Journal of Applied Psychology
2006, Vol. 91, No. 3, 613–621

0021-9010/06/$12.00 DOI: 10.1037/0021-9010.91.3.613

The Impact of Corrections for Faking on the Validity of Noncognitive Measures in Selection Settings

Neal Schmitt and Frederick L. Oswald
Michigan State University

In selection research and practice, there have been many attempts to correct scores on noncognitive measures for applicants who may have faked their responses somehow. A related approach with more impact would be identifying and removing faking applicants from consideration for employment entirely, replacing them with high-scoring alternatives. The current study demonstrates that under typical conditions found in selection, even this latter approach has minimal impact on mean performance levels. Results indicate about .1 *SD* change in mean performance across a range of typical correlations between a faking measure and the criterion. Where trait scores were corrected only for suspected faking, and applicants not removed or replaced, the minimal impact the authors found on mean performance was reduced even further. By comparison, the impact of selection ratio and test validity is much larger across a range of realistic levels of selection ratios and validities. If selection researchers are interested only in maximizing predicted performance or validity, the use of faking measures to correct scores or remove applicants from further employment consideration will produce minimal effects.

Keywords: faking, noncognitive measures, corrections for faking, mean performance, selection ratios and validity

Research on personality tests in personnel selection and their use in applied contexts has increased rather dramatically since the publication of the meta-analysis by Barrick and Mount (1991) that indicated that some personality measures displayed generalizable and practically useful validities in predicting job performance. Along with the increased interest in personality measurement came concerns about the degree to which faking on these measures might influence both validity and the decisions made about individual job applicants. The general concerns about faking are certainly not new in personality measurement, as most of the clinically based instruments that were developed more than a half century ago, such as the Minnesota Multiphasic Personality Inventory and 16 Personality Factor (Cattell, Cattell, & Cattell, 1993), include lie and fake scales. Responses to these scales are often used to make corrections to individuals' scale scores on personality traits of interest or, in some instances, to discount completely the individuals' responses. Literature on faking has been reviewed extensively by Paulhus (1991), who concluded that two major dimensions accounted for most of the variance in various social desirability measures, though intercorrelations between the various measures of social desirability were often low. A *self-deception* factor was considered a "normal" and nondeliberate tendency to present oneself positively, whereas *impression management* was a deliberate attempt to present oneself in a particular manner to achieve some desirable outcome, such as a job offer in a personnel selection context.

Neal Schmitt and Frederick L. Oswald, Department of Psychology, Michigan State University.

Correspondence concerning this article should be addressed to Neal Schmitt, Department of Psychology, Michigan State University, East Lansing, MI 48824-1116. E-mail: schmitt@msu.edu

There seems to be little doubt that personality measures can be faked. Ones and Viswesvaran (1998) reported the results of a meta-analysis of studies in which the scores of examinees who were instructed to "fake good" were compared with those of examinees who were given normal test instructions. Mean scores for the fake-good examinees were about .60 *SD*s higher than they were for those given normal instructions. Within-subjects designs, in which the same participants responded under normal and fake-good conditions, produced slightly higher mean differences of .72 *SD* units on measures of the Big Five constructs. One would expect that real-world data comparing mean applicant and incumbent responses to personality tests would produce smaller mean differences than in these lab conditions but that differences should still exist: Applicants tend to be motivated to get jobs and do actually respond in ways that inflate their scores relative to incumbent groups that would not have the same motivation. Research on this issue is somewhat mixed, however. Hough (1998) found a great deal of variability in mean applicant–incumbent differences across scales and across three different samples, with effect sizes ranging from approximately zero to .50 *SD*s. These differences are consistent with Hough, Eaton, Dunnette, Kamp, and McCloy's (1990) comparison of applicant and incumbent groups.

In addition to presenting themselves favorably on personality instruments, applicants have also been shown to present themselves in a favorable manner on other noncognitive measures such as biodata (Lautenschlager, 1994). Given that faking exists in these self-report measures, an obvious question relates to the impact such faking has on the use of these measures as decision-making tools in selection, and if faking can be identified, what the impact would be from implementing corrections for faking.

Effects of Faking

There are at least two concerns about the presence of faking when noncognitive measures are used. First, researchers are concerned that criterion-related validity will be adversely affected when examinees are faking. This concern has led to the practice of statistically partialing out the effects of social desirability when estimating the relationship between predictors and criteria. This practice is predicated on the notion that faking is a suppressor variable and partialing involves the removal of unwanted trait variance to provide better, and higher, estimates of criterion-related validity. However, some investigators have suggested that aspects of social desirability are related to substantive job-related personality traits such as adjustment, conscientiousness, and emotional stability and to desirable criterion variables (Cunningham, Wong, & Barbee, 1994; Ones, Viswesvaran, & Reiss, 1996; Zerbe & Paulhus, 1987). If this is the case, then computing a partial correlation or otherwise correcting for social desirability might lead to partialing out predictor-relevant and/or criterion-relevant variance, which consequently would distort criterion-related validity estimates.

Actual research comparing corrected and uncorrected criterion-related validities suggests that measures of social desirability typically do not have any great impact on criterion-related validity. Most such comparisons yield differences between corrected and uncorrected validity coefficients much less than .10 (e.g., Barrick & Mount, 1996; Christiansen, Goffin, Johnston, & Rothstein, 1994; Hough, 1998; Hough et al., 1990). In a meta-analysis directed to an estimation of the impact of desirability, Ones et al. (1996) concluded that social desirability is related to emotional stability and conscientiousness but that it does not serve as a useful predictor of job performance (with the possible exception of training performance), nor does it serve as a moderator or suppressor of criterion-related validity.

Although the impact of social desirability may have little effect on criterion-related validity, recent research does indicate that it may have significant impact on who gets hired in some situations. That is, employers may be hiring individuals who would not be hired if their scores were not influenced by some aspect of socially desirable responding. The focus of the current study is to estimate directly what differences in the quality of the workforce might be if employers could remove individuals whose high scores on selection instruments are likely the result of some form of response distortion (whether that takes the form of social desirability or some other form).

Impact of Corrections for Faking on Selection Decisions

Perhaps the first effort to examine the effect of correcting personality test scores on individual selection decisions was performed by Christiansen et al. (1994). Although they found little impact on validity when scores were corrected with the standard 16 Personality Factor formula for correction or with partial correlation analyses, they did find that corrections resulted in different hiring decisions in about 15.0% of the cases when selection ratios were less than .25. When the selection ratio was larger, however, the differences in hiring decisions were smaller when based on corrected scores as opposed to observed scores. Likewise, rank ordering the top candidates produced different results when based on corrected personality test scores as opposed to a rank ordering based on uncorrected scores.

In a similar study, Rosse, Stecher, Miller, and Levin (1998) examined the effect of response distortion among applicants for positions in a property-management firm. Defining response distortion as a score of at least three standard deviations above the mean of a group of incumbents for the same position, Rosse et al. found that at selection ratios of .05, .10, and .20; 88.0%, 56.0%, and 44.0% of the selected applicants, respectively, would have had their Conscientiousness scores classified as suspect. These proportions dropped significantly as the selection ratios increased, just as in the Christiansen et al. (1994) study.

Zickar, Rosse, Levin, and Hulin (1996) approached the same question by using a simulation. Manipulating the degree of faking by adding values to unfaked latent scores, the correlation between the test and the criterion, and the percentage of examinees who faked, Zickar et al. found that even with moderate amounts of faking, the mean of the predictor in the faking group was .30 to .40 *SD*s higher than the mean of the group that was honest. Mean differences between faking and honest groups on the criterion were between .10 and .20 *SD*s. The simulation also confirmed findings from previous work that indicated little decrement in criterion-related validity in the presence of even extreme faking.

Hough (1998) presented results from three large samples that compared two strategies for removing examinees suspected of faking. In the first strategy, the examinees' scores on the substantive trait of interest were reduced if they scored in the top 2.5% on an Unlikely Virtues scale designed by Hough et al. (1990). The second strategy was more stringent than the first, removing the top 5.0% of the examinees on the Unlikely Virtues scale. Use of these corrections resulted in trait score means that were more similar to those of an incumbent group. Particularly at low selection ratios, a very different set of applicants would be selected if the corrections were applied. Both strategies reduced the impact of assumed distortion with no change in criterion-related validities. Hough provided some caveats in the use of these strategies. First, one must assume that the faking scale one uses is actually a measure of the degree to which applicants distort their responses. Second, if the distortion scale correlates with job performance (Hough suggested that a validity of .15 be used as the cutoff), then the corrections should not be used, presumably because one may be removing applicants whose predicted job performance is high also.

Ellingson, Sackett, and Hough (1999) used a regression-based estimate of examinee faking to adjust personality scale scores. In a repeated-measures design, they compared responses under an honest condition, under instructions to fake in a "job-desirable" manner, and under a condition in which the faked responses were corrected. The corrections did produce mean estimates of trait scores that were very nearly the same as the mean trait scores obtained under honest conditions. However, when they compared the actual examinees who would be selected by using corrected and uncorrected scores as well as under varying selection ratios and assumptions about the proportion of the sample distorting its scores, they observed no consistent pattern indicating that the correction improved the proportion of honest respondents who would actually be selected.

In the most recent attempt to evaluate the impact of faking corrections on actual selection decisions, Mueller-Hanson, Heggestad, and Thornton (2003) designed an experiment in which

participants provided answers to a personality instrument under either control or incentive conditions. In contrast with the control condition, the incentive condition received instructions designed to mimic the motivational set in an actual employment situation. The mean difference on the personality measure across the two groups was .41 *SD*s. By using the personality measure scores to rank order and select from the entire group combined, Mueller-Hanson et al. then compared the proportion of the honest group that would be selected at different selection ratios as well as their mean performance on a subsequent performance task that served as a criterion. As the selection ratio decreased, the proportion of honest, or control group, examinees (36.0% at selection ratio of .10) became substantially less than the proportion of the incentive group (64.0%). Differences in mean performance also tended to increase as the selection ratio decreased (Cohen's $d = .56$ at selection ratio of .10). These differences between control and incentive condition may have been exaggerated in this study because criterion-related validities in the two conditions were also different; as has been shown, previous research indicates little evidence for a validity difference between applicant and incumbent conditions.

Summary

These studies provide the basis for several conclusions. First, faking does not seem to influence criterion-related validity in studies other than those in which respondents are directed to fake (Hough, 1998). Second, attempts to correct for faking may seem to work at an aggregate level in that means of corrected and incumbent, or honest, study participants are similar. Third, faking can produce very different decisions about the specific individuals selected, and attempts to correct measures of substantive traits by using scores on faking measures are not effective (Ellingson et al., 1999). The degree to which faking influences the quality of these decisions is a function of the selection ratio and likely some interaction of the criterion-related validity of the trait measure and the correlations between any measure of faking with both the predictor and the criterion.

Study Purpose and Hypotheses

The purpose of the present study was to add to this small body of research in two ways. First, we evaluate directly the impact that corrections for faking have on mean performance levels in what we consider a range of realistic situations in which faking corrections might be applied. If an organization uses a noncognitive fakable measure and a measure of faking such as the Unlikely Virtues scale used by Hough (1998), one approach is to eliminate some portion of those who score highest on the faking measure. These individuals are then replaced with persons who have the highest possible scores on the trait measure yet whose scores on the faking measure do not indicate that they are distorting their answers to make themselves look like more worthy candidates than they are. Another common approach is to correct the trait scores by some amount on the basis of the respondent's score on the faking measure. It should be noted that the former approach of removing and replacing applicants is the most extreme form of trait score correction; the latter approach of correcting applicant scores leads to smaller effects on the expected mean performance of those selected. Even though previous research has indicated that faking corrections have little impact on validity, they continue to be used by those using personality tests (Goffin & Christiansen, 2003) on the common presumption that "validity will be improved by using score corrections" (p. 343). The impact of the faking correction in the present study was defined as the difference between (a) mean performance based on a rank order of scores on the trait measure and (b) mean performance based on those same scores, except that faking applicants are removed and replaced by nonfaking applicants with the highest scores. This strategy is similar to Hough's (1998) second strategy, although it is unclear whether faker applicants were replaced with the next-highest-scoring individuals on the trait measure when estimating the impact of corrections. Furthermore, previous studies have examined only the mean performance of those selected and have not included the next-highest-performing nonfakers to replace those suspected of some form of faking (e.g., Mueller-Hanson et al., 2003; Zickar et al., 1996). We believe that the impact of faking corrections will be much smaller when viewed in this light than as portrayed in previous studies. Further, we also believe that this represents a more realistic scenario than has been applied in some previous studies of the impact of faking and faking corrections.

Second, we examine simultaneously the role of five factors that seem to play a role in the impact that faking has on selection decisions, namely the selection ratio, validity of the predictor variable, the correlation between a faking measure and the predictor, the correlation between the faking measure and the criterion, and the proportion of the candidates considered to be faking. These factors are evaluated across different proportions of people whose responses are considered distorted. The role that these factors play in selection outcomes is generally well known, but the interactions of these factors with those factors that represent the characteristics and use of a faking measure are either unknown or have not been systematically evaluated. Although we believed that all factors, independent of validity and selection ratio, would have only small impact on expected performance, the previous literature and practice provided the basis for the following expectations:

1. The selection ratio will affect the impact of faking corrections such that lower selection ratios will lead to larger differences in mean performance between a group in which no applicant is removed for suspected faking and a group in which some proportion is removed for suspected faking.

This proposition makes sense if you consider the extreme case: In the case where the selection ratio is 1.0 and all applicants are selected, then any adjustment of predictor scores would have no influence on mean performance.

2. The greater the criterion-related validity of the test, the more positive will be the impact of a faking correction on mean performance.

This should be the case because faking has a greater potential to affect mean performance in a negative fashion when validity is high than when validity approaches zero.

3. When the correlation of the faking measure and the predictor is positive, we expect that the impact of a

faking correction, holding everything else constant, will be to reduce mean performance. When the correlation between the faking measure and the predictor is negative, we expect the impact of the correction will be to increase mean performance.

Note that a positive faking–predictor correlation is thought to occur in the case of some personality constructs (e.g., Conscientiousness, Emotional Stability). A negative faking–predictor correlation is thought to occur when those with lower trait scores are motivated to fake to make up their deficiencies. Partialing out the unwanted trait variance associated with faking is one solution used to get the best estimate of criterion-related validity in this instance.

4. Positive correlations between the faking measure and the criterion, after applying a correction measure, will tend to reduce mean performance because the correction for faking will thus have the potential to remove highly qualified applicants.

5. The proportion of people identified as fakers will be associated with reduced mean performance that results from the correction for faking. This assertion is consistent with previous studies (e.g., Zickar et al., 1996).

We also expected that one or more of the five expectations just specified would interact with another to affect mean performance (i.e., mean performance would depend on the levels of more than one factor). However, there is no previous literature or analyses on which to base any predictions that are more focused, and therefore analyses involving interactions would be exploratory in nature. In general, we believed that the correction for faking could not have any large or practically significant impact on expected performance given a wide range of realistic scenarios. We should point out that the impact of removing and replacing applicants from further consideration for selection must be larger than the impact of any partial correction to the trait measure used to make decisions. It is also the case that our analyses implicitly assume that the faking measure is a perfectly construct-valid measure, as does similar earlier work cited above on the impact of faking corrections. However, independent of the theoretical importance for the construct validity of measures (whether of faking, predictors, or criteria), our simulations reflect a range of typical conditions in which the use of faking measures is often justified by concerns for criterion-related validity.

Method

Simulation Conditions

Levels within each of six factors in the simulation design (see Table 1) were completely crossed with the levels of all other factors, yielding a total of 2,304 conditions. Within each condition, the simulation was replicated 1,000 times to yield an estimate of sampling error variance for mean performance. Results for each condition reflect averages across replications. The six factors were as follows:

Sample size. Sample sizes were 50, 100, 250, and 500, which in selection research represent a range from relatively small to relatively large. We did not suspect that sample size would have a systematic influence on results but rather that it would be inversely related to the amount of sampling-error variance found in mean performance. The emphasis of this study is on the pattern of results across conditions that would typically be found, not on the sampling-error variance of the estimates; however, those results are available by request from Neal Schmitt.

Table 1
Simulation Conditions

Parameter	Values
N	50, 100, 250, 500
Criterion-related validity	.10, .20, .40
Predictor–faking correlation	−.20, .00, .20, .40
Faking–criterion correlation	−.10, .00, .10, .30
Selection ratio	.10, .20, .50
Proportion removed for suspected faking	.00, .05, .15, .30

Note. There are 2,304 simulation conditions in a fully crossed design of the values listed above. Each simulation condition was replicated 1,000 times, taking the average mean performance across replications within each condition.

Criterion-related validity. Given that the simulations focus on criterion-related validities for noncognitive predictors that are subject to faking in high-stakes selection settings, simulation values for criterion-related validity ranged from relatively small to moderately large: r_{xy} = .10, .20, and .40, where x represents the predictor and y the criterion.

Predictor–faking correlation. As already mentioned, it is reasonable to suspect that the type of faking on noncognitive measures can be either (a) negatively related to the predictor (e.g., applicants are motivated to impression manage because they seek to cover up deficiencies on the constructs being measured), (b) unrelated to the predictor (e.g., impression management has nothing to do with a high or low standing on the construct being measured), or (c) positively related to the predictor (e.g., applicants who successfully engage in impression management may also tend to do well on a measure that purports to predict customer service because a customer-service orientation requires managing the impressions of others). These cases suggested negative, zero, and positive predictor–faking correlations in the simulations: r_{xs} = −.20, .00, .20, and .40, where s represents the measure of faking.

Faking–criterion correlation. The same line of reasoning led to simulation values for the faking–criterion correlation: r_{sy} = −.10, .00, .10, and .30. Correlations are slightly lower to reflect the likelihood that faking may have more distant or indirect effects on the criterion because the criterion may be more cognitive or complex in nature than the noncognitive predictor and because the nature of faking is, in many ways, likely to be more predictor-specific than criterion-specific.

Selection ratio. Selection ratios ranged from selective to moderate: 10.0%, 20.0%, and 50.0%. Clearly more than three values for the selection ratio could have been chosen, but choosing more levels of any factor leads to a geometric increase in the number of simulation conditions needed. We hoped that selecting across a range of possible selection ratios as we did would result in a pattern of findings that would serve to indicate the values one might expect at other selection ratios.

Proportion removed for suspected faking. In the simulations, we set the proportion of applicants removed because they were suspected of faking at either 0.0%, 5.0%, 15.0%, or 30.0% of top scorers on s, the measure of faking. It should be pointed out that when no identification of fakers is attempted or accomplished, mean performance on the criterion is a function of the selection ratio and test validity and can be estimated with tables attributed to Brogden (Brown & Ghiselli, 1953). These expected mean-performance levels constitute a baseline against which efforts to assess the effect of removing suspected fakers is evaluated.

Data Generation Procedure

Each simulation reflected a unique combination of values for sample size, r_{xy}, r_{xs}, r_{sy}, selection ratio, and percentage identified as faking. We

generated individual scores for *x*, *y*, and *s* to be multivariate normal (with means of 0 and standard deviations of 1) to reproduce these correlations for each cell in the design within the bounds of sampling error. Then, the top scorers on *s* who fell under the percentage to be identified as faking were flagged. In other words, when the percentage of scorers to be identified as faking was higher, the cutoff score on *s* was lower (i.e., the cutoff was the value suggested by the standard normal curve plus any sampling error).

The fact that we have varied levels of the various factors in our simulation independently of each other does not imply any particular causal model of the relationship between these factors. What we have done is to vary levels of each of the simulation factors across values that are representative of those observed in the empirical literature; various combinations of these factors are not equally likely to occur in selection situations. The fact that our simulations allow the correlations between the predictor and the faking measure to vary, without influencing the value of the correlation between the predictor and the criterion, does not mean that we are assuming the independence of these correlations. For example, it is likely that the motivation to distort one's responses to a noncognitive measure influences those responses, which in turn tells us something different about subsequent performance than it would if no such motivation or attempts to fake existed.

Selection Procedure

Next, for each simulation condition, selection occurred in a typical top-down manner on the rank-ordered predictor scores until the specified selection ratio was satisfied. Applicants were removed if they should have been selected with the top-down rank order but were also identified as fakers. They were replaced by those applicants who were the next-highest scorers on the predictor yet were not identified as fakers. This procedure serves the simultaneous goals of removing the influence of faking while maintaining the highest mean performance possible, given the desired selection ratio. Simulations were written in the SAS/IML programming language (1999).

Results

Mean performance was regressed on five design factors (all factors except sample size) and their two- and three-way interactions. As can be seen in Table 2, these main effects and interactions accounted for over 99.0% of the variance in mean performance; the three-way effects accounted for only 0.2% of the variance so we assumed that any higher order interactions would be relatively inconsequential. Statistical significance in these simulations is not meaningful because the number of replications per conditions can be chosen arbitrarily such that sampling error is as small as desired (we chose 1,000 replications per condition). Because design factors are independent of one another, squared standardized regression weights for the main effects represent the variance accounted for by each factor, which Table 2 shows. Another index of the importance of each factor is the mean performance of those selected for each level of the main effects of our design, as Table 3 shows. It should be noted that we could change the impact of these factors had we chosen other levels of the factors considered. However, we chose representative levels of all factors across realistic ranges found in existing literature and practice.

Values in Tables 2 and 3 indicate that the validity of the predictor has the greatest impact, accounting for nearly 59.0% of the total variance in criterion performance. The mean performance across the three levels of selection ratio ranges from .103 to .470. This finding is, of course, consistent with research on the utility of valid selection procedures conducted many decades ago (Brown & Ghiselli, 1953).

Table 2
Hierarchical Linear Regression: Predicting Mean Performance From Simulation Conditions

Factor	B	R^2	ΔR^2
Step 1			
r_{xs}	−.016		
r_{sy}	−.188*		
r_{xy}	.771*		
SR	−.480*		
prf	−.179*	.892*	
Step 2			
$r_{xs} \times r_{xy}$	.008		
$r_{xs} \times SR$	.011*		
$r_{xs} \times prf$	−.019*		
$r_{xs} \times r_{sy}$	−.066*		
$r_{sy} \times r_{xy}$	−.002		
$r_{sy} \times SR$	.027*		
$r_{sy} \times prf$	−.243*		
$SR \times prf$	.060*		
$SR \times r_{xy}$	−.686*		
$prf \times r_{xy}$	−.128*	.991*	.099*
Step 3			
$r_{xs} \times r_{sy} \times r_{xy}$	.002		
$r_{xs} \times r_{xy} \times SR$	−.009		
$r_{xs} \times r_{xy} \times prf$	.013*		
$r_{xy} \times SR \times prf$	−.082*		
$r_{xs} \times r_{sy} \times SR$	.051*		
$r_{xs} \times r_{sy} \times prf$	−.061*		
$r_{xs} \times SR \times prf$	.017*		
$r_{sy} \times SR \times prf$	.020*	.993*	.002*

Note. Regression weights are standardized and are the values at that particular hierarchical step. Mean performance is averaged across 1,000 simulations per condition. r_{xs} = predictor–faking correlation; r_{sy} = faking–criterion correlation; r_{xy} = criterion-related validity; *SR* = selection ratio; *prf* = proportion removed for suspected faking.
* $p < .01$.

Also well established is the impact of the selection ratio on utility estimates. Results in Tables 2 and 3 show that the selection ratio accounts for approximately 23.0% of the variance and that average performance ranges from .140 to .372 across the three levels of validity considered. Results for selection ratio and validity are as expected and, as mentioned above, consistent with a substantial literature on the utility of selection tests (Boudreau, 1991).

The impact of the correlations of a faking measure with a predictor, a faking measure with the criterion, and the proportion of examinees removed because they were thought to be faking has not been previously assessed systematically, even though the body of literature reviewed in our introduction is based on the premise that corrections for faking can make a substantial difference in performance outcomes and that practitioners continue to use such corrections in making decisions about applicants (Goffin & Christiansen, 2003). Results of the regression analyses indicate that the correlation of a faking measure with the criterion (Hypothesis 4) and the proportion identified as faking (Hypothesis 5) each account for slightly more than 3.0% of the variance in average criterion performance. Examination of cell means shows that there is a .092 standard deviation difference in performance (i.e., .304 –

Table 3
Main Effects of Simulation Conditions on Mean Performance

Parameter	Performance (*M*)
r_{xy}	
.10	.103
.20	.226
.40	.470
r_{xs}	
−.20	.272
.00	.267
.20	.263
.40	.263
r_{sy}	
−.10	.310
.00	.285
.10	.260
.30	.210
SR	
.10	.372
.20	.288
.50	.140
prf	
.00	.304
.05	.290
.15	.260
.30	.212

Note. Main effects are collapsed across the values of all other parameters in the simulations. There are 2,304 simulation conditions total in a fully crossed design of the values listed above. Each simulation condition was replicated 1,000 times, taking the average mean performance across replications within each condition. r_{xy} = criterion-related validity (768 conditions each); r_{xs} = predictor–faking correlation (576 conditions each); r_{sy} = faking–criterion correlation (576 conditions each); *SR* = selection ratio (768 conditions each); *prf* = proportion removed for suspected faking (576 conditions each).

.212) when no applicants are removed for faking as opposed to the case when 30.0% are removed. Mean performance differences across the range of our manipulation of the correlation between a faking measure and the criterion are .100 (i.e., .310 – .210). As expected, when the correlation between the faking measure and the outcome measure is negative, mean performance does improve, though such improvement for a change in correlation from .00 to −.10 was only .03 *SD*s. These two effects, then, are relatively small. When the correlation between the faking measure and the criterion is .30 as opposed to zero, removal of applicants decreases expected performance by .075 (i.e., .285 – .210).

The impact of the correlation between the predictor and the faking measure on mean performance was very small. The difference between the largest and smallest mean performance was only .01 *SD*s. Thus the impact from removing those thought to be faking is almost nonexistent, which might be expected given that the impact on criterion performance is indirect, and the validity of observed personality measures, as modeled here, tends to be low to moderate.

In addition to the main effects, three of the three-way interactions were relatively large and statistically significant even though they accounted for minimal variance in expected mean performance (see Table 2). Because these three interactions also involve the variables in the significant two-way interactions, we present tables that involve the three-way interactions only. The largest of these interactions involved the selection ratio, the proportion of examinees removed for faking, and the validity of the predictor. As Table 4 shows, the proportion removed because of suspected faking has the largest effect on mean performance with a high selection ratio, a high proportion faking, and high test validity, as compared with the condition when none are removed on the basis of the faking measure and selection ratio and validity are high. The difference in expected performance is .165 (.317 – .152). A similar comparison for a low-selection ratio yields a mean difference of .104 (.693 – .589). The higher the validity and the higher the selection ratio and proportion removed for faking, the more likely it is that individuals whose expected performance levels are high will be eliminated with the faking measure. Expected performance of the replacements will not be as high. As was evident from the examination of the main effects, validity and selection ratio have the greatest impact on expected performance.

The second three-way interaction involved the selection ratio, the faking–predictor correlation, and the faking–criterion correlation. Specifically, Table 5 shows that mean performance was greatest when the selection ratio was low, the faking–predictor correlation was high, and the faking–criterion correlation was negative. In the case of the faking–criterion correlation, the impact was highest when the selection ratio was .10 and the predictor–faking correlation was .40. In this case, the mean performance difference was .171 (i.e., .442 – .271) comparing conditions in which r_{sy} was −.10 versus .30. Use of a faking measure correlated with the criterion has a direct impact on the levels of criterion performance, and this is reflected in the greater impact of the faking–criterion correlation relative to the faking–predictor correlation.

The interaction of the proportion removed for faking with the faking–predictor and faking–criterion correlations (see Table 6) seemed to indicate that the impact of the faking–criterion correlation was greatest when the proportion faking was large and the

Table 4
Mean Performance by Selection Ratio, Proportion Removed for Suspected Faking, and Criterion-Related Validity

prf and r_{xy}	*SR* = .10 (*M*)	*SR* = .20 (*M*)	*SR* = .50 (*M*)
.00			
.10	.173	.139	.079
.20	.347	.278	.159
.40	.693	.554	.317
.05			
.10	.160	.126	.068
.20	.333	.264	.142
.40	.676	.538	.295
.15			
.10	.143	.107	.045
.20	.308	.238	.112
.40	.644	.501	.245
.30			
.10	.114	.078	.009
.20	.273	.197	.057
.40	.589	.436	.152

Note. Mean performance is in the *z*-score metric. Each result reflects average mean performance across 64 simulation conditions with 1,000 replications per condition. *SR* = selection ratio; *prf* = proportion removed for suspected faking; r_{xy} = criterion-related validity.

Table 5
Mean Performance by Selection Ratio, Predictor–Faking Correlation and Faking–Criterion Correlation

r_{xs} and r_{sy}	$SR = .10$ (M)	$SR = .20$ (M)	$SR = .50$ (M)
−.20			
−.10	.405	.320	.175
.00	.388	.305	.157
.10	.374	.289	.138
.30	.348	.261	.103
.00			
−.10	.409	.326	.178
.00	.390	.305	.156
.10	.366	.282	.134
.30	.321	.240	.091
.20			
−.10	.422	.334	.181
.00	.390	.307	.157
.10	.359	.278	.132
.30	.299	.221	.084
.40			
−.10	.442	.350	.186
.00	.399	.312	.159
.10	.354	.277	.133
.30	.271	.203	.077

Note. Mean performance is in the z-score metric. Results reflect average mean performance across 48 simulation conditions with 1,000 replications per condition. SR = selection ratio; r_{xs} = predictor–faking correlation; r_{sy} = faking–criterion correlation.

predictor–faking correlation was negative as opposed to high positive. All these factors, however, produced relatively small differences in mean criterion performance; the largest difference across all conditions in Table 6 was .29 *SD*s.

Across all three interactions, effect size differences based on the correlation between the faking index and either the criterion or the predictor were relatively small; the main impact in all these interactions was that due to the selection ratio and test validity. Because these other interaction effects are relatively small, we do not examine or report the pattern of these interactions.

Discussion

For the past several decades, researchers have been concerned about the possibility that respondents to personality, biodata, and other noncognitive measures can and do fake their responses in a manner that is most likely to result in their selection for some desired outcome, such as a job or admission to an educational program. One attempt to deal with this threat to the utility of these measures has been to use scores from measures of faking (e.g., measures of lying or social desirability) to correct applicants' scores for faking or to remove from consideration those applicants who are thought to be faking. Results reported in this article show that, given typical levels of faking–predictor and faking–criterion correlations, using measures of faking to remove applicants from consideration for selection—even up to 30.0% of them—has very little impact on the mean performance in typical selection situations (i.e., typical selection ratios, validity, and proportion of fakers to be replaced). Our simulation assumed that fakers in the applicant pool would be rejected outright; if predictor scores were only corrected in some way without necessarily removing applicants, the impact of the use of a faking measure would be even less. The simulation also involved the implicit assumption that the faking measure used was perfectly accurate; this is highly improbable. The degree to which such measures are inaccurate would further diminish the impact of their use. However, there may be situations in which the standard deviation of the worth of some outcome is very large. In those instances, even the relatively small effects (<.10) reported here may be critical to organizational functioning.

It is important to note that our conclusion that faking measures cannot make much difference, given typical selection scenarios, applies only when considering mean performance. At the individual level, however, individuals can and will be affected significantly if the treatment for which they are being considered is a desirable one. Because none of the faking measures are a perfectly valid measure of a person's tendency to fake or produce socially desirable responses on noncognitive predictor measures, faking corrections may also affect unfairly those persons identified as fakers whose responses are not inflated by a motivation to fake. It is possible, of course, that some applicants will react positively to organizational efforts to remove fakers even though the measures of faking are imperfect. This would be one reason to continue using the corrections. Because they have little impact on organizational outcomes, from an organizational perspective, this could be a reasonable strategy. That said, we know of no research that addresses applicant reactions to corrections to their test scores for faking.

The fact that the use of measures to remove suspected fakers may have little impact on expected performance or validity should

Table 6
Mean Performance by Proportion Removed for Suspected Faking, Predictor–Faking Correlation, and Faking–Criterion Correlation

r_{xs} and r_{sy}	$prf = .00$ (M)	$prf = .05$ (M)	$prf = .15$ (M)	$prf = .30$ (M)
−.20				
−.10	.304	.303	.302	.291
.00	.304	.296	.281	.251
.10	.303	.291	.263	.211
.30	.304	.281	.227	.137
.00				
−.10	.305	.307	.307	.297
.00	.304	.297	.283	.250
.10	.303	.287	.254	.198
.30	.304	.266	.201	.100
.20				
−.10	.304	.314	.318	.313
.00	.303	.299	.283	.252
.10	.304	.283	.246	.192
.30	.304	.254	.178	.071
.40				
−.10	.306	.325	.337	.335
.00	.307	.302	.290	.260
.10	.305	.282	.246	.188
.30	.305	.238	.151	.041

Note. Mean performance is in the z-score metric. Results reflect average mean performance across 36 simulation conditions with 1,000 replications per condition. prf = proportion removed for suspected faking; r_{xs} = predictor–faking correlation; r_{sy} = faking–criterion correlation.

not obviate concern for the impact of faking on the construct validity of the predictors we use. When scores on the predictors are a function of applicant attempts to fake responses, the psychological interpretation of these scores may be very different than in instances when no attempts to fake occur. In fact, one desirable outcome of our study would be that researchers redirect their attention from concerns about mechanically correcting test scores for faking to concerns about the constructs underlying measures of faking and predictors, given settings that vary in their influence on applicant motivation to fake. Empirical relationships between predictor and criterion, such as those shown in these simulations, reflect just one of many important types of evidence concerning construct validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Our results do affirm what has been found in many other studies of the utility of selection procedures; that is, the most important determinants of mean performance are test validity and the selection ratio. Over the conditions we studied, validity accounted for 59.0% of the variance in mean performance and selection ratio for 23.0% of the variance. Our results are also consistent with empirical research cited in the introduction that indicates that uncorrected and faking-corrected criterion-related validities differ very little (Christiansen et al., 1994; Hough, 1998). In fact, the results of the simulations demonstrate that in normally encountered situations, such corrections cannot possibly make much difference.

We have at various points mentioned the utility of these faking corrections that lower expected performance only slightly. From a utility perspective, we have not considered the cost of collecting information that would identify suspected fakers. If an inexpensive paper-and-pencil measure of social desirability like those mentioned in the introduction is used, the cost would be minimal. If, however, an extensive background check and/or polygraph measure is used, the cost of collecting this information would be substantial. As mentioned above, it is also likely that any available faking index has far less than perfect accuracy. Checking accuracy is difficult, if not impossible, as employers usually have no way of verifying the truth of an applicant's claims in many employment situations. In those instances when an accuracy measure is available or attempts to construct one are made, substantial inaccuracies are found (e.g., for the case of the polygraph in field settings, see Pollina, Dollins, Senter, Krapohl, & Ryan, 2004). This inaccuracy would further diminish any positive effect that might result from use of a faking index. A final "utility" question that cannot be easily quantified is the potential for negative applicant reactions to procedures that some may interpret as evidence that the employing organization does not trust them and must take special steps to verify their honesty.

Having made these general statements, it is the case that when a faking measure has a positive correlation with the criterion, removing suspected fakers does result in a small decrease (i.e., .10 *SD*s) in mean performance across all conditions. Performance increases slightly when the faking–criterion correlation is negative. Mean performance gets increasingly lower as this correlation becomes positive and larger. The correlation of the faking measure with the predictor has almost no impact on mean performance. Examining cell means across the three-way interactions suggests the impact of the faking–predictor and faking–criterion correlations varies most as a function of validity and selection ratio but not as a function of the correlations themselves.

The major conclusion of this study is that researchers who use some form of correction for faking with the goal of increasing the validity of a test or increasing the mean performance resulting from use of a test should not expect much difference given selection situations with realistic parameters such as those we used in our simulations. In fact, such corrections usually lower expected performance by small amounts. At the same time, as stated previously, even though the results from corrections do not change much at the aggregate level, corrections can still have significant impact on selection decisions for individual applicants.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. *Personnel Psychology, 44,* 1–26.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 83,* 261–272.

Boudreau, J. W. (1991). Utility analysis for decisions in human resource management. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 621–746). Palo Alto, CA: Consulting Psychologists Press.

Brown, C. W., & Ghiselli, E. E. (1953). Percent increase in proficiency resulting from use of selective devices. *Journal of Applied Psychology, 37,* 341–345.

Cattell, R. B., Cattell, A. K., & Cattell, H. E. (1993). *Sixteen Factor Questionnaire* (5th ed.). Champaign, IL: Institute for Personality and Ability Testing.

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

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

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

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

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

Hough, L. M., Eaton, N. R., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook* (pp. 391–420). Palo Alto, CA: Consulting Psychologists Press.

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

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability

and faking on personality and integrity assessment for personnel selection. *Human Performance, 11,* 245–270.

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

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

Pollina, D. A., Dollins, A. B., Senter, S. A., Krapohl, D. J., & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of Applied Psychology, 89,* 1099–1105.

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

SAS Institute. (1999). *SAS/IML User's Guide, Version 8.* Cary, NC: Author.

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12,* 250–264.

Zickar, M. J., Rosse, J. G., Levin, R. A., & Hulin, C. L. (1996, April). *Modeling the effects of faking on personality tests.* Paper presented at the 11th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Received September 15, 2004
Revision received March 22, 2005
Accepted April 29, 2005 ■