PERSONNEL PSYCHOLOGY
2007, 60, 1029–1049

ARE WE GETTING FOOLED AGAIN? COMING TO TERMS WITH LIMITATIONS IN THE USE OF PERSONALITY TESTS FOR PERSONNEL SELECTION

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

We recently published an article in which we highlighted a number of issues associated with the use of self-report personality tests in personnel selection contexts (Morgeson et al., 2007). Both Ones, Dilchert, Viswesvaran, and Judge (2007) and Tett and Christiansen (2007) have written responses to this article. In our response to these articles we address many of the issues raised by Ones et al. and Tett and Christiansen. In addition to a detailed response, we make the following 4 key points: (1) Our criticisms of personality testing apply only to the selection context, not to all research on personality; (2) the observed validities of personality tests predicting job performance criteria are low and have not changed much over time; (3) when evaluating the usefulness of using personality tests to select applicants, one must not ignore the observed, uncorrected validity; and (4) when discussing the value of personality tests for selection contexts, the most important criteria are those that reflect job performance. Implications for personality testing research and practice are discussed.

We recently published an article in *Personnel Psychology* in which we highlighted a number of issues associated with the use of self-report personality tests in personnel selection contexts (Morgeson et al., 2007). This article grew out of a panel discussion conducted at the 2004 Society

Correspondence and requests for reprints should be addressed to Frederick P. Morgeson, Michigan State University, N475 North Business Complex, East Lansing, MI 48824-1122; morgeson@msu.edu

TABLE 1
Conclusions Reached by Morgeson et al. (2007)

1. 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.
2. 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.
3. Corrections for faking do not appear to improve validity. However, the use of bogus items may be a potentially useful way of identifying fakers.
4. 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.
5. 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.
6. 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.
7. Future research might focus on areas of the criterion domain that are likely to be more predictable by personality measures.
8. 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.

Note. Adapted from Morgeson et al. (2007). Reconsidering the use of personality tests in personnel selection contexts. PERSONNEL PSYCHOLOGY, *60*, 683–729. Reprinted with permission, Blackwell Publishing.

for Industrial and Organizational Psychology conference in Chicago. This panel session provoked some strong reactions from the audience, as has the recently published article. Both Ones, Dilchert, Viswesvaran, and Judge (2007) and Tett and Christiansen (2007) have written responses to our recently published article. Although they take very different approaches (and do not always agree in their criticism of our original paper or with each other), both articles express concerns about some of the points made in our article. This article seeks to address the points made by Ones et al. and Tett and Christiansen.

The best way to begin our response to Ones et al. (2007) and Tett and Christiansen (2007) is to start where we ended, that is, with the eight specific conclusions that were reached (see Table 1). There were three general threads in our original panel discussion and article. We first started out with a discussion focused primarily on "faking," (hence the title of the SIOP panel discussion, "Won't Get Fooled Again"), which is summarized in Conclusions 1–3 in Table 1. Then, someone pointed

out that faking was the least of the problems in this particular area of research, which initiated a discussion of alternative reasons why the uncorrected validities in operational contexts are so low (and perhaps not improving), summarized in Conclusion 4. This was followed by a discussion of what might be done to improve the situation, summarized in Conclusions 5–8.

With this as a starting point, much of what was written in response by Ones et al. and Tett and Christiansen can be seen as tangential, if not irrelevant. For example, the ability to predict "career satisfaction" or "leader emergence" from personality traits (and many of the rest of the entries in Table 1 of Ones et al., 2007) were not topics that were covered in our panel discussion, which was instead focused on observed job proficiency validities. The only rows in that table that were pertinent to our discussion would have been the uncorrected values for "overall job performance" and "objective performance," which we did not really need to see again because we all knew them from previous analyses. Similarly, Tett and Christiansen offer numerous reasons why they disagree with our conclusion that self-report personality measures have low validity. A number of these reasons (e.g., using Personality Oriented Job Analysis or narrow traits or multiple traits or interactions among traits) could also be viewed as methods that may help improve the overall performance of traits.

The conclusions and suggestions of Tett and Christiansen are potentially viable, but as yet we have very little supportive data. It is possible that averaging test validities and ignoring the signs of validity coefficients may underestimate validity, but as stated in several of the subsequent conclusions in this section, the specific measures (or constructs) and direction must be specified a priori, not after data are collected and examined as has often been the case. It is not clear whether the types of predictions suggested by Tett and Christiansen are driven by formally developed theory. For example, in citing the Christiansen et al. (1994) paper, they cite four traits that they think should be more highly related to success at upper-level supervisory positions than other 16PF traits and cite higher validities for measures of these traits. Perhaps this is the case, but were these traits specified as the target traits a priori and would other experts looking at the jobs and measures come to similar conclusions about the relevance of 16PF measures to these supervisory jobs? What is the underlying theoretical model specifying the linkages? The same concerns apply to the selection of narrow band versus broad band measures and job analysis. How many measures were examined in coming to the conclusion that some narrow band measures are relevant? The Tett and Christiansen suggestions that theory and configural use of trait measures may be more valid predictors are intriguing, but these are only possibilities at this point.

They cite no data on these issues, but some studies are encouraging (e.g., Barrick, Mount, & Strauss, 1993).

Although there may be disagreements about what avenues of future research are more or less promising, the most significant differences of opinion centered around our Conclusions 1– 4. That is, does the ability to fake responses to personality items explain why the correlations are so low, and are the correlations really that low? As with our original article, the current article gathers the opinions of former journal editors. Despite this diversity, the panelists have some common reactions to the Ones et al. and Tett and Christiansen articles. We begin our response by first discussing some overall conclusions about the use of personality tests in selection contexts and then provide some additional detail about our reactions to the Ones et al. and Tett and Christiansen articles.

Overall Conclusions

Conclusion 1: Our criticisms of personality testing applied only to the selection context, not to all research on personality.

Our original paper focused specifically on the validity of personality inventories for personnel selection and argued that, for this purpose, personality measures showed disappointingly low levels of validity. Scores on personality inventories are likely to be correlated with a number of variables that are of interest to organizational researchers, and we did not argue and do not believe that personality inventories are irrelevant for understanding behavior in organizations. We did argue, however, that they are poor predictors of criteria such as job performance and are difficult to justify as a basis for making high-stakes decisions about individuals.

Conclusion 2: The observed validities of personality tests predicting job performance criteria are low and have not changed much over time.

One curious fact about the validity of personality tests is that the average uncorrected validity has changed little over time. Table 2 details the uncorrected average correlations between personality measures and job performance criteria. Two conclusions are warranted based on this data. First, the magnitude of the correlation between various personality measures and job performance is quite low (ranging from −.02 to .15). Second, the magnitude of the correlation has been surprisingly consistent across time. In the first meta-analytic summary, Schmitt, Gooding, Noe, and Kirsch (1984) found an overall observed correlation between personality measures and performance ratings of .206, perhaps because they

TABLE 2
Uncorrected Average Correlations Between "Big Five" Personality Measures and Job Performance Criteria

Personality Measure	Hurtz and Donovan (2000)	Salgado (1997)	Barrick and Mount (1991)
Conscientiousness	.15	.10	.13
Extraversion	.06	.06	.06
Agreeableness	.07	−.00	.04
Emotional Stability	.09	.08	.04
Openness to Experience	.03	.00	−.02

included only data published in peer-reviewed outlets. In 1991, Barrick and Mount found that the correlations between the Big Five and job performance range from −.02 to .13. In 2000, Hurtz and Donovan found that the correlations between the Big Five and job performance range from .03 to .15. Little has changed in this time. There is significant controversy about the interpretation of corrected validities, but there is little controversy that the uncorrected validities of personality inventories as predictors of job performance reported in several existing meta-analyses are close to zero. These empirical data are something with which the field of industrial and organizational psychology must come to terms.

Although both Ones et al. and Tett and Christiansen provide various hypotheses regarding ways in which personality test validity might be enhanced (e.g., use of theory and job analyses to select predictors, using multiple versus single personality predictors, use of narrow traits), the cumulative data on these "improvements" is not great. In producing that database, researchers must be clear as to what data resulted from a priori hypotheses and what was based on post hoc attempts to interpret inconsistent and disappointing data.

Conclusion 3: When evaluating the usefulness of using personality tests to select applicants, one must not ignore the observed, uncorrected validity.

As noted in Conclusion 2, the observed validity of personality tests has changed little over time. What have changed, however, are the types of corrections that have been applied to the observed validity. This is no doubt responsible for the optimism expressed by many in the personality-testing field about the usefulness of personality tests for personnel selection. Yet, when organizations use personality tests, they do not correct the scores. They use the observed scores. Thus, any discussion of the viability of personality tests in the context of personnel selection should not ignore

the low levels of observed validity. In this context, it should be noted that we feel corrections for range restriction are appropriate when data are available indicating such restriction has occurred. Corrections for unreliability are usually much smaller and most agree are appropriate only when considering unreliability in the criterion.

Conclusion 4: When discussing the value of personality tests for selection contexts, the most important criteria are those that reflect job performance.

Given our exclusive focus on the selection context, the relationships between personality measures and other criteria (e.g., motivation, attitudes, leadership) do not constitute support for the use of self-report personality inventories in personnel selection. The most important criteria are those that reflect actual performance in the job.

The Validity of Personality Inventories for Predicting Job Performance

There are several points of disagreement between our original article and the responses presented by Ones et al. (2007) and Tett and Christiansen (2007), but the most important have to do with the validity of personality inventories as predictors of performance. Our analysis of univariate validities leads to the conclusion that personality inventories show disappointingly low validity. Multiple correlations presented by Ones et al. (2007) and Tett and Christiansen (2007) suggest a more optimistic set of conclusions. It is worth describing in some detail why we disagree with these more optimistic conclusions.

First, the objective nature of the averaging process of observed validities results in pretty clear consensus about what the actual uncorrected values are prior to any number of adjustments one would like to make after that (e.g., correct for measurement error, estimate multiple regressions, compound traits, employ absolute values, corrections for construct validity, and so on). According to Barrick and Mount (1991), Conscientiousness fares the best at .13, and Extraversion is second at .06.

Second, it is sometimes difficult to compare, in any straightforward way, univariate correlations and multiple correlations, in part because the multiple *R* sometimes produces weighting schemes that make no sense in substantive terms (e.g., negative weighting of some personality traits). In addition, such a "shot-gun" approach where every trait is used even if it cannot be theoretically linked to the nature of the work via a formal job analysis is a step backward to our much criticized "dustbowl" empiricism days. Thus, an approach that emphasizes trait relevance is an important

suggestion offered by Tett and Christiansen but one that has been ignored by Ones et al. (2007).

Third, the multiple *R*s reported by Ones et al. for the Big Five are not necessarily as impressive as they might appear. For example, Ones et al. (2007; Table 1) report *corrected* unit-weighted *R* values for all five factors as .23 and .20 for criteria of overall performance and objective performance, respectively. But later in Table 2 they report the corrected validity for Conscientiousness *all by itself* for those two criteria as .23 and .19, respectively. Thus, the correlation goes up by .01 points (and only for objective performance) when one adds the other remaining four variables, which does not strike us as a great trade off of degrees of freedom (4) and predictive accuracy (.01). In addition, the corrected *R* values do not seem all that impressive when compared to the corrected univariate validity estimates reported by Schmidt and Hunter (1998) for cognitive tests (.51), work samples (.54), structured interviews (.51), job knowledge tests (.44), peer ratings (.49), or job tryout procedures (.44). Even if we accept every correction and every assumption made by Ones et al. (2007), the conclusion that an optimal combination of the entire Big Five accounts for approximately 7% of the variance in overall performance and 5% of the variance in objective performance strikes us as a ringing confirmation of our belief that normal personality measures do not seem to have much value as predictors of job performance.

Fourth, as low as the multiple correlations between personality and performance are, there are good reasons to believe that these figures and estimates of incremental validity are overestimates. The meta regressions cited in these papers rely on an unrealistically low estimate of the intercorrelations among the Big Five. The estimated population intercorrelations among the Big Five are typically taken from the unpublished dissertation of Ones and are described as ranging from .00 to .27 (Ones, Viswesvaran & Reiss, 1996) with an average intercorrelation of only .15. There is clear evidence that the Big Five factors are not as orthogonal as suggested by these low intercorrelations (Block, 1995; Funder, 2001). For example, the revised NEO Personality Inventory (Costa & McCrae, 1992) reports domain–scale intercorrelations as high as .53 in self-ratings. The intercorrelations among the five factors appear even stronger in applicant samples (McManus & Kelly, 1999). Schmit and Ryan (1993) found that the five-factor structure fit a student sample but not an applicant sample. In the student sample the correlations among the five personality factors ranged from $-.28$ to .09 (mean = .014). In contrast, the six factors found for the applicant sample ranged from .53 to .04 (mean = .34). The factor with the most item-composite loadings in the applicant sample consisted of a group of items from four of the NEO subscales that reflected an "ideal applicant" self-presentation.

Fifth, corrections for range restriction or criterion unreliability are more credible when data about the extent to which there actually is range restriction or unreliability in the sample is available. In particular, corrections for range restriction are often made based on some assumed value or distribution based on data from a very small number of often unrelated studies or on the basis of a selection ratio in which it is the assumption that top down selection has occurred and all the top scoring candidates accepted the offered positions. Clearly, anyone who has worked in an organization realizes these assumptions regarding the selection ratio are unrealistic. Similarly, corrections for unreliability are made based on assumed distributions or values from similar studies. We do have more data on the typical reliability of performance measures, but there is continuing controversy about the conceptual adequacy of some operationalizations of reliability and the appropriateness of their use in the employment context (Murphy & DeShon, 2000a).

Sixth, we will not attempt here to resolve the ongoing debate over the best method of correcting for unreliability in the criterion, but there are conceptual arguments as to the appropriate estimate (Murphy & DeShon, 2000a). We will simply note that any problems that might be present when correcting a single correlation coefficient will necessarily be compounded when moving into the realm of multiple regression. For example, the corrected multiple correlation between the Big Five and overall job performance is a function of 15 separate corrected correlations (i.e., five univariate validity estimates and 10 estimates of intercorrelations among Big Five measures). It takes a great deal of faith in one's corrections to believe that all 15 are done correctly.

Seventh, there is a legitimate and substantive disagreement over the best way of correcting for measurement error when interpreting correlations (e.g., Murphy & DeShon [2000a, 2000b] present a detailed critique of the psychometric models typically used to correct for measurement error in validity generalization analyses). We would not characterize the statistical corrections as "magical" or "fanciful," as claimed by Ones et al. Rather, we would simply note that the appropriateness of the corrections used so freely in many of the meta-analyses on personality measures is still open to debate and criticism. Quantitative reviews are not free of subjectivity but require multiple judgment calls. Interestingly, even the authors of these two rebuttals apparently disagree on the extent of support for personality as a selection procedure and the appropriateness of the procedures used in their own respective meta-analyses (see Ones, Mount, Barrick, & Hunter, 1994; Tett, Jackson, Rothstein, and Reddon, 1994).

Eighth, the idea to use "compound traits" seems an unreasonable strategy (Ones et al., 2007) because in many ways it is a conceptually and

empirically weaker way to combine traits relative to just using the multiple correlation approach. In the introduction to his book on psychometric theory, Nunnally (1978, p. 4), speaks directly to the idea of "compound measures" noting that:

> "As this book will show in detail, each measure should concern some one *thing*, some distinct, unitary attribute. To the extent that unitary attributes should be combined to form an overall appraisal, e.g., of adjustment, they should be rationally combined from different measures rather than haphazardly combined within one measure."

Thus, Ones et al. (2007), contend that integrity is a compound trait made up of three different subtraits, including Conscientiousness, Emotional Stability, and Agreeableness. If this is true, then the best approach to combine them is to leave them independent and enter them into a regression equation to predict some criterion, not simply add up the polyglot of items into a single score. The regression approach is preferable because it maintains the uniqueness of each separate variable, it provides the optimal weight for each variable, and can unambiguously show how much each variable contributes (or fails to contribute) to prediction (yet recognize that such an approach can sometimes produce difficult to interpret regression weights). All of these desirable features are lost with the use of compound variables, and hence, results will generally be weaker and harder to interpret.

In our view, the analyses presented by Ones et al. (2007) lead to the same conclusion that we voiced in our original paper. Even if one makes the most optimistic assumptions about the low correlations among the Big Five and about the correctness of the entire string of corrections needed to reach the conclusion that the entire span of normal personality accounts for about 5% of the variance in job performance, one is left with the conclusion that about 95% of the variance in performance appears to have nothing to do with normal personality, as measured by currently available methods. This strikes us as an argument against relying too heavily on personality in selection.

Faking and Other Response Distortion

Ones et al. (2007) suggest that response distortion does not harm predictive validity and also question the generalizability to actual selection situations of "directed faking" studies with nonapplicants. For their part, Tett and Christiansen question some of the concerns about personality item ambiguity raised by Morgeson et al. We address each of these issues in turn.

First, there are questions about the frequency with which real applicants applying for real jobs in real settings actually attempt to fake personality inventories and the impact of faking on the validity of these measures. Ones et al. (2007) criticize Table 1 of Morgeson et al. (2007) for including directed faking or lab studies with nonapplicants. Although we agree that laboratory simulations will not provide an answer to this question, commonsense suggests that some applicants do fake their responses. As noted by Rosse, Strecher, Miller, and Levin (1998), "Personality testing . . . provides an almost ideal setting for dissimulation: Job applicants are motivated to present themselves in the best possible light; transparency of items makes it possible to endorse items that will make them look good, and there is little apparent chance of being caught in a lie." In addition, a recent meta-analysis comparing applicants in real selection situations to nonapplicants has shown that applicants do appear to inflate their scores on self-report personality inventories on job-relevant dimensions, and this inflation is more pronounced on direct measures of the Big Five than on indirect measures (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). Across all job types, applicants scored significantly higher than nonapplicants on Extraversion, Emotional Stability, Conscientiousness, and Openness.

A separate question is whether personality inventories can be faked, and on this point the findings of the laboratory simulations appear clear and consistent. They show that college students with little or no motivation to fake can, through simple instructions to look good, substantially elevate their scores on a personality inventory. Such vulnerability to distortion casts serious doubt on the soundness of such inventories in selection situations where the test takers are often very motivated to do well on the tests.

In their response, Tett and Christiansen defend commercial personality inventories and rebut Dipboye's comments about item ambiguity by pointing to the rigor of the psychometric scrutiny underlying their construction. In so doing, they really ignore Dipboye's primary point that outright faking is only one of the potential sources of response distortion on self-report personality inventories. The emphasis on faking oversimplifies a general and complex problem. A depiction of the test taker as faking assumes that the test taker responds to each item using the same frame of reference as the author of the test. Although some do define the item in a way consistent with the investigator's latent-trait continuum, there are a variety of potential dynamics at work including not only lying but also projection of an idealized concept of oneself, a lack of self-insight, and the projection of an image that could be true in the future or in a different setting even if it may not be true across all situations in the past. Kuncel and Kuncel (1995) estimate on the basis of previous research that a quarter to half

of response processes on personality inventories are inappropriate from the frame of reference of the investigator. Moreover, they characterize as "farfetched" the assumption that respondents faithfully adopt the frame of reference of the test author for each item as they proceed through an inventory. A more accurate characterization of test takers is that they are "struggling with the behaviors and feelings described within the item" using a frame of reference idiosyncratic to each test taker and the feelings and cognitions that have accumulated from answering previous items in the inventory (Kuncel & Kuncel, 1995, p. 189). Nunnally (1978) advised that those using a personality inventory subject them to protocol analysis in which the manner that respondents are interpreting questions is carefully explored. He noted that "when that is done, one is rather disturbed by the differences in meaning held by different subjects and by the extent to which all subjects are somewhat confused by some items" (p. 559). With a few exceptions (Fiske, 1968; Robie, Brown, & Beaty, 2007), this advice seems to have been ignored in favor of a black box, psychometric approach. In light of the complexity of the cognitive and interpersonal processes associated with personality test taking, it is not surprising that attempts to improve the low validities found in the prediction of job proficiency by reducing or correcting faking have yielded such disappointing results.

Tett and Christiansen also assert that "validity in an applicant context is likely to be compromised by informing applicants of what traits are being measured" (Conclusion #13; see also Conclusion #14). This is an apparent response to the speculations of Morgeson et al. (2007) that administering items from the same scale as a set might improve the structural validity of test items. Rather than agreeing or disagreeing with Tett and Christiansen, we would suggest that the jury is out on this question. The jury is also out on the broader issue of just how much to disclose to applicants about what is being measured. In partial support of disclosure, Hough, Eaton, Dunnette, Kamp, and McCloy (1990) reviewed research on the validity of subtle versus obvious items on personality inventories and concluded that "subtle items, often considered a unique virtue of external scale construction, are often less valid than obvious items, and may reduce scale validity" (p. 582). Based on these previous findings, Hough et al. (1990) report that the items on the ABLE inventory were written to "consist of obvious, rather than subtle, items that can readily be distorted" (p. 593). In addition, Johnson (2004) found in a nonselection context that item validity as measured by agreement between self- and other ratings is actually greater when the items are transparent indicators of the trait being measured. Some have even speculated that the testing context should be changed to gain the trust and cooperation of test takers (Aronson & Reilly, 2006; Fiske, 1968; Lovell, 1967). Given the low validities found with inventories that

disguise the constructs they purportedly measure, it would be interesting to empirically test whether open and honest test formats influence the quality of responses to personality inventories.

The Impact of Personality Testing on Applicant Attitudes

Throughout the Ones et al. and the Tett and Christiansen articles, they depict self-report personality inventories as having utility and usefulness in personnel selection. For instance, Tett and Christiansen conclude that "Commercial self-report personality tests yield useful validity in relations with job performance" and Ones et al. assert that "Even validities of .20 translate to substantial utility gains." Neglected in their analysis is the fact that usefulness and utility of a predictor is a reflection of a variety of factors in addition to validity (Jayne & Rauschenberger, 2000). Another of these dimensions of value is applicant reactions. Even if all of the claims of validity for personality inventories were true, we cannot ignore the potential for personality inventories to offend applicants, including those we wish to recruit. In a meta-analysis, Hausknecht, Day, and Thomas (2004) conclude that personality tests ($M = 2.88$), along with biodata ($M = 2.81$), personal contacts ($M = 2.51$), honesty tests ($M = 2.47$), and graphology ($M = 1.76$), are among the least favorably evaluated selection techniques. In their survey of managers, Smither, Reilly, Millsap, Pearlman, and Stoffey (1993) found that a relatively innocuous personality inventory (the Managerial Potential Scale of Gough, 1984) ranked the *lowest* of the 14 measures (mean favorability of 2.84) with only 35.3% agreeing and 42.5% disagreeing that it was job related. Tied with personality for the least favorably rated instrument was a biodata measure containing self-reports of temperament. It is also interesting to note that in these types of surveys, integrity tests are usually among the least favorably received, especially when they are in the form of personality tests (Whitney, Diaz, Mineghino, & Powers, 1999). In light of evidence that applicant perceptions about selection are related to evaluations of the organization, intentions to accept job offers, and recommend the employer to others (Hausknecht et al., 2004), negative reactions to personality inventories must be taken as seriously as validity data. Of course, there are a variety of factors that should be evaluated in assessing the utility, usefulness, or value of a selection procedure including legal defensibility, cost of purchasing and administering the procedure, process flexibility, alignment with diversity and affirmative action goals, candidate flow statistics, selection ratios, and cycle time to fill a position (Jayne & Rauschenberger, 2000). To our knowledge there are no rigorous and comprehensive evaluations of the utility of self-report personality inventories. However, it is our opinion that the low predictive validities reported by Ones et al.

and Tett and Christiansen combined with the adverse effects on recruiting and organizational image of negative applicant reactions pose substantial barriers to demonstrating the "usefulness" of self-report inventories.

Does Faking Have a Major Impact on Criterion-Related Validities?

With respect to whether faking has a major impact on the criterion-related validity, it is clear that the two articles that were written in response to our article come to different conclusions. Ones et al. (2007) dismiss faking and social desirability as factors that have a major influence on the criterion-related validity of personality traits. Tett and Christiansen (2007) see faking as more of a problem. As is often the case, the data on this topic do not "speak for itself," and thus, two informed parties like Ones et al. and Tett and Christiansen could look at the same data and come to different conclusions.

Within our panel discussion, this issue came up when Campion stated in his review of the 18 studies that examined this issue, 8 of 18 articles found that distortion affected validities. Hollenbeck asked whether "affected" in this case meant that the validity in one case was tested against the other and the difference between the two correlations was statistically significant, because this is rarely if ever shown directly. Sampling error alone will assure that the correlations are not the same to two decimal places, so one has to ask if the difference is larger than what might be expected due simply to sampling error. This is a difficult test for this literature to pass because it would require very substantial changes in rank orderings between conditions to generate correlations that would differ from each other by that amount given the usual sample sizes available. In this context, a recent article by Schmitt and Oswald (2006) is relevant. They showed that, given the usual level of observed personality test validity and the correlations between faking measures and personality and faking measures and criteria, it is statistically impossible to recognize any great difference between observed test validity and validities corrected for the measured tendency to fake on the part of test takers.

Tett and Christiansen report the results of two meta-analyses that deal with this. A meta-analysis by Tett, Jackson, Rothstein, and Reddon (1999) shows that the uncorrected validities were higher for applicant samples (where faking is a concern) versus incumbent samples (where faking should not be a concern), where the difference was .20 versus .15 in favor of applicants. In contrast, however, a meta-analysis by Hough (1998) reports that the effect for incumbents was larger than the effect for applicants .09 to .06. Although one could examine these two meta-analyses to try to ascertain which one is really right, or one could just split the difference, the key is to study the numbers themselves and how they relate to our original

question. When you examine the size of the values and their differences, and then return to the question "is faking the reason why validities are so low?" one comes to the conclusion that this is clearly not *the* problem. This also explains why our panel discussion left this issue behind pretty early in the session.

In making his point about this, Hollenbeck suggested that faking may "basically be a constant" that shifts the entire score distribution up by a constant amount, and Tett and Christiansen (2007) show that this cannot be the case because people low on the trait have more to gain from faking versus people high on the trait. This may be true, but even this example helps illustrate why it is unlikely for faking to grossly distort validities. If there were 20 Conscientiousness items and a perfectly conscientious individual responded to all 20 the right way, he or she indeed would not be able to gain anything from faking. However, it is also true that no one could ever pass this person in any rank ordering sense (he or she would always be at the top) and that it would be impossible for this person to work their way to the bottom of the distribution. The distance between this person and others may get closer, and some shifting may occur in the middle, but the people at the bottom are still likely to be lowest on the trait (because they do not have it and cannot fake it).

One last point that needs to be raised is that the argument is often shifted away from the impact that faking has on criterion-related validities and instead focused on the outcomes experienced by any single applicant. That is, depending upon where the cut is made, any shift in rank ordering could wind up hurting an individual applicant who failed to intentionally or unintentionally fake relative to another individual. Sometimes this argument is used to justify the need to find ways to eliminate faking by those who recognize it does not impact criterion-related validities. Sometimes these arguments are used by those who believe that there are no feasible alternatives for eliminating all forms of faking and, thus, wish to rule out the use of these measures altogether. This is a shift in orientation from examining the value of testing to an institution that makes a large number of decisions to the value of testing to an individual applicant who is the victim of a single decision.

When one makes this shift, it is important to remember that the standard error associated with *any single prediction* from regression models is very, very high, even with multiple correlations much larger than those being discussed here. Readers who do not believe this should insert various values in the formula for calculating this standard error (see Cohen, Cohen, West, & Aiken, 2003, p. 95 for this formula). Over many, many decisions, even small correlations can have significant value for institutions, but currently, given the typical validities we see in selection contexts, the odds for any one individual are not greatly

affected for any single decision by even our most effective selection methods.

Suggestions for Future Research and Practice Using Personality Measures as Selection Tools

Looking forward, there are several suggestions we would make for future research on personality measures as selection tools.

First, we would strongly urge complying with the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) in the reporting of data. "Both uncorrected and corrected values should be presented when corrections are made for statistical artifacts such as restriction of range or unreliability of the criterion (p. 52)." We applaud the decision of Tett et al. and Ones et al. to include uncorrected as well as corrected correlations in their summaries. Further, it should be clear which corrections are made based on assumptions and which corrections are made based on actual data collected on the applicants and test under consideration.

Second, we feel that contextualizing personality measures or developing and using custom-developed tests might solve some of the problems associated with the current use of self-report personality tests. The idea of contextualizing the items by adding "at work" to each strikes us as a potential way to achieve better empirical results (see Schmit, Ryan, Steirwalt, and Powell, 1995). Regardless of whether one feels it would work or not, it is one technique that could be tested in future research. It would certainly make it easier to defend personality inventories to applicants and would perhaps reduce the negative reactions to these instruments.

It is possible to take this even further and write custom-developed tests specifically for the job and organization in question. With some exceptions, commercially available tests are not necessarily superior to tests developed by in-house psychologists or consultants for client organizations. A custom-developed test strategy is extremely valuable because it may open up content validation as a potential validation strategy (which is often very difficult with commercial tests). The ability to be able to fall back on a content validation strategy is valuable because even if one were to take the corrected validities reported by Ones et al. (2007), in most operational contexts, there will not be enough people to provide sufficient statistical power to get results that some critic (or lawyer) would not be able to attribute to sampling error. That is, if one takes the corrected r value of roughly .20 for Conscientiousness alone or the corrected R of roughly .25 for the all five variables together (again, these values were taken from the rows dealing with overall performance and objective performance from Ones et al. Tables 1 and 2) to have statistical power of .80, to be able

to detect these effects would require 193 and 240 research participants respectively. Many organizations do not hire that many people a year, let alone that many people for the same job category. Thus, sample sizes this large for individual jobs will not be available in most operational contexts. Furthermore, because custom-made tests can be made to appear more job related, candidate reactions may be more positive. Many traditional commercial tests contain questions that seem irrelevant to candidates and can embarrass the organization. Importantly, the recommendation to use custom-developed tests was made over 40 years ago in the discussion in Guion and Gottier's (1965) review of personality test validity. Regardless of whether you use a commercially available test or a custom-developed test, however, it is important to evaluate the quality of the instrument based on the evidence available and the qualifications of the test developer.

Third, we would limit the criterion measures to those most relevant to the question of using personality inventories to select among applicants for jobs. Ones et al. (2007) exaggerate the support for the use of personality inventories in personnel selection in their Tables 1 and 2 by expanding the criterion domain far beyond what can be used to judge the validity and usefulness of personality variables for personnel selection. For example, Table 2 reports meta-analyses examining the relationship of personality to "motivational criteria," "leadership criteria," "work attitudes," and other criteria measured with highly subjective methods. Although they may be distally related to the types of criteria of primary concern to those using personality measures for selection, one cannot conclude from findings of a relationship to personality that personality is a means of selecting effective employees.

Take, for example, the leadership criteria used in the meta-analysis of Judge, Bono, Ilies, and Gerhardt (2002) and reported in Table 2. As an example of what they coded as leadership effectiveness, they cite a study using a measure of perceived influence in which subordinates indicated "how much influence they felt their supervisor had on the productivity and overall effectiveness of their unit" (Judge et al., 2002, p. 769). Obviously, perceived influence is not equivalent to effectiveness, and showing that there is a correlation of a personality dimension with perceived influence does not provide a strong basis for use of this measure to select managers who will be effective. Another example is in the meta-analysis of Judge and Ilies (2002), which is reported in Ones et al. Table 2. Judge and Ilies (2002) found a corrected relationship to personality of .21 with expectancy. To infer from this finding that we are justified in using personality to select employees is a long stretch indeed given the low relationship of expectancy to performance (only .21 according to the meta-analysis of Van Eerde & Thierry, 1996).

Fourth, we would further suggest limiting the criterion measures to those that are objective and based on reports independent of the predictor measures. The use of self-reports to measure both the predictor (personality) and the criteria runs the risk of inflating the estimates of validity. For instance, Organ and Ryan (1995) found that Conscientiousness is positively related to employee altruism but only when self-reports of altruism were used. With self-reports the level of association was .449 and compared to only .043 with independent ratings.

Fifth, only studies that use a predictive model with actual job applicants should be used to support the use of personality in personnel selection. We base this recommendation on previous research showing important differences between applicants and nonapplicants on personality measures (e.g., Birkeland et al., 2006; Schmit & Ryan, 1993; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Although Ones et al. cite research to support their position that criterion and construct validity generalizes across applicant and nonapplicant settings, the research so far constitutes a mixed bag that provides little basis for strong conclusions.

This is the case for at least three reasons. First, many of the studies Ones et al. cite used a single personality questionnaire that was not designed to measure the Big Five (e.g., CPI, Personal Preferences Inventory, 16PfI). As Smith, Hanges, and Dickson (2001, p. 313) noted, "although we have interpreted the results to indicate support for the invariance of the FFM [five factor model], the FFM cannot be disentangled from the instrument designed to measure it (in this case a reduced version of the HPI)." Second, most studies have used multigroup confirmatory factor analysis procedures, which may not be the most appropriate approach to testing the impact of faking on construct validity or the invariance of the FFM (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Stark et al., 2001). Third, in most of the cited studies, the conditions of inventory administration were of doubtful relevance to realistic selection settings where external applicants are competing for jobs in an organization. Ellingson, Sackett, and Connelly (2007, p. 394) specifically noted as a potential limitation "the fact that the majority of the selection assessments were conducted for internal selection purposes" and this "may have altered the degree of distortion observed if motives to distort differ between internal and external selection contexts."

Based on this research, we do not share Ones et al.'s optimism that construct and criterion-related validities of personality measures generalize from nonapplicants (e.g., students and current employees) in nonselection settings to applicants in realistic selection situations. Rather than assuming that findings will generalize, we believe at this stage of the research that the prudent approach is to use applicant samples.

Conclusion

Our fundamental purpose in writing these articles is to provide a sobering reminder about the low validities and other problems in using self-report personality tests for personnel selection. Due partly to the potential for lowered adverse impact and (as yet unrealized) increased criterion variance explained, there seems to be a blind enthusiasm in the field for the last 15 years that ignores the basic data. There is considerable evidence to suggest that when predictive validation studies are conducted with actual job applicants where independent criterion measures are collected, observed (uncorrected) validity is very low and often close to zero. This is a consistent and uncontroversial conclusion. Although numerous meta-analytically based corrections may increase validity estimates, not all scholars agree with the legitimacy of these corrections. Some of the co-authors on this article believe that in light of these problems, Guion's (1965, p. 379) comments from over 40 years ago are still true today: "In view of the problems. . .one must question the wisdom . . . of using personality as instruments of decision in employment procedures." Other co-authors are somewhat more optimistic and believe that personality tests may have value in some situations and with proper research methods. However, all of us would agree with Guion's further comments that, "Research must continue, but it should be basic research defining and classifying traits and discovering how a job applicant's personality relates to the personality he reveals later on the job." It is our hope that the Morgeson et al. (2007) article and the debate it caused prompts scholars to conduct new, theory-driven empirical research into the issue of why personality test validities have been so historically low. All of the response articles provide numerous ideas for future research. Instead of simply concluding whether personality tests are useful or not, fundamental research should be conducted on the issues raised in this series of articles to advance the state of scientific knowledge on this topic.

REFERENCES

Aronson ZH, Reilly RR. (2006). Personality validity: The role of schemas and motivated reasoning. *International Journal of Selection and Assessment, 14*, 372–380.

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, Strauss JP. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology, 78*, 715–722.

Birkeland SA, Manson TM, Kisamore JL, Brannick MT, Smith MA. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–334.

Block J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*, 187–215.

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

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

Costa PT, McCrae RR. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Ellingson JE, Sackett PR, Connelly BS. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*, 386–395.

Fiske DW. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research, 3*, 393–401.

Funder DC. (2001). The really, really fundamental attribution error. *Psychological Inquiry, 12*, 21–23.

Gough HG. (1984). A managerial potential scale for the California Psychological Inventory. *Journal of Applied Psychology, 69*, 233–240.

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

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

Hausknecht JP, Day DV, Thomas SC. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. PERSONNEL PSYCHOLOGY, *57*, 639–683.

Hough LM. (1998). Personality at work: Issues and evidence. In Hakel M (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Hillsdale, NJ: Erlbaum.

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.

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

Jayne MEA, Rauschenberger JM. (2000). Demonstrating the value of selection in organizations. In Kehoe J (Ed.), *Managing selection in changing organizations* (pp. 123–157). San Francisco: Jossey-Bass.

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

Judge TA, Bono JE, Ilies R, Gerhardt M. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765–780.

Judge TA, Ilies R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology, 87*, 797–807.

Kuncel RB, Kuncel NR. (1995). Response-process models: Toward an integration of cognitive-processing models, psychometric models, latent-trait theory, and self-schemas. In Shrout PE, Fiske ST (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 183–199). Hillsdale, NJ: Erlbaum.

Lovell VR. (1967). The human use of personality tests: A dissenting view. *American Psychologist, 22*, 383–393.

McCrae RR, Zonderman AB, Costa PT, Bond MH, Paunonen SV. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566.

McManus MA, Kelly ML. (1999). Personality measures and biodata: Evidence regarding heir incremental predictive value in the life insurance industry. PERSONNEL PSYCHOLOGY, *52*, 137–148.

Morgeson FP, Campion MA, Dipboye RL, Hollenbeck JR, Murphy K, Schmitt N. (2007). Reconsidering the use of personality tests in personnel selection contexts. PERSONNEL PSYCHOLOGY, *60*, 683–729.

Murphy KR, DeShon RP. (2000a). Inter rater correlations do not estimate the reliability of job performance ratings. PERSONNEL PSYCHOLOGY, *53*, 873–900.

Murphy KR, DeShon RP. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? PERSONNEL PSYCHOLOGY, *53*, 913–924.

Nunnally JC. (1978). *Psychometric Theory* (2nd ed). New York: McGraw-Hill.

Ones DS, Dilchert S, Viswesvaran C, Judge TA. (2007). In support of personality assessment in organizational settings. PERSONNEL PSYCHOLOGY, *60*, 995–1027.

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

Organ D, Ryan K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behaviors. PERSONNEL PSYCHOLOGY, *48*, 775–802.

Robie C, Brown DJ, Beaty JC. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489–509.

Rosse JG, Strecher 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.

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

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.

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.

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.

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, 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.

Smith DB, Hanges PJ, Dickson MW. (2001). Personnel selection and the five-factor model: Reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology, 86*, 304–315.

Smither JW Reilly RR, Millsap RE, Pearlman K, Stoffey RW. (1993). Applicant reactions to selection procedures. PERSONNEL PSYCHOLOGY, *46*, 49–76.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the Validation and Use of Personnel Selection Procedures* (4th ed). Bowling Green, OH: Author.

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

Tett RP, Christiansen ND. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. PERSONNEL PSYCHOLOGY*, 60*, 967–993.

Tett RP, Jackson DN, Rothstein M, Reddon JR. (1994). Meta-analysis of personality-job performance relations: A reply to Ones, Mount, Barrick, and Hunter (1994). PERSONNEL PSYCHOLOGY*, 47*, 157–172.

Tett RP, Jackson DN, Rothstein M, Reddon JR. (1999). Meta-analysis of bi-directional relations in personality-job performance research, *Human Performance, 12*, 1–29.

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

Whitney DJ, Diaz J Mineghino MAE, Powers K. (1999). Perceptions of overt and personality-based integrity tests. *International Journal of Selection and Assessment, 7*, 35–45.