12, 14, 15, 22, 24, 32, 33, 37, 38, 41, 42, 43, 44, 47, 49, 50, 64, 66, 67, 90, 99, 103, 104, 108)

Note. 112 articles were identified up to May of 2002. Some studied multiple questions. The specific articles addressing each topic are shown in parentheses, which correspond to numbers shown in parentheses after each reference in the reference list.

(f) Does faking affect criterion-related validity? There were 18 studies, with 8 findings that it does and 10 findings that it does not.
(g) Does faking affect factor structure? There were four studies, with two finding an effect and two not.
(h) Can response distortion be detected? There were 33 studies, with 12 finding some success, 7 finding no success, and 14 proposing methods that *may* be useful for detecting distortion.
(i) Can response distortion be mitigated? There were 10 studies, with 3 finding some success, 2 finding mixed success, 2 finding no success, and 3 others offering advice for mitigating distortion.
(j) There were 13 studies that explored other questions.

Four overall conclusions can be drawn from this review of the research literature on faking in personality tests. First, the total number of studies on the topic is large, suggesting that faking has been viewed as an important problem. Second, people can and apparently do fake their responses on personality tests. Third, almost half the studies where criterion-related validity was studied found some effect of faking on criterion-related validity. Fourth, there has been substantial research devoted to techniques for detecting and mitigating faking, but no techniques appear to solve the problem adequately.

Robert Dipboye. I believe we have to begin this discussion with a clear definition of what we mean by faking on a personality inventory. Paul Ekman (2001) defines a lie or deceit as a deliberate attempt by an individual to mislead another person by concealing or falsifying information. Faking on a personality test is usually described as lying but in this context it would seem to assume three preconditions. First, the individual taking the personality test has sufficient self-insight to accurately describe him or herself on an item. Second, the individual taking the test understands the question and interprets it as the author of the test intended. Third, the individual taking the test willfully and consciously deviates from the truth in the answer to the question so as to create a desired impression. Outright lies do occur, but I suspect that they do not account for the low validities so often found for personality tests. Guion (1965) identified "attitudinal variables that may systematically influence scores" that "render personality measures meaningless" (p. 357). He included in this list those who are honest but lack self-insight and those who desire to present an idealized concept of self in addition to those who engage in "out-and-out faking" (p. 357).

It seems important to include these other types of responders in addition to the "liars" when attempting to identify the distortions and noise that prevent the valid measurement of personality. There are test takers

who do not understand the question or define it in a manner that differs from the intent of the author of the test. When I have given students in my classes the major personality tests that we use in personnel selection, many express confusion on how they are to interpret the items. When the response alternatives force the individual into extremes (e.g., true or false) or contain indeterminate anchors such as *many*, *often*, or *sometimes*, the confusion and the protests are especially pronounced. Take, for example, items such as "I seldom toot my own horn," "I read a lot," "I don't talk a lot," "Rarely get irritated," "I often get headaches," "I am always prepared," or "I seldom get emotional." In other cases there is confusion about not only the response alternatives but also the item itself such as "I don't understand things," "I speak ill of others," "I get annoyed by others' behavior," or "I worry about things." The common request is for a definition of "things" and "others," the situation to which the item refers, and the meaning of seldom, often, always, and rarely. There are also those who know what is meant by the question (or at least think they do) but provide an answer that is aspirational. In response to the question "I am a leader," individuals might honestly see themselves as having the potential to become leaders even if they have not had the opportunity to demonstrate their leadership abilities. Is it faking to project an image that one sincerely believes is possible to achieve? I think not. There are also those who understand the question and know the answer that is the "correct" self-report but who believe that the employer or psychologist has no right to that information. I could expand this list. The primary point here is that self-report personality tests provide very poor indicators of the underlying constructs because of failures in test content and the testing context. On the basis of my own reading of the literature and experience with personality testing, I would hypothesize that outright lying is a small and insignificant part of a much larger problem.

A lot of effort has been wasted in futile attempts to identify the "fakers" so as to adjust their scores. It seems fair to conclude that these attempts have failed and may even lessen rather than improve the predictive potential of the measures. I find especially offensive the conclusion that has been drawn from some research that so-called faking is related to the underlying conscientiousness or integrity of the person. Given the nature of personality tests and the testing context, what we get is often what we deserve, and we should not blame the respondent for our failures. Alternatives to traditional self-report such as forced choice and conditional reasoning appear promising. I suspect that these are unlikely to yield permanent, realistic solutions in mass testing situations. The time, effort, and cost of developing such measures and the difficulty of maintaining their security are likely to remain as major impediments.

I suggest a more radical approach to measuring many of our noncognitive constructs. When measuring maximum performance such as in ability testing and work sampling, keeping the examinee uninformed of the nature of the test and closely monitoring for cheating seem unlikely to detract from the validity of measurement. There may be other adverse consequences (e.g., harm to recruiting) when examinees are treated as persons not to be trusted, but when maximum performance is being tested, faking is not one of them. In attempting to measure through self-report what people "typically" do, however, the accuracy of the tests depends on the cooperation of the examinees and their willingness to disclose. In these situations, I suggest that we need to engage them in a more open process where we disclose what we are looking for and gain the trust of test takers rather than playing paper-and-pencil games with them. An open process might include telling applicants the constructs we are attempting to measure and the items that are indicators of these constructs. It might also include providing them an opportunity to elaborate on their responses to inventory items. In the end, such openness may not be the solution to the low validities obtained for personality tests. But I find it hard to believe that such options could worsen the already low validities and am convinced that this approach would provide a more honest and ethical approach to using personality measures for selection.

Issue 2: What Is the Criterion-Related Validity of Personality Measures?

Kevin Murphy. I am less concerned with faking because I think there are much more serious problems than faking that we need to worry about. The problem with personality tests is not faking; it is that the validity of personality measures as predictors of job performance is often disappointingly low. A couple of years ago, I heard a SIOP talk by Murray Barrick that pulled together several meta-analytic measures of the Big Five, which is thought to be a fairly comprehensive and well-accepted taxonomy of normal personality (these five factors are often labeled Neuroticism, Extroversion, Openness to Experience, Agreeableness, and Conscientiousness; Conscientiousness is the factor that has been shown to be correlated with performance in the widest range of jobs). He said, "If you took all the Big Five, measured well, you corrected for everything using the most optimistic corrections you could possibly get, you could account for about 15% of the variance in performance." I think that is an overestimate, but even if the figure is correct, it is not very impressive. You are saying that if you take normal personality tests, putting everything together in an optimal fashion and being as optimistic as possible, you'll leave 85% of the variance unaccounted for. The argument for using personality tests to

predict performance does not strike me as convincing in the first place. In contrast, for example, simple measures of general cognitive ability are thought to account for about 20–25% of the variance in job performance (Schmidt & Hunter, 1998); the figure is likely higher in more complex jobs.

Because personality inventories overlap only minimally with measures of cognitive ability, there is some potential for these inventories to provide some incremental validity, and they may therefore have some practical value. Nevertheless, it does seem to me that it is very important to understand that an optimal combination of the information from a presumably comprehensive taxonomy of broad personality factors just does not tell us much about who will perform well or poorly on the job. Common sense tells you that broad personality factors *should* be important, but the data suggest that they are not.

Neal Schmitt. One of the things that Kevin said struck a chord with me is that if we have a personality measure and are concerned about faking, faking only makes a difference if the measure is valid. In 1965, Guion and Gottier (1965) published a paper in which the average validity of personality tests was .09. Twenty-five years later, Barrick and Mount (1991) published a paper in which the best validity they could get for the Big Five was .13. They looked at the same research. Why are we now suddenly looking at personality as a valid predictor of job performance when the validities still haven't changed and are still close to zero? One of the reasons is that we have applied corrections for range restriction, criterion unreliability, and, in some instances, predictor unreliability, which are fine if you are interested in looking at the constructs and their relationships to other constructs. But the point of fact is that when we go to use personality tests, we do not correct those scores. We use the observed scores and reliability and validity information for observed scores should be considered when examining the impact of faking.

Some might argue that the validity of personality/integrity tests is substantial (Ones, Viswesvaran, & Schmidt, 1993). I am very familiar with the paper by Ones et al. (1993), because I did deal with it during the editorial process. That set of data, if you look just at the average validity (.37), is very much inflated for the simple reason that many of the integrity tests involved integrity items that asked individuals to essentially admit theft or dishonest behavior, and the outcome was an admission of theft or behavior. The integrity test and the criteria used to validate the test are one and the same; that is, self-report admissions of wrongdoing. In these cases, the correlations reported are more like test–retest reliability estimates than criterion-related validity coefficients. If you examine Table 11 in the meta-analysis provided by Ones et al. (1993), you will see one

cell in that table that includes predictive studies of applicants with an externally measured criterion. That cell includes a summary of seven studies with an average validity of .09 (corrected value equal to .13) and a 90% credibility interval that includes zero. I realize that predictive studies in which theft criteria are collected are very difficult to conduct, but I do not believe we should mislead clients with "inflated" estimates of predictive validity using studies that are based on self-report criteria or concurrent studies. The latter are the nature of most of the validity coefficients reported in that table. So, I do not think you can just look at this whole body of literature uncritically, come up with an average value, and say that represents the predictive validity of integrity tests. I think the validity of integrity tests against objective theft measures is much lower than the average values reported in the paper by Ones et al. (1993). To quote Ones et al. (1993) "when the criterion was the much narrower one of (externally measured) theft alone, the mean observed validity from predictive studies conducted on applicants was considerably smaller at .13" (p. 691). When a broader set of outcomes is considered or the results of studies employing other research designs or participants, the validities are substantially higher (in the .20s), but I suspect that when most company personnel think of using integrity tests they are hoping to avoid theft, sabotage, or other serious counterproductive behavior. The validities for this narrower set of criteria are substantially lower and may in fact be zero, if we consider the confidence interval reported by Ones et al. (1993) for these studies.

Kevin Murphy. I am going to disagree to some extent and agree to some extent. I think the integrity test validities are a little higher than Neal thinks. I think they are a little lower than Deniz Ones thinks. The problem is that it is a grab bag. If you look at these tests, what they include, and what they are all about, it is hard to say anything coherent about integrity. There is no coherent theory about what these tests are supposed to measure. The fact that some of them do pick up not only counterproductivity criteria but also supervisory evaluation performance criteria is interesting and important—but I think quite a bit of work is needed to figure out why these tests work. There was a time when we thought these were personality measures that tapped personality dimensions that were well understood, like Conscientiousness. I do not really think that is true. I think that Deniz's work and the work that I have done shows some of the same things—that these measures are a morass. Integrity test scores seem to be related to several of the Big Five dimensions that are supposedly distinct and nearly orthogonal—and it is not clear to me what integrity tests are really measuring. So I think putting these tests in the same grab bag (by using meta-analysis to combine validity estimates) makes it look

like we are going back to Bob Guion's 1965 paper where he said "There is a whole bunch of stuff here and we do not know exactly what it is all about. Some of them work sometimes and we do not quite know why, and some don't work at all." It is déjà vu all over again.

Michael Campion. I have two related concerns about criterion-related validity. First, I reviewed the meta-analyses that have been published on personality tests and cognitive tests (the strongest predictor of job performance) in order to compare their validities. This comparison highlights the overall validity of two commonly used tests and enables a relativistic comparison of the two types of tests. This review found 13 meta-analyses for cognitive tests and 12 meta-analyses for personality tests. Some studies focused only on cognitive or personality tests, whereas others included both types of tests.

The meta-analyses are summarized in Tables 2 and 3, and Table 4 provides an overall summary in terms of medians and ranges. For cognitive ability measures predicting proficiency criteria, the median uncorrected validity was .20 with an interquartile range (i.e., middle 50% of the values) of .16–.25. The median corrected validity was .40 with an interquartile range of .28–.52. For predicting training criteria, the median uncorrected validity was .38 with an interquartile range of .30–.42, and the median corrected validity was .62 with an interquartile range of .54–.67. For personality measures predicting proficiency criteria, the median uncorrected validity was .10 with an interquartile range of .05–.17. The median corrected validity was .18 with an interquartile range of .09–.26. For predicting training criteria, the median uncorrected validity was .11 with an interquartile range of .08–.15, and the median corrected validity was .23 with an interquartile range of .10–.30.

As this shows, the validity of personality tests is typically quite low, supporting the points made by Kevin and Neal concerning the generally low validity of personality tests. I should point out, however, that personality tests could potentially add a small amount of incremental validity to a battery of cognitive tests. The only caveat is to make sure they do not replace the cognitive tests or get an inordinate amount of weight in the composite score or hiring decision because this would drastically reduce the overall validity.

In addition, several other observations are worth making. The corrections applied in the meta-analyses of cognitive ability measures are typically only for range restriction and criterion unreliability. The corrections applied in the meta-analyses of personality measures commonly include range restriction, criterion unreliability, *and* predictor unreliability. Correcting for predictor unreliability is typically *not* done for employment

TABLE 2

Summary of Meta-Analyses Investigating the Validity of General Cognitive Ability Tests

			Proficiency criteria			Training criteria		
Author	Job type(s)	Test/Construct type	Total *N*	$\bar{r}^a$	$\bar{\rho}^b$	Total *N*	$\bar{r}$	$\bar{\rho}$
Barrett, Polomsky, & McDaniel (1999)	Firefighters	Cognitive tests	2791	.20	.42	2007	.50	.77
Vinchur, Schippmann, Switzer, & Roth (1998)	Sales people	Overall cognitive ability	1,770	.18	.31	–	–	–
		General cognitive	1,231	.23	.40	–	–	–
		Verbal ability	597	.08	.14	–	–	–
Levine, Spector, Menon, Narayanan, & Cannon-Bowers (1996)	Craft jobs in utility industry	Cognitive	12,504	.25	.43	5,872	.38	.67
Hirsh, Northrop, & Schmidt (1986)	Law enforcement occupations	Memory	3,028	.05	.11	801	.22	.40
		Reasoning	3,175	.08	.18	4,374	.33	.57
		Verbal	2,207	.09	.22	3,943	.37	.62
Hunter (1986)	High complexity	GATB	2,455	–	.58	1,863	–	.50
	Medium complexity		12,933	–	.51	3,823	–	.57
	Low complexity		14,403	–	.40	575	–	.54
	Setup		1,114	–	.56	235	–	.65

(Continued)

TABLE 2
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
	Feeding/offbearing		1,219	–	.23	–	–	–
	Mechanical	General cognitive ability	–	–	–	156,143	–	.62
	Clerical		–	–	–	42,832	–	.58
	Electronic		–	–	–	92,758	–	.67
	General technical		–	–	–	180,806	–	.62
(Reanalysis of Ghiselli, 1973)	Manager	General cognitive ability	–	–	–	N/A	–	.51
	Clerk		–	–	–	N/A	–	.71
	Salesperson		–	–	–	N/A	–	–
	Protective professions worker		–	–	–	N/A	–	.87
	Service worker		–	–	–	N/A	–	.66
	Trades and crafts worker		–	–	–	N/A	–	.65
	Elementary industrial worker		–	–	–	N/A	–	.61
	Vehicle operator		–	–	–	N/A	–	.37
	Sales clerk		–	–	–	N/A	–	–
Hunter & Hunter (1984) (Reanalysis of Ghiselli, 1973)	Manager	General cognitive ability	N/A	–	.53	–	–	–
	Clerk		N/A	–	.54	–	–	–

(Continued)

TABLE 2
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
	Salesperson		N/A	–	.61	–	–	–
	Protective professions worker		N/A	–	.42	–	–	–
	Service worker		N/A	–	.48	–	–	–
	Trades and crafts worker		N/A	–	.46	–	–	–
	Elementary industrial worker		N/A	–	.37	–	–	–
	Vehicle operator		N/A	–	.28	–	–	–
	Sales clerk		N/A	–	.27	–	–	–
Schmitt, Gooding, Noe, & Kirsch (1984)	Average across 6 occupational groups	General mental ability	3,597	.22[c]	–	–	–	–
Schmidt, Hunter, & Caplan (1981)	Petroleum Industry Operator	General intelligence	1,486	.18	.26	–	–	–
	Petroleum Industry Maintenance	General intelligence	821	.21	.30	–	–	–
Pearlman, Schmidt, & Hunter (1980)	Clerical	General mental ability	17,539	.26	.52	32,157	.44	.71
		Verbal ability	39,187	.18	.39	44,478	.39	.64
		Reasoning ability	11,586	.18	.39	4,928	.22	.39

(Continued)

TABLE 2
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
		Memory	7,764	.18	.38	–	–	–
Schmidt, Gast-Rosenberg, & Hunter (1980)	Computer programmer	Programmer Aptitude Test (PAT; number series)	535	.20	.43	–	–	–
		PAT (Figure analogies)	535	.22	.46	–	–	–
		PAT (total score)	1,299	.38	.71	1,635	.64	.91
Schmidt, Hunter, Pearlman, & Shane (1979)	First-line supervisor	General mental ability	5,143	–	.64	–	–	–
Ghiselli (1973)	Managerial	Intellectual abilities	more than 10,000	.27	–	500–999	.30	–
	Clerical	Intellectual abilities	more than 10,000	.28	–	more than 10,000	.47	–
	Sales	Intellectual abilities	5,000–9,999	.19	–	–	–	–
	Protective	Intellectual abilities	1,000–4,999	.22	–	1,000–4,999	.42	–
	Service	Intellectual abilities	1,000–4,999	.27	–	1,000–4,999	.42	–
	Vehicle operators	Intellectual abilities	1,000–4,999	.16	–	1,000–4,999	.18	–
	Trades & crafts	Intellectual abilities	more than 10,000	.25	–	more than 10,000	.41	–
	Industrial	Intellectual abilities	more than 10,000	.20	–	5,000–9,999	.38	–

Note. Several studies also reported information on quantitative-oriented tests (e.g., arithmetic reasoning, quantitative ability), mechanical aptitude (e.g., spatial/mechanical ability, mechanical comprehension), and perceptual/motor ability (e.g., perceptual speed, motor ability). These tests are not reported in this table because the focus is on the verbal oriented measures of cognitive ability or overall measures of cognitive ability. Only meta-analyses of primary data are reported. Studies that reported validities based on meta-analyses conducted by others are not included.

[a]Average (uncorrected) correlation across studies.

[b]Fully corrected correlation across studies. This represents all the corrections made and reported. Although the specific corrections made varies depending on the study, authors typically correct for range restriction and criterion unreliablility.

[c]This result is for performance ratings only. Validity averaged across all reported criteria was .25 ($N = 40{,}230$).

TABLE 3

Summary of Meta-Analyses Investigating the Validity of Personality Tests

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
Hurtz & Donovan (2000)	Averaged across 4 occupational groups	Conscientiousness	7,342	.15	.24[c]	741	.02	.03[c]
		Extraversion	5,809	.06	.09[c]	644	.12	.19[c]
		Agreeableness	5,803	.07	.12[c]	644	.12	.21[c]
		Emotional Stability	5,027	.09	.15[c]	644	.06	.09[c]
		Openness to Experience	4,881	.03	.06[c]	644	.08	.14[c]
Mount, Barrick, & Stewart (1998)	Jobs involving interpersonal interactions	Conscientiousness	1,586	.17	.23	–	–	–
		Extraversion	1,507	.09	.12	–	–	–
		Agreeableness	1,586	.14	.18	–	–	–
		Emotional Stability	1,586	.12	.16	–	–	–
		Openness to Experience	1,507	.11	.14	–	–	–
Vinchur, Schippmann, Switzer, & Roth (1998)	Sales people (performance ratings)	Conscientiousness	2,186	.11	.21	–	–	–
		Extraversion	3,112	.09	.18	–	–	–
		Agreeableness	2,342	.03	.06	–	–	–
		Emotional Stability	3,134	.05	.10	–	–	–
		Openness to Experience	804	.06	.11	–	–	–

(Continued)

TABLE 3
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
		Affiliation	2,389	.06	.12	–	–	–
		Potency	2,907	.15	.28	–	–	–
		Achievement	1,319	.14	.25	–	–	–
		Dependability	1,702	.10	.18	–	–	–
	Sales people (sales criterion)	Conscientiousness	1,774	.17	.31	–	–	–
		Extraversion	2,629	.12	.22	–	–	–
		Agreeableness	918	−.02	−.03	–	–	–
		Emotional Stability	2,157	−.07	−.12	–	–	–
		Openness to Experience	951	.03	.06	–	–	–
Salgado (1997)	Averaged across 5 Occupations in the European community	Conscientiousness	2,241	.10	.26[d]	324	.15	.39[d]
		Extraversion	2,799	.06	.14[d]	383	.01	.03[d]
		Agreeableness	2,574	−.00	−.02[d]	415	.12	.31[d]
		Emotional Stability	2,799	.08	.18[d]	470	.11	.27[d]
		Openness	1,629	.00	.02[d]	477	.11	.26[d]
		Affiliation	279	.08	.15	–	–	–
		Potency	2,278	.15	.26	–	–	–
		Achievement	1,269	.23	.41	–	–	–
		Dependability	359	.10	.18	–	–	–
Mount & Barrick (1995)[e]	All occupations	Conscientiousness	31,275	.18	.31[d]	4,106	.17	.30[d]
		Achievement	16,570	.19	.33	2,731	.20	.35
		Dependability	26,234	.17	.30	3,221	.21	.36

(Continued)

TABLE 3
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
Ones, Viswesvaran, & Schmidt (1993)	All occupations	Integrity tests	68,772	.21	.34			
	Low job complexity	Integrity tests	1,633	.28	.45			
	Medium job complexity	Integrity tests	16,200	.19	.32			
	High job complexity	Integrity Tests	858	.28	.46			
Hough (1992)		Dependability	46,116	.08	–	4,710	.11	–
		Achievement	2,811	.15	–	1,160	.21	–
		Intellectance	11,297	.01	–	8,744	.02	–
		Potency	65,876	.10	–	8,389	.07	–
		Affiliation	3,390	.00	–	–	–	–
		Adjustment	35,148	.09	–	8,685	.12	–
		Agreeableness	22,060	.05	–	988	.08	–
		Rugged individualism	3,007	.08	–	1,614	.03	–
		Locus of control	12,580	.11	–	225	.28	–
		All personality	202,285	.08	–	34,515	.08	–
Barrick & Mount (1991)	Averaged across 5 occupational groups	Conscientiousness	12,893	.13	.23[c]	3,585	.13	.23[c]
		Extraversion	12,396	.06	.10[c]	3,101	.15	.26[c]
		Agreeableness	11,526	.04	.06[c]	3,685	.06	.10[c]
		Emotional Stability	11,635	.04	.07[c]	3,283	.04	.07[c]

(Continued)

TABLE 3
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
		Openness to Experience	9,454	−.02	−.03[c]	2,700	.14	.25[c]
Tett, Jackson, & Rothstein (1991)	Averaged across all occupations	Personality	13,521	.16	.24[c]	–	–	–
		Conscientiousness[f]	450	.12	.18[c]	–	–	–
		Extraversion[f]	2,302	.10	.16[c]	–	–	–
		Agreeableness[f]	280	.22	.33[d]	–	–	–
		Neuroticism[f]	900	−.15	−.22[c]	–	–	–
		Openness[f]	1,304	.18	.27[c]	–	–	–
Hough, Eaton, Dunnette, Kamp, & McCloy (1990)	Averaged across all occupations	Dependability	–	.13	–	–	.11	–
		Achievement	–	–	–	–	.33	–
		Intellectance	–	.01	–	–	.14	–
		Surgency	–	.04	–	–	.08	–
		Affiliation	–	−.01	–	–	–	–
		Adjustment	–	.13	–	–	.16	–
		Agreeableness	–	−.01	–	–	.10	–
		Masculinity	–	–	–	–	.09	–
		Locus of control	–	–	–	–	.29	–
Schmitt, Gooding, Noe, & Kirsch (1984)	Averaged across 6 occupational groups	Personality	4,065	.21[g]	–	–	–	–
Ghiselli (1973)	Managerial	Personality	more than 10,000	.21	–	–	–	–

(Continued)

TABLE 3
(Continued)

Author	Job type(s)	Test/Construct type	Proficiency criteria			Training criteria		
			Total *N*	$\bar{r}$[a]	$\bar{\rho}$[b]	Total *N*	$\bar{r}$	$\bar{\rho}$
	Clerical	Personality	more than 10,000	.24	–	–	–	–
	Sales	Personality	1,000–4,999	.31	–	–	–	–
	Protective	Personality	500–999	.24	–	100–499	−.11	–
	Service	Personality	100–499	.16	–	–	–	–
	Vehicle operators	Personality	100–499	.26	–	–	–	–
	Trades & crafts	Personality	100–499	.29	–	–	–	–
	Industrial	Personality	less than 100	.50	–	–	–	–

Note. Only meta-analyses of primary data are reported. Studies that reported validities based on meta-analyses conducted by others are not included.

[a]Average (uncorrected) correlation across studies.

[b]Corrected correlation across studies. This represents all the corrections made and reported, unless otherwise noted. Although the specific corrections made varies depending on the study, authors typically correct for range restriction and criterion unreliability.

[c]This corrected correlation also includes a correction for predictor unreliability. Corrections for predictor unreliability have traditionally not been conducted because it overstates operational validity (i.e., the fact that tests are used as they are, with a given level of unreliability).

[d]This corrected correlation includes a correction for predictor unreliability and construct invalidity. The correction for construct invalidity is based on the assumption that in the meta-analyses of Tett et al. (1991) and Barrick and Mount (1991) the validity of single scales from the personality inventories were examined, not the Big Five. As such, the scales were only elements or facets of the construct. Thus, if interest is at the theoretical construct level and not in the level of scales, validity must be corrected for imperfect construct assessment. See Salgado (1997), page 33, for a complete explanation. Such a correction was criticized by Hurtz and Donovan (2000) who only included studies that used personality measures that were explicitly designed to measure the Big Five personality dimensions. The lower validity estimates that they observed suggested that the corrections made by Mount and Barrick (1995) and Salgado (1997) for construct invalidity are inappropriate because they inflate the validity estimate.

[e]The specific nature of the corrections made in this study are not well specified.

[f]Only includes "confirmatory" studies (i.e., studies in which the researchers indicated either an explicit or implicit rational for examining specific traits in relation to performance in a particular job). Such a practice may incorrectly estimate the true validity (Ones, Mount, Barrick, & Hunter, 1994).

[g]This result is for performance ratings only. Validity averaged across all reported criteria was .15 ($N = 23{,}413$).

TABLE 4
Summary of Meta-analytic Validity Estimates for Cognitive and Personality Tests

Statistic	Cognitive tests	Personality tests
Predicting proficiency criteria:		
• Median uncorrected validity	.20	.10
• Inter quartile range of uncorrected validities	.16–.25	.05–.17
• Median corrected validity	.40	.18
• Inter quartile range of corrected validities	.28–.52	.09–.26
Predicting training criteria:		
• Median uncorrected validity	.38	.11
• Inter quartile range of uncorrected validities	.30–.42	.08–.15
• Median corrected validity	.62	.23
• Inter quartile range of corrected validities	.54–.67	.10–.30

Note. Based on 13 meta-analyses for cognitive tests and 12 for personality tests (see Tables 2 and 3). Corrections are explained in the notes to those tables.

tests because it will overstate the validity of the test when it is actually used. That is, because a test must be used as it is (i.e., with a given level of reliability), any correction for test unreliability will overestimate the operational validity of the test. The amount of overestimation is difficult to predict because few personality meta-analyses report validities corrected for range restriction and criterion unreliability as well as validities corrected for range restriction, criterion unreliability, and predictor unreliability. Mount, Barrick, and Stewart (1998), however, did report both figures. When validities were only corrected for range restriction and criterion unreliability, the validity was .23. When validities were corrected for range restriction, criterion unreliability, and predictor unreliability, the validity was .26. This suggests that correcting for predictor unreliability had a non trivial effect on validity estimates (i.e., 5.3% of the variance explained vs. 6.8% of the variance explained).

Mount and Barrick (1995) and Salgado (1997) also made corrections for "construct invalidity" (i.e., the fact that many of the personality measures coded as a particular personality dimension are poor measures of the underlying construct). This further overestimates the validity of personality tests as used in practice because tests must be used as they are. Moreover, recent research by Hurtz and Donovan (2000) criticized this correction. They conducted a meta-analysis on studies that used personality tests that were explicitly designed to measure the underlying Big Five personality dimensions. They found that the validity for Conscientiousness was .24, which was smaller than Mount and Barrick (1995) at .31 or Salgado (1997) at .26, suggesting that corrections for construct invalidity

overestimate criterion-related validity. Interestingly, the validity figure of .31 was used by Schmidt and Hunter (1998) in their influential article that examined the incremental validity of a number of constructs beyond that of cognitive ability. This suggests that they overestimated the incremental validity of conscientiousness.

One other comment about these meta-analyses is that cognitive ability measures typically predict training criteria very well. This is especially important as knowledge-based work becomes more common in organizations and the training demands of such jobs increase. In contrast, personality tests are not very effective at predicting training performance.

Furthermore, the meta-analyses on integrity tests are different from most other meta-analyses in a couple of important ways. For one, unlike most other meta-analyses of selection procedures, these meta-analyses on integrity tests relied more than usual on articles or data from companies that produced those tests. And compared to other meta-analyses in other construct domains, that is more of a problem with the integrity test meta-analyses. These tests were originally developed for their commercial potential, not just for research. My concern is not the "file drawer" problem (i.e., studies that are written but never published). I believe that non supportive results were never even documented.

I have a vivid memory of being called by a test publisher over 25 years ago. We had purchased some of their tests, and they were following up to see how our validation study turned out. They clearly stated on the phone that they were only interested in the significant results. I gave them the data on the several significant correlations, but they did not even want to hear about the many non significant correlations. In all fairness, I do not believe they were trying to deceive the public. They were simply trying to gather information that would help them market their tests. They were not thinking about the potential influence on meta-analyses that might be conducted in 20 years.

Another concern is the quality of the source studies. There was great interest in the early 1980s in the integrity test industry to get articles published on integrity tests. This is when many of the tests were originally developed. Their researchers submitted many articles. I was a reviewer during these days and was frequently assigned these articles. Very few of these studies were accepted for publication because they were not very good quality. Gathering all of these low quality unpublished articles and conducting a meta-analysis does not erase their limitations. We have simply summarized a lot of low quality studies. Therefore, the findings of the meta-analyses cannot be believed uncritically. I think they overestimate the criterion-related validity due to methodological weaknesses in the source studies.

Issue 3: What Effect Does Faking Have in Personality Measurement?

John Hollenbeck. One thing that seems clear is that faking does not distort the criterion-related validities of these tests.[1] Although there is empirical support for this, it is almost unnecessary because for faking to really affect the validity, given how low the correlations are to begin with, it would have to cause radical changes in rank orders. I think this is unlikely because for most applicants it is just a constant that shifts the entire score distribution up. There may be some individual differences in things like self-monitoring or social desirability that perturb this general pattern a little, but it would not radically alter rank orders. This is because it would take radical changes in rank orders to change the correlation at a level of magnitude that would allow one to find a statistically significant difference between correlations. Small changes in rank orderings could affect individual decisions depending upon where the cut score was set, but this is a different matter. The standard error around any one estimated criterion score given the correlations one typically sees in the area of selection is very large (see Cohen & Cohen, 1983, pp. 63–63). Therefore, even with a cognitive ability test, there is not a lot of predictive gain for any one specific decision, but rather the utility is derived from making a large number of predictions, all of which are slightly better. When aggregated in this fashion, small gains in predictive accuracy at the individual level translate into non trivial differences at the organizational level. Therefore, the fear of making an error for any one person, whether due to faking or sampling error around one's predicted criterion score based upon a personality or cognitive ability test, will never be eliminated given the current state of this technology, even at its best.

Moreover, self-monitoring is probably a good thing in most social contexts, suggesting that whatever contributes to faking may also contribute to job performance—especially when one employs a supervisory rating as the criterion as is so often the case. Therefore, faking is probably not a

[1]This conclusion highlights a key feature of this article: These comments reflect the different opinions and interpretations of the panelists. Readers will note that Mike Campion suggested (in Table 1) that 8 of 18 articles indicate that response distortion did affect the validity of the measures. In this footnote, John addresses these apparently different conclusion about the effects of faking on criterion-related validity: It is not clear what standards Mike is using as a criterion for the statement "it affects" criterion-related validity. If the criterion is that the correlations are *significantly different from each other*, then I suspect Mike would conclude that the field is 0 for 18. If he uses the "single group validity comparison," where one correlation is statistically different from zero but another one is not, then perhaps you could get 8 differences, but that is a suspect test. With a sample size of 84, a correlation of .30 is significant but .29 is not, but those hardly differ. If the criterion is the correlations are not the same to the second decimal place, then I am sure that given sampling error, all 18 would find differences.

factor that is suppressing the validities. Finally, whatever construct drives faking, it is probably correlated with conscientiousness, emotional stability, and agreeableness, and thus is probably redundant with much of what is driving the personality scores to begin with. Thus, for a whole host of reasons, faking is not a big problem with respect to criterion-related validity of personality tests.

On the other hand, once you grant that it might change rank orders a little, then it does have an effect at the individual level in the sense that those who fake well will tend to be more likely to get selected—especially if the cut off is drawn around the mean where many similar scores pile up closely. Of course, even validity coefficients in the .40s and .50s do not offer much comfort to any single individual in terms of enhancing the probability that an accurate decision will be made in their own individual case, so this is not something that is new or unique to these kinds of measures, but low personality test validity makes it a little more salient.

Neal Schmitt. In the context of the incidence of faking and the utility of corrections for faking, comments about the validity of personality tests are appropriate even if those validities are .30 or .40. That is, in order for corrections to make an impact on criterion-related validity or standardized performance measures, the faking measure that you are using has to be correlated with the outcome, the predictor, or both. And, in most cases, they do not or those correlations are relatively small. So, the faking correction will not have an impact on criterion-related validity and will not have an impact on standardized outcome performance in most instances. That is why the relatively low level of criterion-related validity of personality measures is relevant in the context of a discussion about solutions to the problem of faking. Faking corrections are not going to have much of an impact even when the validities are substantially higher. This is the context under which low validity comes into play. Since this symposium was held, Fred Oswald and I (Schmitt & Oswald, 2006) have demonstrated the points made here in a simulation. The major conclusion of that work is that if we are interested in improving validity alone, then the use of "faking" scales is not going to have more than minimal effects in all situations that represent reasonable estimates of the correlations among the variables involved and the situations in which tests are used.

Issue 4: How Can Personality Tests be Used in a More Effective Way?

John Hollenbeck. My colleagues and I have frequently used personality measures in our research on teams, and we obtain effects for various traits all the time. In fact, across our program of research over the last 6 to 8 years, we have found an effect for each of the constructs specified

by the five-factor model for some aspect of team-related behavior, but these effects are often contingent on other variables. We have been able to link Conscientiousness to team performance, but we found this link is contingent upon how team performance is operationalized (additive vs. conjunctive indices; LePine, Hollenbeck, Ilgen, & Hedlund, 1997). We have been able to link Extraversion to helping behavior, but this link was contingent upon the legitimacy of the help request (Porter et al., 2003). We have been able to link Agreeableness to performance, but this link was contingent upon the nature of the reward structure (cooperative or competitive; Beersma et al., 2003). We have found that Emotional Stability predicts performance in team contexts but that this link is contingent upon the degree of stress created by the situation (Hollenbeck et al., 2002). We have even been able to find relationships between Openness to Experience and performance when team members were using non traditional communication technologies (e-mail and chat) but not when they were interacting face-to-face (Colquitt, Hollenbeck, Ilgen, LePine, & Sheppard, 2002).

In addition to exploring contingencies, we tend to have better success when we are linking some specific trait whose selection was theory driven, and then we link it to some really specific behavior—like "helping" or "information sharing" or "collective deduction" (Ellis et al., 2003). Moreover, this narrower and more selective criterion is then often measured objectively in a tightly controlled context where the task is held constant. We do not rely on different supervisor evaluations of different people doing different tasks because when you do, the problem is not so much on the predictor side but on the criterion side. We also do not rely exclusively or solely on self-reports of behavior that are notoriously inaccurate in team contexts. For example, in one study, we found effects for certain variables when communication was measured objectively but not when it was measured by subjective self-reports. The problem with the self-reported behavior in this case was that the low communication teams did not really know how much communication was actually occurring in high communication teams. Therefore, in terms of making these tests more effective, much of the battle has to be fought on the criterion side of the prediction equation.

Neal Schmitt. I would second that, and if you are going to use personality measures, make sure you know what the outcome is and direct your personality measure development toward that outcome, then you not only are more likely to have validity, but you are also more likely to be able to defend the use of that test if it is challenged.

John Hollenbeck. Another approach that my colleagues and I have explored in terms of selection for our MBA program was using a forced-choice format for Conscientiousness items. In this research, we paired each Conscientiousness item with some other five-factor model item (e.g.,

Emotional Stability or Agreeableness) that demonstrated the exact same endorsement probability when the two items were presented independently. Therefore, each of the two items was equal in its social desirability operationalized in terms of endorsement rates. Thus, for one to endorse a Conscientiousness item, the person would have to not endorse an item that taps Emotional Stability that is equally socially desirable. Similarly, failing to endorse an item that reflects low Conscientious requires the individual to endorse an item that reflects low Agreeableness at the very same level of social desirability. Across the whole test, this essentially forces the individual to establish how central Conscientiousness is to his or her self-concept *relative* to other socially desirable traits. Thus, effective faking in this context would have to be very, very specific, and sophisticated, in that in order to fake Conscientiousness, one would have to be willing to simultaneously fake looking Emotionally Unstable.

The result of this research with respect to Conscientiousness measured in the traditional manner was to lower the mean by roughly one half of a standard deviation and enhance the variance by just under 50%. This is what you might expect would happen if it made it more difficult for people to fake because items that are easy to fake would result in higher means and less variance relative to true scores. However, again, in this study, it did not affect the criterion-related validity very much. That is, when we tested the correlations between Conscientiousness and GPA for Conscientiousness measured in alternative ways, those correlations were not different from each other.

As I noted earlier, this is very unlikely to ever happen because although this intervention did affect the nature of the distribution of scores, it had only minor effects on rank orders within that distribution. As an aside, this study was rejected for publication because as an intervention, the consensus was that this did not really "solve" the faking problem, where "solving the problem" was defined as creating a significantly higher criterion-related validity for the Conscientiousness measure. Perhaps that is too strict a criterion but is a criterion that we have seen reviewers apply.

Issue 5: What Are Some of the Other Concerns About Faking?

Michael Campion. A whole separate issue is whether faking hurts other candidates. The fact that some candidates fake means other candidates are denied jobs. For many of my clients, they are less concerned with the validity costs of faking than the fact that some candidates give extremely desirable answers and are displacing other candidates. For example, in one situation we saw that 5% got perfect scores. It was very unlikely anyone could be this virtuous. The hiring context was very selective, and so this group of likely fakers was displacing an equal number of more

honest candidates. I suppose that it is possible that faking might be job related, but this client felt that it was lying. And for that group that was denied jobs because of the lying of others, it cost them at a personal level. Therefore, I have an ethical problem with using selection procedures that are easily fakable because some candidates are going to be damaged. The fact that we cannot figure out exactly who they are does not change the fact that somebody is being denied employment because somebody else is lying.

Kevin Murphy. I think one of the things that makes this debate difficult is that we often fail to define what we are talking about, so that people are talking about faking and have a whole variety of different constructs and different sorts of behaviors in mind. One of the definitions of faking is "saying what you think you ought to say rather than what you really want to say." We have a word for that—"civilization."

I am more concerned about the failure to fake. Suppose you are in a situation where you know what you are supposed to do (such as giving a socially desirable or socially acceptable answer) and you cannot or will not do it. We should not be wringing our hands about faking. I think we should be thinking very seriously about the people who give responses that are not socially adaptive in a high stakes situation where they know what they are supposed to do. People who do not know when they should give honest answers and when they should fake might lack a skill of adaptation to a social world.

Issue 6: Can We Detect Faking in Personality Tests?

Neal Schmitt. I do not have much confidence in social desirability scales either. From data that I have collected over the years, they do not seem to do what they are supposed to do. If we want to use faking or lying scales at all, what we should do is to construct a set of items that are obviously false. If you are selecting a computer programmer, use an item that says, "I program in Exit." And, if you have four or five items like that and a single respondent answers all of them (or some small number of them) in the affirmative, they are lying. Those persons should be taken out of the applicant pool. You will have something that is pretty decently defendable—both to the rejected person, him or herself, and to the organization for which you are working.

Michael Campion. With regard to faking scales, I want to concur with Neal and add one more point. A recent review I conducted of faking scales led me to believe that many social desirability scales are actually positive attributes, and I would have a hard time explaining to clients or to candidates why endorsing those items is bad. So, I came down exactly

where Neal did on using items that cannot be true. I called them "bogus" statements. They represent one possible approach to detecting faking. The only concern I had was that candidates might endorse bogus statements due to carelessness. As such, you may not be able to differentiate faking from carelessness. When people take many of our surveys and questionnaires, they are not always paying attention as we think they are.

In fact, some personality tests have such scales and refer to them as validity or carelessness scales. They are usually items with known answers or answers that everyone should respond to in the same way (e.g., everyone should endorse statements such as "Computers can make some jobs more efficient" or "I have completed all the items on this test"). When people do not endorse these items in the proper way, you have evidence that they have not responded to the test carefully and the scores may not be valid.

Robert Dipboye. I too am skeptical about the use of social desirability scales to detect faking on personality inventories. Unfortunately, I am just as skeptical about the options. As Neal suggests, items could be included in some types of tests that if endorsed would be obvious lies. But what would be an obviously incorrect item in the context of personality testing where we are asking for self-reports of values, preferences, sentiments, and the like? Take, for example, some of the items I listed previously such as "I read a lot." Perhaps we could ask for the specific number of books they have read, but who is to verify the accuracy of the response? The inclusion of carelessness scales, as mentioned by Mike, is a potentially good approach to weeding out those applicants who are sloppy or random in their responses, but deceiving is not the same as careless responding. The research on lie detection in the interview, or interview types of settings, suggests other possible approaches. For instance, there is evidence that response latency to personality test items is an indicator of honest responding (Holden & Hibbs, 1995). The findings of this research suggest that we use computer administered personality inventories and look for the slow responders in the attempt to identify or correct for faking. Paul Ekman's (2001) research shows that lying is reflected in microexpressions and people can detect lying in an interview if they are trained to recognize the microexpressions that accompany lying. Microexpressions are very subtle changes in facial expression and posture, such as shrugs and smiles, that occur in less than a fraction of a second and leak the emotion behind statements. These are fascinating approaches to lie detection, but neither the use of response latency nor microexpression detection is likely to prove useful in dealing with faking in many selection situations. Questions remain about the relation of time of response on personality inventory items to faking (Holden, Wood, & Tomashewski,

2001) and training in identification of micro expressions is only relevant to one-on-one interviews.

So what's the solution? Again, I have to get back to the point of why should we try to catch people lying? Why not start with the reality that applicants obviously will try to convey a favorable image of themselves in a selection situation. If they do not, then that is probably more of a reason for concern than if they do. Decades of attempts to detect faking on self-report personality inventories have failed to produce a technique of catching the liars or correcting for faking that is practical for purposes of selection. I would suggest that our efforts would be better spent on creating a testing environment in which applicants want to describe themselves as honestly as possible in the interest of determining whether they fit the position. This is asking a lot of applicants who need work and understandably want to present themselves positively. In my opinion, it is still a more practical alternative than all forms of lie detection when selection is based on self-reports of typical behavior.

I should note here that my viewpoint may appear similar in some respects to others who have argued that faking is not a major concern in personality testing. Some have argued that faking is of no concern because there is little convincing evidence that it lowers criterion-related validity. But these arguments are usually based on the premise that personality inventories do a good job of predicting job performance. In response to this position I would say that validities are already so low that there may be little room to go lower. And although faking may not account for low validities, there is evidence that they can distort the rank order of applicants in the decision process (Rosse et al., 1998).

Others have argued that faking is of no concern because personality is all about conveying impressions and establishing reputation. From this perspective, ambiguity in personality assessment could be a virtue, not a problem (Johnson, 2004). Thus, according to Hogan (1991, p. 902), responses to personality inventory items are "automatic and often non-conscious efforts on the part of test-takers to negotiate an identity with an anonymous interviewer (the test author)." Consequently, it is insignificant whether a person who agrees with an item such as "I read a lot" in fact reads a lot. What matters is that a person who endorses this item is showing an automatic, unconscious tendency to the endorsement of open and intellectual behavior characteristic of this personality trait. The problem I have with this viewpoint is that it seems to fail in distinguishing the selection situation from everyday interactions where we negotiate identities. Negotiation implies a give-and-take that is missing when an applicant responds to a personality inventory in a selection context. Also, automatic, unthinking responding to items seems less likely in my opinion than hypervigilance and acute self-consciousness. In short, I fail to see the taking

of a personality inventory in selection as a normal extension of everyday life but rather as a bizarre circumstance that resembles in few ways the real world. In responding to items applicants may lie about themselves but many others respond as best they can in an attempt to cope with what is essentially an unrealistic self-report task.

Issue 7: What Recommendations Would You Give About the Use of Personality Tests in Selection Contexts?

Neal Schmitt. I would make three specific recommendations. First, avoid published personality measures in almost all instances. The only time I would use them is if they are directly linked in a face-valid way to some outcome. Second, I would construct my own measures that are linked directly to job tasks in a face-valid or relevant fashion. Now, you might say that is when you are actually going to get faking. Fine. If you think that is going to be the case, I would include a carelessness or faking scale that detects individuals who are careless or those who are obviously lying or faking. I mentioned as an example earlier, "programming in Exit." An item we sometimes use to catch obviously careless respondents is "tomato plants are always larger than trees." If people answer those kinds of items frequently, then you know something is wrong. The end result is that you are most likely to have a valid test, and you will also be able to explain why you ask certain questions, and you are less likely to have to answer embarrassing questions about why in the world you are asking about people's sexual behavior when they are applying for a job. You could also explain why the liars are being removed on the basis of their responses to these obviously idiotic questions, and somebody else will accept it. Third, I also agree with the notion that we often spend too much time trying to fool our respondents in most of this research. One of the recommendations I give to my students is that if you have a personality measure or a multi dimensional construct of any kind, label the items measuring various constructs, and put all the items measuring a single construct in the same place in your instrument so respondents will know what you are asking. Too often, we try to get individuals to contribute to our research efforts, and then we spend all the time they are responding to our questions trying to fool them about what we want to know. Ask them directly and several times, then we will be more likely to get quality answers.

Michael Campion. I fully agree with Neal's point about developing personality tests that are clearly job related and that avoid questions that are ambiguous (at best) or embarrassing (at worst).

I would like to second Neal's point about labeling our constructs and add a reference. There are a couple of articles by Chet Schreisheim (e.g., Schreisheim, Kopelman, & Solomon, 1989; Schreisheim, Solomon, & Kopelman, 1989) in which he found that grouping and labeling items on the same construct helped the factor structure of the instrument. I believe this really helps. Whatever you want to measure, tell respondents what it is and put all the questions together. They have a great deal of difficulty with what must appear to be subtle differences between our esoteric constructs. We should make it easier for them, or we will get meaningless data. This is especially true in research contexts. In selection contexts, I am not so sure. Telling people that you are trying to measure their Conscientiousness may lead to more faking. So I would not label the scales so clearly in a selection context.

Research Issues

Issue 8: Should Personality Research Focus on an Expanded Criterion Domain?

Kevin Murphy. If you break down performance measures into contextual versus task-related, you find higher levels of validity for personality inventories as predictors of contextual performance. It is not markedly higher; it is not as high as it should be.

Regardless of the criterion we are trying to predict, I think taking a self-report at face value is bad practice. We usually understand that we ought to adjust what we hear in other social settings and take the context into account. Then we get to personality tests and we act like it is a problem if the context influences the answers. So, I think it goes back to the weakness of self-report as a one and only method. I think that if we are going to do better here, you need to consider multiple methods. That being said, I do think you will find higher validities for some facets of performance than for others.

Michael Campion. I have tried to distinguish "team performance" from "task performance" in many research contexts. This is close to the contextual versus task performance distinction. I have been very frustrated by my inability to distinguish between these different aspects of job performance. I also have seen this problem in the research conducted by others. In our profession, we are very poor at teasing apart different job performance constructs, especially with the same measurement method. Therefore, I agree with Kevin's point about the importance of focusing on aspects of the criterion domain that may be more logically related to

personality, but I am concerned about our ability to distinguish it from other aspects of job performance.

Robert Dipboye. Expanding the criterion domain is a good idea. Limitations in the criterion domains used in validation are a major problem not only in the validation of personality tests but also the interview and other non cognitive measures. But I do not think that this is going to be the answer because you are still going to be confronted with the fact that the measures of the predictor are crude. Assuming we were able to identify and measure all aspects of the criteria important to a job, we still would lack the personality measures that would allow valid prediction of these criteria.

Issue 9: What Kind of Future Research Should Be Done?

John Hollenbeck. I would actually like to see a study that focuses a little more closely on the low scorers in this context, especially because most of the traits that might be employed in selection contexts have some degree of social desirability associated with them. That is, it is probably better to be Conscientious, Emotionally Stable, Agreeable, and so on. Thus, although a high score might be ambiguous (is the person high or just faking?) a low score can only have one interpretation—the person is so low on the trait that he or she cannot even fake it. Now we all know if we were to take a correlation between any trait or any characteristic and cut it in half, we could use the statistical formulas for saying, "well, how much did this shrink?" I would like to see if a lot of the predictive validity that you see for personality measures really occurs at the low level.

I had a student one time that came to my office because he did not do very well on a test in my class and he said, "I can't understand it because I studied during the whole basketball game—even halftime. When the teams took a break, not me, I kept working." This was a student that quite honestly could not even fake Conscientiousness. Likewise, to the extent that anyone taking an overt honesty test that basically asks you the question, "on the average, how much have you stolen from previous employers?" answers "$212.00" then I think that is diagnostic. So, I would really like to see studies of people that score at the low end of some of these scales.

In addition, I would like to see a more thorough breakdown of the jobs that are being studied in this context. A lot of jobs require faking as part of effective job performance. If you work at Disney, you are supposed to be having a good day every single day. Again, if you cannot fake a personality item, how are you going to work at Disney everyday, smiling at these young children stomping on your feet? I think we need an analysis of this behavior.

Finally, I think conditional reasoning might be another approach that is really promising in this regard. These tests are less direct than traditional measures and focus on how people solve problems that appear, on the surface at least, to be standard inductive reasoning items. In reality, however, the items tap implicit biases (e.g., rationalizing violent behavior) that are more attractive to a responder if they are high on some focal trait, such as aggressiveness, as opposed to low on this trait (James et al., 2005). These measures have been found to be psychometrically sound in terms of internal consistency and test–retest estimates of reliability, and their predictive validity against behavioral criteria, uncorrected, is over .40 (James, McIntyre, Glisson, Bowler, and Mitchell, 2004). As others have suggested here, we need to divorce ourselves a little bit from straight self-reports of how people think they behave and instead focus more on how they actually behave regardless of how they perceive it. Conditional reasoning tests, which focus on how people make decisions and judgments, are an interesting step in that direction.

Neal Schmitt. Another issue that is sometimes raised is that we may not change criterion-related validity much when we correct for faking, but I am not sure that is the right thing to be looking at anyway if we are interested in organizational outcomes. We should be looking at the change in standardized outcome performance. There is only one unpublished paper by Zickar, Rosse, Levin, and Hulin (1996) in which researchers have proceeded in that fashion. If we had a good faking index, we would examine the scores of individuals who are at the top of our distribution—the ones that would normally be selected if we were going in top-down fashion. We would then identify the top-scoring persons that have high faking scores, remove them from consideration, and replace them with other individuals. So, what we need to do is to take a look at the average standardized performance of the individuals who are on the original list and compare it with the standardized performance of those who are going to be on the replacement list. Nobody has ever done that, with the exception of Zickar et al. Given the validity of personality, the correlation of faking measures with criterion and predictor, and the scenario described above, I cannot believe that the standardized criterion performance will be substantially affected one way or the other if "fakers" were removed. These points have been confirmed and demonstrated in the paper mentioned earlier (Schmitt & Oswald, 2006).

Another point someone mentioned is that for faking corrections to make any difference, there have to be individual differences in faking. If everybody fakes the same amount, we just move everybody up so there must be individual differences in faking if corrections are to have any effect. There is almost no literature on this. McFarland and Ryan's (2000)

paper suggested that individuals who were low in Conscientiousness, low in integrity, high in Neuroticism, and bright, do fake more, or at least these are the individual difference measures that are correlated with faking measures. But, all of those correlations were fairly small and are unlikely to have a huge effect on the bottom line.

Kevin Murphy. I have more faith in the constructs than in the measures. And I think that the problem is that we have been relying on self-report measures for the last 100 years or so. We should look at other ways of assessing personality. If you want to know about someone's personality, just ask his or her coworkers. There are a variety of ways of finding out about people's stable patterns of behavior. The validity results for reports from other people, at least in my opinion, have been much more encouraging for the most part. I think that self-report is the problem, not the personalities. The idea that personalities are not related to behaviors in the workplace is almost by definition wrong. So, I think that we have a method of measurement that is, in my view, not salvageable, and therefore, dealing with the faking issue is low on my totem pole.

Michael Campion. I would resonate with the recommendation that is emerging here—we are really interested in studying these personality constructs, but we are very concerned with the self-report method of measurement due to all of its limitations. Whether it is criterion performance measurement or the measurement of personality, I think one theme I hear is let's think about different ways of measuring these constructs. Let's not abandon the construct, but instead abandon self-report measurement and think about new and innovative ways of measuring the constructs. For a good recent review of the evidence of how poor we are at self-assessment across a range of domains, see Dunning, Heath, and Suls (2004).

Robert Dipboye. Rather than abandoning self-report perhaps we can improve them. Here, I might disagree to some extent with Kevin by suggesting that one avenue for research is to explore how to use self-report in a manner that solicits more accurate information. The typical administration of a personality inventory in a selection situation makes a person feel as though he or she is in an interrogation. It is a one-way communication in which the individual knows little about what is happening or why. The person is kept uninformed about the objectives of the test and why he or she needs to answer the questions that are asked. What if we were more open about the nature of the personality inventory and allowed people to participate to some degree in the testing process? It may seem radical to disclose to respondents what is being measured in a personality inventory but some research suggests that such disclosure may yield better validity. For instance, there was an article in *Applied Psychology: International*

Review (Kolk, Born & van der Flier, 2003) that examined the impact of making assessment center dimensions available to applicants versus not making them available. They found some improvement in construct validity when the applicants knew the dimensions on which they were being tested.

Another strategy is to allow people to elaborate on their responses to personality items. Neal Schmitt and Charles Kunce (2002) found in a study published in *Personnel Psychology* that having individuals elaborate on their responses to biodata items was associated with less faking. This is just one example of creative ways of administering personality inventories to improve the quality of responses. Whether these strategies are practical in mass testing remains to be seen, but we ought to explore alternatives to the interrogation that constitutes the typical personality testing situation. Chris Argyris in his 1960s *Psychological Bulletin* article on the unintended consequences of rigorous research warned that when we engage participants in the one-way authoritarian relationship characterizing many social psychology experiments, experimenters should not expect open and frank responses from participants (Argyris, 1968). Likewise, we should not expect job applicants to be open and frank in their responses to personality inventory items when the focus of measurement becomes ferreting out the liars and fakers.

Conclusion

Looking across the topics covered and the opinions expressed, the following conclusions are drawn based on areas of agreement among the panelists. Differences among the authors are so noted.

(a) Faking on self-report personality tests should be expected, and it probably cannot be avoided, although there is some disagreement among the authors on the extent to which faking is problematic.
(b) Faking or the ability to fake may not always be bad. In fact, it may be job-related or at least socially adaptive in some situations.
(c) Corrections for faking do not appear to improve validity. However, the use of bogus items may be a potentially useful way of identifying fakers.
(d) We must not forget that personality tests have very low validity for predicting overall job performance. Some of the highest reported validities in the literature are potentially inflated due to extensive corrections or methodological weaknesses.
(e) Due to the low validity and content of some items, many published self-report personality tests should probably not be used for personnel selection. Some are better than others, of course, and when those

better personality tests are combined with cognitive ability tests, in many cases validity is likely to be greater than when either is used separately.

(f) If personality tests are used, customized personality measures that are clearly job-related in face valid ways might be more easily explained to both candidates and organizations.

(g) Future research might focus on areas of the criterion domain that are likely to be more predictable by personality measures.

(h) Personality constructs certainly have value in understanding work behavior, but future research should focus on finding alternatives to self-report personality measures. There is some disagreement among the authors in terms of the future potential of the alternative approaches to personality assessment currently being pursued.

REFERENCES

References followed by numbers in parentheses are cited by number in Table 1.

Abbott RD. (1975). Improving the validity of affective self-report measures through constructing personality scales unconfounded with social desirability: A study of the Personality Research Form. *Educational and Psychological Measurement, 35*, 371–377. (70)

Abbott RD, Harris L. (1973). Social desirability and psychometric characteristics of the Personal Orientation Inventory. *Educational and Psychological Measurement, 33*, 427–432. (80)

Alliger GM, Dwight SA. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59–72. (4)

Argyris C. (1968). Rigorous research designs and unintended consequences. *Psychological Bulletin, 70*, 185–197.

Austin JS. (1992). The detection of fake good and fake bad on the MMPI-2. *Educational and Psychological Measurement, 52*, 669–674. (29)

Baer RA, Wetter MW, Berry DT. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12*, 509–525. (30)

Bagby RM, Gillis JR, Toner BB, Goldberg J. (1991). Detecting fake-good and fake-bad reasoning on the Millon Clinical Multiaxial Inventory-II. *Psychological Assessment, 3*, 496–498. (34)

Barrett GV, Polomsky MD, McDaniel MA. (1999). Selection tests for firefighters: A comprehensive review and meta-analysis. *Journal of Business and Psychology, 113*, 507–513.

Barrick MR, Mount MK. (1991). The Big Five personality dimensions and job performance: A meta-analysis. PERSONNEL PSYCHOLOGY, *44*, 1–26.

Barrick MR, Mount MK. (1996). Effects of imppression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261–272.

Barrick MR, Mount MK. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261–272. (13)

Barrick MR, Ryan AM. (2003). *Personality and work: Reconsidering the role of personality in organizations*. San Francisco: Jossey-Bass.

Bass BM. (1957). Faking by sales applicants of a forced choice personality inventory. *Journal of Applied Psychology, 41*, 403–404. (109)

Beersma B, Hollenbeck JR, Humphrey SE, Moon H, Conlon DE, Ilgen DR. (2003). Cooperation, competition, and team performance: Towards a contingency approach. *Academy of Management Journal, 46*, 572–590.

Borkenau P, Amelang M. (1985). The control of social desirability in personality inventories: A study using the principle-factor deletion technique. *Journal of Research in Personality, 19*, 44–53. (54)

Borofsky GL. (1992). Assessing the likelihood of reliable workplace behavior: Further contributions to the validation of the Employee Reliability Inventory. *Psychological Reports, 70*, 563–592. (31)

Braun JR. (1962). Effects of a top management faking set on the Gordon Personal Inventory. *Psychological Reports, 10*, 611–614. (100)

Braun JR. (1963). Effects of positive and negative faking sets on the survey of interpersonal values. *Psychological Reports, 13*, 171–173. (97)

Braun JR, Farrell RM. (1974). Re-examination of the fakability of the Gordon Personal Inventory and Profile: A reply to Schwab. *Psychological Reports, 34*, 247–250. (77)

Braun JR, Gomez BJ. (1966). Effects of faking instructions on the Eysenck Personality Inventory. *Psychological Reports, 19*, 388–390. (94)

Braun JR, LaFaro D. (1968). Fakability of the Sixteen Personality Factor Questionnaire, Form C. *Journal of Psychology, 68*, 3–7. (92)

Braun JR, Smith M. (1973). Fakability of the self-perception inventory: Further investigation. *Psychological Reports, 32*, 586. (81)

Braun JR, Tinley JJ. (1972). Fakability of the Edwards Personality Inventory booklets, IA, II, and III. *Journal of Clinical Psychology, 28*, 375–377. (84)

Christiansen ND, Goffin RD, Johnston NG, Rothstein MG. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. PERSONNEL PSYCHOLOGY, *47*, 847–860. (20)

Cohen J, Cohen P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen J, Lefkowitz J. (1974). Development of a biographical inventory blank to predict faking on personality tests. *Journal of Applied Psychology, 59*, 404–405. (78)

Colquitt JA, Hollenbeck JR, Ilgen DR, LePine JA, Sheppard L. (2002). Computer-assisted communication and team decision-making accuracy: The moderating effect of openness to experience. *Journal of Applied Psychology, 87*, 402–410.

Comrey AL, Backer TE. (1975). Detection of faking on the Comrey personality scales. *Multivariate Behavioral Research, 10*, 311–319. (71)

Costello RM, Schneider SL, Schoenfeldt LS. (1993). Applicants' fraud in law enforcement. *Psychological Reports, 73*, 179–183. (23)

Cunningham MR, Wong DT, Barbee AP. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies of the Reid report. *Journal of Applied Psychology, 79*, 643–658. (21)

Dicken C. (1963). Good impression, social desirability, and acquiescence a suppressor variables. *Educational and Psychological Measurement, 23*, 699–720. (98)

Dicken CF. (1960). Simulated patterns on the California Psychological Inventory. *Journal of Counseling Psychology, 7*, 24–31. (105)

Doll RE. (1971). Item susceptibility to attempted faking as related to item characteristic and adopted fake set. *Journal of Psychology, 77*, 9–16. (86)

Dunnett S, Koun S, Barber PJ. (1981). Social desirability in the Eysenck Personality Inventory. *British Journal of Psychology, 72*, 19–26. (59)

Dunnette MD, McCartney J, Carlson HC, Kirchner WK. (1962). A study of faking behavior on a forced-choice self-description checklist. PERSONNEL PSYCHOLOGY, *15*, 13–24. (101)

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

Edwards AL. (1965). Measurement of individual differences in ratings of social desirability and in the tendency to give socially desirable responses. *Journal of Experimental Research in Personality, 1*, 91–98. (95)

Edwards AL. (1990). Construct validity and social desirability. *American Psychologist, 45*, 287–289. (38)

Edwards AL, Edwards LK. (1992). Social desirability and Wiggins's MMPI content scales. *Journal of Personality and Social Psychology, 62*, 147–153. (32)

Ekman P. (2001). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: W. W. Norton & Company.

Ellingson JE, Sackett PR, Hough LM. (1999). Social desirability correction in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166. (7)

Ellingson JE, Smith DB, Sackett PR. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*, 122–133. (2)

Elliot AGP. (1981). Some implications of lie scale scores in real-life selection. *Journal of Occupational Psychology, 54*, 9–16. (60)

Ellis A, Hollenbeck JR, Ilgen DR, Porter CO, West B, Moon H. (2003). Team learning: Collectively connecting the dots. *Journal of Applied Psychology, 88*, 821–835.

Esposito JL, Agard E, Rosnow RL. (1984). Can confidentiality of data pay off? *Personality and Individual Differences, 5*, 477–480. (55)

French EG. (1958). A note on the Edwards Personal Preference Schedule for use with basic airmen. *Educational and Psychological Measurement, 18*, 109–115. (107)

Furnham A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385–400. (50)

Furnham A, Craig S. (1987). Fakeability and correlates of the Perception and Preference Inventory. *Personality and Individual Differences, 8*, 459–470. (47)

Gayton WE, Ozmon KL, Wilson WT. (1973). Effects of faking instructions on the Psychological Screening Inventory. *Psychological Reports, 32*, 516–518. (82)

Ghiselli EE. (1973). The validity of aptitude tests in personnel selection. PERSONNEL PSYCHOLOGY, *26*, 461–477.

Gillis JR, Rogers R, Dickens SE. (1990). The detection of faking bad response styles on the MMPI. *Canadian Journal of Behavioural Science, 22*, 408–416. (39)

Goldsmith RE, Matherly TA. (1986). The Kirton Adaptation Innovation Inventory, faking, and social desirability: A replication and extension. *Psychological Reports, 58*, 269–270. (51)

Gordon ME, Gross RH. (1978). A critique of methods for operationalizing the concept of fakeability. *Educational and Psychological Measurement, 38*, 771–782. (67)

Graham JR, Watts D, Timbrook RE. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment, 57*, 264–277. (35)

Guion RM. (1965). *Personnel testing*. New York: McGraw Hill.

Guion RM, Cranny CJ. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 67*, 239–244.

Guion RM, Gottier RF. (1965). Validity of personality measures in personnel selection. PERSONNEL PSYCHOLOGY, *18*, 135–164.

Heron A. (1956). The effects of real-life motivation on questionnaire response. *Journal of Applied Psychlogy, 40*, 65–68. (110)

Hinrichsen JJ, Gryll SL, Bradley LA, Katahn M. (1975). Effects of impression management efforts on FIRO-B profiles. *Journal of Consulting and Clinical Psychology, 43*, 269. (72)

Hirsh HR, Northrop LC, Schmidt FL. (1986). Validity generalization results for law enforcement occupations. PERSONNEL PSYCHOLOGY, *39*, 399–420.

Hogan R. (1991). Personality and personality measurement. In Dunnette MD (Ed.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 873–919). Palo Alto, CA: Consulting Psychologists Press.

Holden RR, Fekken GC. (1989). Three common social desirability scales: Friends, acquaintances, or strangers. *Journal of Research in Personality, 23*, 180–191. (44)

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

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

Hollenbeck JR, Moon H, Ellis A, West B, Ilgen DR, Sheppard L et al. (2002). Structural contingency theory and individual differences: Examination of external and internal person-team fit. *Journal of Applied Psychology, 87*, 599–606.

Hough LM. (1992). The "Big Five" personality variables—Construct confusion: Description versus prediction. *Human Performance, 5*, 139–155.

Hough LM. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209–244. (9)

Hough LM, Eaton NK, Dunnette MD, Kamp JD, McCloy RA. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581–595. (40)

Hough LM, Furnham A. (2003). Use of personality variables in work settings. In Borman WC, Ilgen DR, Klimoski RJ (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 131–169). Hoboken, NY: Wiley.

Hsu LM, Santelli J, Hsu JR. (1989). Faking detection validity and incremental validity of the response latencies to MMPI subtle and obvious items. *Journal of Personality Assessment, 53*, 278–295. (45)

Hunter JE. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340–362.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hurtz GM, Donovan JJ. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879.

Ironson GH, Davis GA. (1979). Faking high or low creativity scores on the Adjective Check List. *Journal of Creative Behavior, 13*, 139–145. (64)

Jackson DN, Messick S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*, 243–252. (108)

Jackson DN, Wroblewski VR, Ashton MC. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388. (5)

James LR, McIntyre MD, Glisson CA, Bowler JL, Mitchell TR. (2004). The conditional reasoning measurement system for aggression: An overview. *Human Performance, 17*, 271–295.

James LR, McIntyre MD, Glisson CA, Green PD, Patton TW, LeBreton JM et al. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods, 8*, 69–99.

Jeske JO, Whitten MR. (1975). Motivational distortion of the sixteen personality factor questionnaire by persons in job applicants' roles. *Psychological Reports, 37*, 379–382. (73)

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

Kanfer R, Ackerman PL, Murtha T, Goff M. (1995). Personality and intelligence in industrial and organizational psychology. In Saklofske DH, Zeidner M (Eds.), *International handbook of personality and intelligence* (pp. 577–602). New York: Plenum.

Kirchner WK. (1962). "Real-life" faking on the Edwards Personal Preference Schedule by sales applicants. *Journal of Applied Psychology, 46*, 128–130. (102)

Kolk NJ, Born MPh, Van Der Flier H. (2003). The transparent centre: The effects of revealing dimensions to candidates. *Applied Psychology: An International Review, 52*, 648–669.

Kroger RO, Turnbull W. (1975). Invalidity of validity scales: The case of the MMPI. *Journal of Consulting and Clinical Psychology, 43*, 48–55. (74)

Kroger RO, Wood LA. (1993). Reification, "faking," and the Big Five. *American Psychologist, 48*, 1297–1298. (24)

Krug SE. (1978). Further evidence on 16PF distortion scales. *Journal of Personality Assessment, 42*, 513–518. (68)

Lanyon RI. (1993). Development of scales to assess specific deception strategies on the Psychological Screening Inventory. *Psychological Assessment, 5*, 324–329. (25)

Lautenschlager GJ. (1986). Within-subject measures for the assessment of individual differences in faking. *Educational and Psychological Measurement, 46*, 309–316. (52)

Leary MR, Kowalski RM. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*, 34–47. (41)

LePine JA, Hollenbeck JR, Ilgen DR, Hedlund J. (1997). Effects of individual differences on the performance of hierarchical decision making teams: Much more than *g*. *Journal of Applied Psychology, 82*, 803–811.

Levine EL, Spector PE, Menon S, Narayanan L, Cannon-Bowers JA. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance, 9*, 1–22.

Lilienfeld SO, Andrews BP, Stone-Romero EF, Stone D. (1994). The relations between a self-report honesty test and personality measures in prison and college samples. *Journal of Research in Personality, 28*, 154–169. (22)

Mahar D, Colognon J, Duck J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. *Personality and Individual Differences, 18*, 605–609. (16)

Marlowe D, Crowne DP. (1961). Social desirability and response to perceived situational demands. *Journal of Consulting Psychology, 25*, 109–115. (104)

Match J, Wiggins N. (1974). Individual viewpoints of social desirability related to faking good and desirability estimation. *Educational and Psychological Measurement, 34*, 591–606. (79)

McClelland JN, Rhodes F. (1969). Prediction of job success for hospital aides and orderlies from MMPI scores and personal history data. *Journal of Applied Psychology, 53*, 49–54. (90)

McCrae RR, Costa PT Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting & Clinical Psychology, 51*, 882–888. (58)

McFarland LA, Ryan AM. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821. (6)

Meehl PE, Hathaway SR. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30*, 525–564. (112)

Messick S. (1960). Dimensions of social desirability. *Journal of Consulting Psychology, 24*, 279–287. (106)

Michaelis W, Eysenck HJ. (1971). The determination of personality inventory patterns and intercorrelations by changes in real-life motivation. *Journal of Genetic Psychology, 118*, 223–234. (87)

Morgeson FP, Campion MA, Dipboye RL, Hollenbeck JR, Murphy K, Ryan AM et al. (2004, April). *Won't get fooled again? Editors discuss faking in personality testing*. Panel discussion conducted at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Mount MK, Barrick MR. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In Rowland KM, Ferris GR (Eds.), *Research in personnel and human resources management* (Vol. 13, pp. 153–200). Greenwich, CT: JAI Press. (17)

Mount MK, Barrick MR, Stewart GL. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145–165.

Nicholson RA, Hogan R. (1990). The construct validity of social desirability. *American Psychologist, 45*, 290–292. (42)

Norman WT. (1963). Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology, 47*, 225–241. (99)

Ones DS, Mount MK, Barrick MR, Hunter JE. (1994). Personality and job performance: A critique of the Tett, Jackson, and Rothstein (1991) meta-analysis. PERSONNEL PSYCHOLOGY, *47*,147–156.

Ones DS, Viswesvaran C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269. (10)

Ones DS, Viswesvaran C, Reiss AD. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679. (14)

Ones DS, Viswesvaran C, Schmidt FL. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology, 78*, 679–703. (26)

Orpen C. (1971). The fakability of the Edwards Personal Preference Schedule in personnel selection. PERSONNEL PSYCHOLOGY, *24*, 1–4. (88)

Palsane MN, Lodhi PH. (1979). Eysenck Personality Inventory scales and social desirability: A correlational and factorial study. *Psychologia, 22*, 236–240. (65)

Paulhus DL. (1984). Two-component models of socially desirable responding. *Journal of Personality & Social Psychology, 46*, 598–609. (56)

Paulhus DL, Bruce MN, Trapnell PD. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin, 21*, 100–108. (18)

Paulhus DL, Reid DB. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60*, 307–317. (36)

Pearlman K, Schmidt FL, Hunter JE. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Peterson GW, Clark DA, Bennett B. (1989). The utility of MMPI subtle, obvious, scales for detecting fake good and fake bad response sets. *Journal of Clinical Psychology, 45*, 575–583. (46)

Porter CO, Hollenbeck JR, Ilgen DR, Ellis AP, West BJ, Moon H. (2003). Backing up behaviors in teams: The role of personality and legitimacy of need. *Journal of Applied Psychology, 88*, 391–403.

Pumroy DK. (1962). Relationship between the social desirability scale and the California Psychological Inventory. *Psychological Reports, 10*, 795–796. (103)

Rahim MA. (1984). The Social Desirability Response Set and the Eysenck Personality Inventory. *Journal of Psychology, 116*, 149–153. (57)

Ramanaiah NV, Martin HJ. (1980). On the two-dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Personality Assessment, 44*, 507–514. (61)

Ramanaiah NV, Schill T, Leung LS. (1977). A test of the hypothesis about the two-dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Research in Personality, 11*, 251–259. (69)

Robbinette RL. (1991). The relationship between the Marlowe-Crowne Form C and the validity scales of the MMPI. *Journal of Clinical Psychology, 47*, 396–399. (37)

Rosse JG, Stecher MD, Miller JL, Levin RA. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644. (11)

Ruch FL, Ruch WW. (1967). The *K* factor as a (validity) suppressor variable in predicting success in selling. *Journal of Applied Psychology, 51*, 201–204. (93)

Ryan A, Sackett P. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1*, 248–256. (48)

Salgado JF. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*, 30–43.

Schlenker BR, Weigold MF. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology, 43*, 133–168. (33)

Schmidt FL, Gast-Rosenberg I, Hunter JE. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65*, 643–661.

Schmidt FL, Hunter JE. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt FL, Hunter JE, Caplan JR. (1981). Validity generalization results for job groups in the petroleum industry. *Journal of Applied Psychology, 66*, 261–273.

Schmidt FL, Hunter JE, Pearlman K, Shane GS. (1979). Further tests of the Schmidt-Hunter Bayesian Validity Generalization Procedure. PERSONNEL PSYCHOLOGY, *32*, 257–281.

Schmit MJ, Ryan AM. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966–974. (27)

Schmit MJ, Ryan AM, Stierwalt SL, Powell AB. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*, 607–620. (19)

Schmitt N, Cortina JM, Ingerick MJ, Wiechmann D. (2003). Personnel selection and employee performance. In Borman WC, Ilgen DR, Klimoski RJ (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 77–105). Hoboken, NY: Wiley.

Schmitt N, Gooding R, Noe R, Kirsch M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. PERSONNEL PSYCHOLOGY, *37*, 407–422.

Schmitt N, Kunce C. (2002). The effects of required elaboration of answers to biodata questions. PERSONNEL PSYCHOLOGY, *55*, 569–587.

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

Schreisheim CA, Kopelman RE, Solomon E. (1989). The effects of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational and Psychological Measurement, 49*, 487–508.

Schreisheim CA, Solomon E, Kopelman RE. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement, 13*, 19–32.

Schwab DP. (1971). Issues in response distortion studies of personality inventories: A critique and replicated study. PERSONNEL PSYCHOLOGY, *24*, 637–647. (89)

Schwab DP, Packard GL. (1973). Response distortion on the Gordon Personal Inventory and the Gordon Personal Profile in a selection context: Some implications for predicting employee tenure. *Journal of Applied Psychology, 58*, 372–374. (83)

Smith DB, Ellingson JE. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*, 211–219. (1)

Stark S, Chernyshenko OS, Chan K, Lee WC, Drasgow WC. (2001). Effects of the testing situation of item responding: cause for concern. *Journal of Applied Psychology, 86*, 943–953. (3)

Stricker LJ. (1969). "Test-wiseness" on personality scales [Monograph]. *Journal of Applied Psychology, 53*, 1–18. (91)

Tett RP, Jackson DN, Rothstein M. (1991). Personality measures as predictors of job performance: A meta-analytic review. PERSONNEL PSYCHOLOGY, *44*, 703–742.

Thornton GC III, Gierasch PF III. (1980). Fakability of an empirically derived selection instrument. *Journal of Personality Assessment, 44*, 48–51. (62)

Timbrook RE, Graham JR, Keiller SW, Watts D. (1993). Comparison of the Wiener-Harmon subtle-obvious scales and the standard validity scales in detecting valid and invalid MMPI-2 profiles. *Psychological Assessment, 5*, 53–61. (28)

Topping GD, O'Gorman JG. (1997). Effects of faking set on validity of the NEO-DDI. *Personality and Individual Differences, 23*, 117–124. (12)

Tunnell G. (1980). Intraindividual consistency in personality assessment: The effect of self-monitoring. *Journal of Personality, 48*, 220–232. (63)

Velicer WF, Weiner BJ. (1975). Effects of sophistication and faking sets on the Eysenck Personality Inventory. *Psychological Reports, 37*, 71–73. (75)

Vinchur AJ, Schippmann JS, Switzer FS, Roth PL. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83*, 586–597.

Viswesvaran C, Ones DS. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational & Psychological Measurement, 59*, 197–210. (8)

Walsh JA. (1990). Comment of social desirability. *American Psychologist, 45*, 289–290. (43)

Waters LK. (1965). A note on the "fakability" of forced-choice scales. PERSONNEL PSYCHOLOGY, *18*, 187–191. (96)

Weinstein D. (1972). Social desirability as related to the success-failure inventory. *Psychological Reports, 31*, 674. (85)

Wiener DN. (1948). Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology, 12*, 164–170. (111)

Winder P, O'Dell JW, Karson S. (1975). New motivational distortion scales for the 16PF. *Journal of Personality Assessment, 39*, 532–537. (76)

Worthington DL, Schlottman RS. (1986). The predictive validity of subtle and obvious empirically derived psychological test items under faking condition. *Journal of Personality Assessment, 50*, 171–181. (53)

Zalinsky JS, Abrahams NM. (1979). The effects of item context in faking personnel selection inventories. PERSONNEL PSYCHOLOGY, *32*, 161–166. (66)

Zerbe WJ, Paulhus DL. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12*, 250–264. (49)

Zickar MJ, Drasgow F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71–87. (15)

Zickar MJ, Rosse JG, Levin RA, Hulin CL. (1996, April). *Modeling the effects of faking on personality tests*. Paper presented at the 11th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.