Psychology Science, Volume 48, 2006 (3), p. 226-246

Relationships between faking, validity, and decision criteria in personnel selection

BERND MARCUS[1]

Abstract

There has been some debate in recent years as to whether faking on personality tests, while apparently not affecting criterion-related validity, still has a detrimental effect on the accuracy of hiring decisions. The present paper is set out to contribute to a clarification of this issue conceptually and empirically. In the conceptual part, statistical parameters of test scores obtained in selection settings that may affect validity and hiring decisions are disentangled. A data set of job incumbents who took an integrity test in a research setting is then used to demonstrate the effects of simulated faking scores with systematically manipulated distributional properties. Results show that, while hiring decisions are more sensitive to manipulations than validity, changes on both decisions and validity depend upon the same parameters, most importantly on variance in faking. Unlike the overlap between decisions based on faked and non-faked scores, the accuracy of these decisions was not more sensitive to faking than validity, regardless of selection ratio. Results are discussed in light of findings on criterion-related validity of personality tests in real-world applicant settings.

Key words: applicant faking, personnel decision accuracy, criterion-related validity, personality testing

[1] Bernd Marcus, Department of Psychology and Bachelor of Management and Organizational Studies Program, University of Western Ontario.
I wish to thank Uwe Wagner for his help with sample recruitment and data collection for the empirical part of this paper and Rick Goffin for many insightful comments on an earlier version of this paper.
Correspondence regarding this article should be addressed to Bernd Marcus, Department of Psychology and Bachelor of Administrative and Commercial Studies Program, University of Western Ontario, London, Ontario, N6A 5C2, Canada; Email: bmarcus3@uwo.ca.

Until recently, it was common sense that psychological assessment tools susceptible to deliberate changes of responses shall not be used in situations where test takers are motivated to manipulate their scores. This conventional wisdom was seriously challenged by meta-analytic findings showing that correcting scores on personality tests for apparent response distortion using “validity scales” of any sort had no effect on actual criterion-related validity in applicant settings (Moorman & Podsakoff, 1992; Ones Viswesvaran & Reiss, 1996; Viswesvaran, Ones & Hough, 2001). However, this failure to improve validity by means of conventional scale score correction may be due to either the invalidity of validity scales for detecting faking or to the insensitivity of validity coefficients to linear statistical corrections with the moderate sizes of correlations found in that domain, or both. Hence, these findings are only weak indication for the absence of – presumably negative – faking effects on criterion-related validity. More direct evidence comes from findings in other metaanalyses that fakable selection instruments are as predictive of job-related criteria if used in applicant settings than they are in non-applicant settings (Hough, 1998 report a detriment of $r_{diff} = .07$ in predictive applicant vs. concurrent incumbent settings; but this was exactly attributable to design alone in Ones, Viswesvaran & Schmidt’s, 1993, larger meta-analysis on integrity tests). Taken together, the accumulated empirical findings so far do not point to a systematic dilution of criterion-related validity of fakable measures in real applicant settings, whereas a substantial drop in validity was consistently reported with instructed faking in the laboratory (Hough, 1998). This seems to imply that processes of faking or self-presentation as well as their outcomes may be considerably more complex in the real world than those of simulated faking studies[2].

While not disputing that unequivocal evidence of negative effects of faking on validity is lacking, several researchers (e.g., Donovan, Dwight & Hurtz, 2003; Rosse, Stecher, Levin & Miller, 1998) expressed their ongoing concerns about the use of personality tests for making hiring decisions. The main argument put forward by these and other authors is the possibility that, especially if selection ratios are low, fakers have a disproportionately good chance of being hired even if overall criterion-related validity remains unchanged. Several studies using various methods and designs appear to confirm the sensitivity of hiring decisions to faking at the high end of the distribution (e.g., Mueller-Hanson, Heggestad & Thornton, 2003; Rosse et al., 1998).

The general objective of the present paper is to examine the validity of this argument from various perspectives. First, I point to some limitations in previous research apparently supporting this argument, using the two studies just cited as exemplars, as both appeared in the leading journal of I/O psychology. Second, I attempt to clarify the relations between (a) faking and hiring decisions, and (b) faking, validity, and decision accuracy conceptually, using hypothetical statistical properties of the variables and concepts involved. Finally, I use

---

[2] A discussion of the full complexity of faking processes in personnel selection is beyond the scope of the present paper (see, e.g., Bozeman & Kacmar, 1997; McFarland & Ryan, 2000, for process models of faking). As rightly pointed out by one anonymous reviewer, and discussed in some detail by the present author elsewhere (Marcus, 2005), a lack of faking effects on criterion-related validity does not imply that faking does not exist or that it has no effect on validity. The overall null findings may instead be the outcome of coexisting positive and negative effects that weigh each other out in aggregates. In the present paper, I conceptualize “faking” purposefully in a simplified manner as some unspecified entity, to clarify how this hypothetical variable affects different operationalizations of decision accuracy in the absence of clear-cut effects on criterion-related validity.

simulated faking scores with systematically varied properties to examine empirically, based on an actual data set, the previously unexplored differences of faking effects on decision accuracy conceptualized either as (a) changes in hiring decisions, or (b) mean scores of hired applicants on job-related criteria.

Exemplary previous research

While the argument of impaired decision accuracy has a striking appeal, there may be several reasons to be skeptical about its soundness. First, showing that hiring decisions change in the presence of faking is one thing, but showing that the changes actually lead to less accurate decisions (or more decision errors) may be a quite different issue. In the absence of an accepted standard of who "deserves" being hired and who doesn't, it is difficult to assess to what extent decision changes can be treated as decision errors. In personnel selection, this standard is usually given by measures of job performance or other job-related criteria (Society for Industrial and Organizational Psychology, 2003). Rosse et al. (1998) did not measure performance or any other job-related criterion, but used an impression management (IM) scale to estimate the amount of faking in applicants' scores on personality scales. If the personality scales are correlated with the IM scale, impression managers identified by that means will naturally be found in high proportions at the high end of personality score distributions, and correcting for IM, of course, leads to changes in the rank order. Given the questionable ability of IM scales to detect low performers, including the scale used by Rosse et al. (see Barrick & Mount, 1996), this is a weak surrogate for actual performance. Accordingly, Rosse et al. based their decision accuracy argument not on measured performance but on the *assumption* that extreme fakers (as measured by IM scores) show a lack of integrity and "would also engage in other forms of dishonesty" (Rosse et al., 1998, p.642). The finding that fakable (Alliger & Dwight, 2000) scores on integrity tests predict dishonest job behavior in applicants (Ones et al., 1993) does not seem to support this speculation, and, as Rosse et al. acknowledged, many scholars of self-presentation and personnel selection disagree with the assumption that IM indicates dishonesty (e.g., Hogan & Hogan, 1998; Ones & Viswesvaran, 1998; Schlenker, Britt & Pennington, 1996).

Whereas Rosse et al.'s study lacked a measure of actual performance, Mueller-Hanson et al. (2003) did assess the performance of their participants on a series of tedious tasks. In line with Rosse et al., Mueller-Hanson and colleagues found that the proportion of fakers (potentially) hired increases as the selection ratio decreases, but they found in addition that hired participants in the induced faking condition had lower performance than those in a control group if the selection ratio was low. So, does this study provide the previously lacking piece of evidence in favor of the decision accuracy argument? Not exactly, I think. While Rosse et al. had compared job applicants with incumbents, Mueller-Hanson et al. used an experimental design set out to simulate an applicant setting. Mueller-Hanson et al. enhanced the realism of their simulated applicant condition by offering participants incentives for successfully passing the personality test and providing them with some additional information. Although this is certainly more realistic than the typical directed faking study, the design lacks at least one potentially decisive feature of actual applicant settings. Real applicants do not only have to expect a gain from passing the test, they also have to consider the long-term costs of accepting a job offer, which, in their own view, would be a misfit to their personality (see

Marcus, 2003, 2005, for more details on this argument). Real applicant settings are like gambles where the prize is precious but stakes are high, facing applicants with a mixture of incentives and deterrents for faking. Evidence reported above suggests that criterion-related validity is not affected under these conditions. Typical directed faking experiments present their participants with a situation where they have absolutely nothing to gain and nothing to lose through faking. Hough (1998) reports evidence from multiple studies that found an almost complete breakdown of validity under these conditions. Mueller-Hanson et al.'s design represents a compromise between these extremes, like a gamble with a small jackpot but no stakes at all. One would expect criterion-related validity to be moderately affected under these conditions, which is in line with the difference in validity between the incentive (r = .05) and control group (r = .17) reported by Mueller-Hanson et al. Thus, this experiment appears to represent a progress over earlier laboratory experiments in terms of realism, but not such a substantial one to fully justify the authors' claim that they combined the control of an experiment (which they did) with the realism of an applicant setting. In particular, this study could not support the assumption that we find lowered decision accuracy in the *absence* of an effect on criterion-related validity.

My second major concern is with the bases for this assumption reported by Rosse et al. (1998) and later cited by several other authors. Rosse et al. wrote "Drasgow and Kang (1984) have shown that correlation coefficients are extremely robust estimators of linear associations between variables but that this robustness comes at the cost of sensitivity to changes in rank order in particular ranges of a bivariate distribution (such as changes among top-scoring applicants). […] Although the observed validity of the test may not change for the whole sample, its validity for the applicants who are at the top end of the predictor distribution […] may approach zero if response distortion occurs primarily among those who receive the highest scores [cites of unpublished studies follow]." (p. 636). Upon inspection, however, Drasgow and Kang (1984) did actually not address the question of changes in rank order in particular ranges of a distribution. These authors addressed a slightly different question, namely the extent to which non-equivalent measurement of a test across defined groups (e.g., the two sexes) translates into differential validity across the same groups. This is equivalent to the question whether group membership moderates validity. In the present case, one might ask, for example, whether personality tests are more valid in persons answering honestly compared with fakers, however fakers might be defined. Unlike this moderator effect, Rosse at al.'s interpretation of Drasgow and Kang refers to the extent to which changes in the high end of *one* bivariate distribution suppress validity only in that particular range. This question is related but not identical to the moderator issue (cf. Borkenau & Ostendorf, 1992). Moreover, validity coefficients for applicant samples are usually based on hired applicants only (the very group for which validity is presumed to approach zero), as criterion data is typically not available for rejected applicants. As the assumption of the coexistence of unaffected validity and affected decision accuracy is essential for the entire discussion given empirical results on criterion-related validity in applicants, there appears to be some need for clarification on this issue.

A third concern to be mentioned here is with the inconsistency between the empirical or logical status of some issues and the way they are interpreted. For example, regardless of differences in substantive interpretations of faking, almost all scholars of the field appear to agree that a dichotomous distinction between "fakers" and "non-fakers" oversimplifies the continuous distribution of this variable (but see Zickar, Gibby & Robie, 2004, for a more complex grouping approach, which, however, does not preclude that transition between groups is smooth). However, a dichotomous concept of faking is still used to discuss the implications of faking on hiring decisions (e.g., Donovan et al., 2003). There is also no disagreement that a mere mean shift of a distribution of faking scores does not matter, whereas the variance of faking potentially does, but there is still much more research on mean effects than on the variance of faking. As a final example, Douglas, McDaniel and Snell (1996; as cited in McFarland & Ryan, 2000) rightly pointed out that the possible amount of faking depends on the true score such that a high true score limits any possible enhancement. Although this notion is true by definition, it can be essentially misleading if not properly interpreted. As discussed in more detail in the sections that follow, all of these issues have a number of potentially important, but often overlooked, implications for the impact of faking on validity and hiring decisions.

To summarize, I mentioned three major concerns with the decision accuracy argument, (1) lack of convincing empirical evidence for impaired decision accuracy in the presence of intact validity, (2) open questions regarding the likelihood of such a co-occurrence, and (3) a range of technical issues that are relevant but not always appropriately considered. To avoid misunderstandings, the word "concern" is used here to denote points I think are in need of clarification, not to imply that proponents of the decision accuracy argument are wrong with their assumptions. These concerns are closely related to each other. An inspection of the way statistical parameters of the distribution of faking scores can be expected to affect hiring decisions and validity may be used to develop hypotheses on simultaneous effects on these outcomes, which may then be tested empirically. Thus, the present paper is set out to address the concerns enumerated above in reversed order. First, I address some of the mechanics of faking effects on hiring decisions and validity conceptually. This discussion is presented in a non-technical fashion, although it addresses essentially technical questions. That is, the variable "faking" is discussed with respect to particular statistical properties, but regardless of any substantive definition. In the subsequent section, an empirical data set is used to systematically vary these distributional properties and examine the effects on outcomes with some real data. Finally, I discuss implications of the present research in light of the original arguments and empirical evidence available from other sources.

Faking, true scores, and hiring decisions

Leaving unsystematic error components and systematic errors other than faking aside, the actual score of an applicant on a fakable measure like a personality test can be conceptualized as composed of the true score on the trait measured and an additional "faking score". Faking itself may consist of several facets (cf. footnote 2), but all these facets occur at the same time under conditions motivating faking. Theoretically, the faking score can take on negative values (a phenomenon known as faking bad), but it seems reasonable to assume that this is unlikely in the selection context. The well-known graphical representation of Taylor and Russell's (1939) classic concept of selection utility can be used to illustrate the effects of the faking score on hiring decisions and its relationship to the true score. Figure 1 depicts an adaptation of this concept to the present issue. In this figure, the left-hand (solid) ellipse represents the bivariate distribution of true scores on the predictor and job performance, and the right-hand (dashed) ellipse represents the respective distribution of applicant scores.

Consistent with the meta-analytic literature mentioned earlier, the applicant distribution shows an elevated mean on the predictor but the same form as the true score distribution, indicating unchanged validity. The regression line of performance regressed on true scores is also provided. Further, the three vertical lines represent, from left to right, a hypothetical cutoff for the case that no faking at all occurs, the actual cutoff or selection ratio in the presence of faking, and the maximum score on the predictor. The first of these lines cuts the distribution of true scores at the same proportion as the selection ratio cuts the distribution of applicant scores. This is based on the assumption that the selection ratio depends on the number of positions available, not on a cutoff provided in the test manual. This condition may not apply to all situations, but changing it has no effect on the general mechanics involved.

The horizontal arrows in Figure 1 may require some explanation. We can define a critical faking score (FS_{crit}) as the amount of faking needed to pass the test under applicant conditions. This value is given by the difference between the applicant cutoff and the individual true score, or $FS_{crit} = C_{AS} - TS$. The critical faking score can be seen as an index of the difficulty of successful faking, or as an indicator of the probability of being hired at a given level

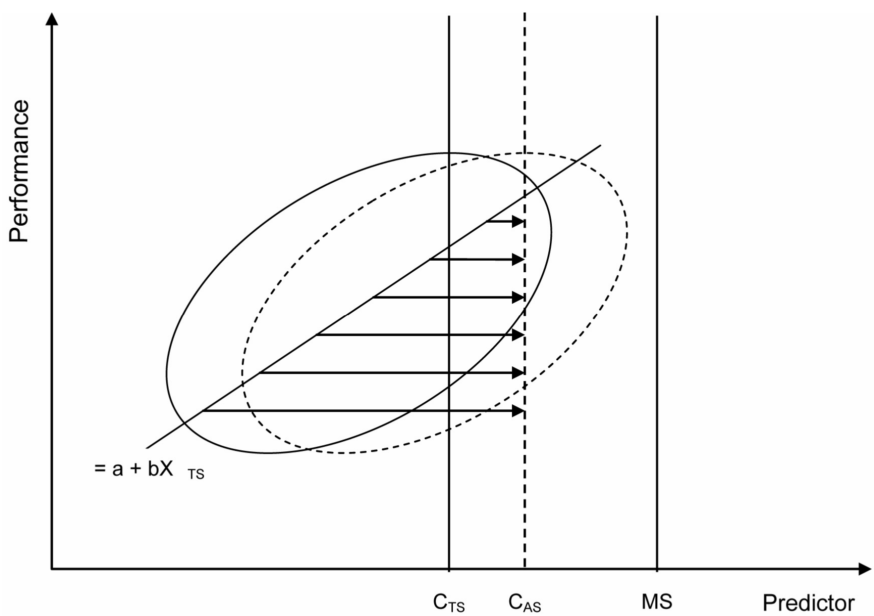

Figure 1:
Effects of True Score Distribution on Hypothetical Hiring Decisions under the Condition of Faking in the Absence of Lowered Validity. (Solid ellipse represents predictor true score distribution; dashed ellipse represents distribution of applicant scores. Sloped line is the regression line for performance regressed on true scores. Horizontal arrows represent critical faking scores (see text). C_{TS} = hypothetical predictor cutoff for true scores; C_{AS} = predictor cutoff for applicant scores (selection ratio); MS = maximum predictor score.)

of the true score if the faking score is unknown. The arrows in Figure 1 depict this value at varying levels of the true score and the corresponding levels of performance predicted from true scores.

There are several things noteworthy about the graphical analysis in Figure 1. Obviously, FS_{crit} is determined by the true score, which implies that the probability of being hired is a positive function of true scores regardless of the selection ratio. However, at decreasing levels of the selection ratio, the impact of the faking component in the applicant scores relative to the true score component on selection decisions becomes stronger. This implies that decisions in highly competitive selection processes are in fact more sensitive to faking than less competitive ones. As long as true scores are positively correlated with performance, FS_{crit} is also negatively related to performance, implying that it is more difficult for low than for high performers to fake the test. If predictor and criterion scores are standardized, we can estimate true scores from performance in the same way as vice versa such that the expected value of TS = X becomes $r_{xy}Y$. We can then write $F\hat{S}_{crit} = C_{AS} - r_{xy}Y$ to estimate critical faking scores from performance. These relationships between true score, performance and hiring probability suggest that the common concern that extreme fakers are likely to be found among hired applicants in high proportions is not nearly as stringent as may appear at first sight.

Also obvious from Figure 1 is that hiring decisions are independent of the maximum predictor score as long as the cutoff does not exceed the maximum score – which, of course, is impossible. It seems extremely unlikely that the cutoff even approaches the maximum score. As an author of several personality tests and user of many others, I have access to thousands of raw scores obtained under varying conditions. In these data sets, not even one participant reached the maximum score in tests containing a sizable number of items. As long as the cutoff score is below the maximum score, the maximum score has no effect at all on hiring beyond the critical faking score (which differs from the maximum score only by a negative constant). The possible amount of faking, also known as opportunity to fake, therefore is an essentially meaningless concept with regard to hiring decisions. Falsely interpreted, the concept of opportunity to fake could be taken as indication that true scores are negatively related to chances of being hired (because low true scorers have more opportunity to fake), which is the reverse of the true relationship.

Finally, the present analysis suggests that under conditions implied by proponents of the decision accuracy argument (low selection ratio plus a sizable mean effect of faking on predictor scores), a large proportion of high true scorers (i.e. those who would have been hired if everyone would take the test honestly) are required to fake to some extent to be hired. In Figure 1, this applies to everyone whose true score is placed between the lines denoted R_{TS} and R_{AS}. Hence, any dichotomous distinction between fakers and non-fakers inevitably confuses true positives with false positives within the apparently homogeneous group of "fakers" as long as there is any mean effect of faking. More generally, conceptualizing faking as a dummy variable defines away all variance within the faking group, which, as I will try to show next, is precisely the parameter we should be most interested in.

The impact of faking on validity and decision accuracy

Aside from any content-related definition of faking (recall that the present analysis does not apply such a definition), its likely impact on validity and hiring decisions can in part be deduced from hypothetical distributional properties. As with any variable, the univariate distribution of "faking" can be described by the major statistical parameters mean and variance, and additional parameters like skewness. In the bivariate or multivariate case, the covariance with other variables is also of interest. For the present purpose, those other variables would be the true score on the predictor trait and the performance criterion.

In the simplest case, faking may be assumed to be normally distributed and uncorrelated with both the predictor and the criterion at the construct level. That is, one can think of faking as a normally distributed random variable added to the true score. Under these conditions, the impact of faking on both validity and decision accuracy can be expected to fully depend on its variance. A random variable would introduce random variance or measurement error into the test score. The larger this random variance, the larger will be its attenuating effect on correlations with outside variables (i.e. validity). At the same time, the probability of changes in the rank order of the original distribution due to the addition of the random variable also increases with that variables' variance. As noted by many authors before (e.g., Lautenschlager, 1994), if there is no variance in faking, there are also no changes in rank order and, thus, hiring decisions or validity.

Unfortunately, there is very little known about the variance in faking in actual applicant settings. For various reasons, the few studies that directly addressed questions of faking variance do not permit an examination of the variance of the true score relative to the faking score in actual applicant settings (e.g., Kluger & Colella, 1993; McFarland & Ryan, 2000; Robie, Born & Schmit, 2001). Some indirect evidence is available from studies comparing personality scores (i.e., the sum of true and faking scores) across applicant and incumbent samples. In the absence of a meta-analysis on this issue (Birkeland, Manson, Kisamore, Brannick & Liu, 2003, report only mean effects), findings from large-scale primary studies appear not very conclusive. Depending on the subscale, Hough (1998) reports partially enhanced and partially unchanged variances in her applicant samples relative to incumbents; Smith, Hanges and Dickson (2001) observed mostly restricted scores in applicants but with some exceptions, and Zickar et al. (2004) found examples of restriction and enhancement along with negligible differences between applicants and incumbents.

It would be important to know more about the variance of faking, because this would allow for indirect deductions of the effects of faking on validity and, in turn, to learn more about the nature of what is referred to as faking itself. If faking were a normally distributed random variable in the sense defined above, it could be expected that the variance of faked scale scores is enhanced, relative to true scores. If two uncorrelated normally distributed variables are summed up, the variance of the composite equals the summed variances of its elements, and the standard deviation follows the equation $s_{(x1 + x2)} = \sqrt{(s_{x1}^2 + s_{x2}^2)}$. For example, if the faking scores were uncorrelated with the true scores and had the same variance, we would expect the standard deviation of the sum of these scores to equal $\sqrt{2} = 1.44$ times the standard deviation of the true scores. If we don't find range enhancement on fakable measures in applicant settings, we have to abandon the preliminary assumption of randomness in terms of univariate and bivariate distributional properties, which has important implications for the questions at hand.

Failure to find the expected range enhancement after adding a faking score can have various causes, but the implications of these causes are likely to be similar. For example, it is possible that faking is normally distributed, but negatively correlated with true scores. Under these conditions, the negative correlation between true scores and faking eliminates the range enhancement effect of summing these variables. Compared to faking as a random variable, the probability of changes in rank order will rise, because the added component is negatively correlated with the original distribution. It is also likely, but not necessary, that faking is negatively correlated with the criterion under these conditions, because true scores are positively related to the criterion. If this is the case, we sum up a positive and a negative predictor, which likely leads to a stronger impairment of validity. If, however, faking is negatively related to true scores but uncorrelated with the criterion, it may even suppress irrelevant variance in true scores and, thus, improve validity. If that happens, defining faking as "response distortion" and the changes in rank order as indicators of lowered decision accuracy becomes highly questionable.

A second possibility to explain a lack of range enhancement (or even restriction) in applicant scores is via ceiling effects. A ceiling effect may occur because the range of possible scores on a test is not wide enough to cover the sum of the true score and the additional faking component for all test takers. This is the place where the concept of opportunity to fake becomes useful. A ceiling effect causes a distribution of the sum of the true and faking scores, which is negatively skewed. Provided that the true scores are distributed normally, the faking scores would have to be extremely negatively skewed to cause such a distribution, which means that there is a majority of extreme fakers with few outliers faking only modestly. This is in line with the logic of experimentally induced faking, which leads to a clustering of total scores at the high end of the distribution because everybody is instructed to score as high as possible. This is likely to lead to very high negative correlations between original (or true, for that matter) scores and faking scores. For example, McFarland and Ryan (2000), who used difference scores to measure faking in an experimental within-subjects design, report correlations around r = -.70 between their original and faking scores. In principle, these negative correlations should have the same effects as mentioned above for the case of normally distributed faking scores, but the size of these effects on validity and decisions should be stronger.

Although Rosse et al. (1998) explicitly expected a negatively skewed distribution of faking scores, there appears to be some inconsistency with their emphasis on a small group of extreme fakers who supposedly are mainly responsible for effects on decision accuracy. Negative skewness implies that the majority of faking scores is close to the upper end of the distribution, and that we find a sizeable tail only at the lower end. The assumption that there are extreme positive outliers on the faking distribution is more consistent with a positive skewness. Consistent with that line of reasoning, Robie et al. (2004) report more positively skewed distributions in applicant than in incumbent samples across all dimensions of the five-factor model of personality. Thus, it is possible that faking scores are in fact positively skewed. Under these conditions, one would still expect a negative correlation between true scores and faking, with all its implications for outcomes of interest, but not to the same extent as with a negative skewness.

In summary, the above considerations suggest that the effects of faking on both decision accuracy and validity depend on the variance of faking, its skewness, and its covariance with the true score and criterion. Further, it can be expected that the effects of these univariate and bivariate properties of faking on validity and decision accuracy generally point into the same direction. More complicated are the questions of (a) how various combinations of distributional properties affect the outcomes, and (b) how sensitive validity and decision criteria are to variations of distributional properties. In the following section, these questions are addressed by systematically varying distributional properties of simulated faking scores and investigating the effects of these variations on outcomes in a real data set.

An empirical example

Method

Sample

The present study was based on a data set from a sample of N = 272 job apprentices in Germany who received a three-year job training in various occupations in the food industry. The training consists of practical elements on-the-job and theoretical lessons taken in a specific school. Thus, all participants were regularly employed job incumbents. The study was conducted under anonymous conditions in school during classes, and participants were instructed to respond honestly (see Marcus & Wagner, 2005, for more details). Mean age was 17.7 years, and 41.5 % of the participants were women.

Measures and procedure

The predictor measure in this study was the *Inventar berufsbezogener Einstellungen und Selbsteinschätzungen* (IBES; Marcus, 2006), a German language integrity test that contains both overt and personality-based scales. An overall score across all subscales is used for the present analysis, as this corresponds to the necessity to aggregate predictor measures in order to arrive at a selection decision in practice. Five different criteria were measured in the present study, representing a fairly broad range of job-related criteria of common interest in I/O psychology. Participants completed self-report measures of overall job performance (3 items assessing performance with respect to absolute and relative standards, $\alpha = .57$), job satisfaction (4 items tapping into various facets of job satisfaction, $\alpha = .63$, Iwanowa, 2004), and counterproductive behavior (17 items tapping into multiple forms of counterproductivity, $\alpha = .83$, see Marcus & Wagner, 2005). In addition, training performance was measured by grades averaged across three commonly taught subjects ($\alpha = .59$), and supervisory ratings of job performance (6 items tapping into task-related, contextual, and overall performance on-the-job, $\alpha = .81$) were obtained for a subsample of N = 175.

Faking was not measured nor experimentally induced in this study. Instead, simulated faking scores with varying properties were created and added to the predictor scores. For each value of the parameters manipulated, analyses were run on a set of five variables each of which resembled the same distributional properties but whose values were assigned differently to participants across replications. This repeated measurement increases the stability of results and permits to examine the variation of findings. The mean of the added variables was always set at .80 standard deviation units of the original score of the integrity test. This choice resembled the mean difference between the present sample and simulated applicant settings for this measure (Marcus, 2006), but has no effect on the findings reported.

First, normally distributed random variables with systematically varied variance were created. The values chosen for this variation were .20, .50, and 1.00, units of the standard deviation of the original predictor score, simulating small, medium, and large variance in faking, respectively. Adding these variables to the original integrity score, of course, led to enhanced variance in the "faked" predictor. To simulate faking scores that do not lead to range enhancement, the random variables with medium (.50 *SD* units) and large (1.00 *SD* units) variance were combined with negatively weighted integrity scores whose weights were chosen such that the variance of the original integrity score plus the added faking score (which now correlated negatively with integrity) resembles that of the integrity score alone. This manipulates the covariance of the faking scores with true scores in addition to the univariate variance manipulation, but not the skewness (with .20 *SD* units, this manipulation would have been trivial and was waived).

Skewness of faking scores was manipulated by first creating a set of positively skewed random variables with the same mean and variance as the normally distributed random variables with medium-sized (.50 SD units) variance. A gamma distribution with the shape parameter 2.56 and the scale parameter 3.2 satisfies these conditions (cf., e.g., Hays, 1977). These skewed random variables were then transformed in the same fashion as described for the normally distributed random variables above to create a composite score that does not differ from the original integrity score in terms of variance (i.e., the skewed faking scores were correlated negatively with the original integrity score). Finally, a negatively skewed distribution of faking scores was created in a similar way by first creating gamma-distributed variables that were either uncorrelated or negatively correlated with integrity, weighting these appropriately to preserve the medium-sized variance in the faking score and the unchanged variance in the total score, and transforming them such that the skewness is reversed. The absolute values for the skewness of gamma distributions varied between .78 and 1.49 across the five random variables, that for the resulting aggregate observed scores between .29 and .58.

Thus, there are seven sets of simulated faking scores, each including five repeated measurements: three variations of variance in faking only, two variations of faking negatively correlated with the true score at different levels of variance, and two variations of skewness, one being positive and one negative, at medium-sized variance of faking combined with negative correlation with the true score. This should allow for a detailed examination of the effects of these parameters on validity and hiring decisions.

Results and discussion

Table 1 shows the effects of the simulated faking scores on the validities of the predictor across criteria. As can be seen in the first data column, the original integrity score correlates with all five criteria at varying levels (r = .18 to .54, all ps $\leq$ .01), which resembles common findings reported in the literature for this type of tests. For each set of manipulated faking

Table 1:
Effects of Simulated Faking Scores on Criterion-Related Validities

Predictor	OS	(1) .2 SD, random			(2) .5 SD, random			(3) 1 SD, random			(4) .5 SD, correlated			(5) 1 SD, correlated			(6) .5 SD, pos. skew			(7) .5 SD, neg. skew		
Criteria	r	r_{mean}	r range	Δr	r_{mean}	r range	Δr	r_{mean}	r range	Δr	r_{mean}	r range	Δr	r_{mean}	r range	Δr	r_{mean}	r range	Δr	r_{mean}	r range	Δr
Training performance (grades)[1]	.235	.238	.23-.25	.003	.225	.22-.23	-.010	.188	.14-.22	-.047	.221	.22-.23	-.014	.106	.05-.14	-.129	.208	.14-.26	-.027	.176	.13-.22	-.059
Job performance (supervisory rating)	.194	.190	.17-.20	-.004	.157	.13-.18	-.037	.131	.12-.15	-.063	.150	.12-.18	-.044	.053	.04-.07	-.141	.143	.12-.18	-.051	.145	.12-.20	-.049
Job performance (self-rating)	.180	.178	.16-.19	-.002	.167	.15-.18	-.013	.126	.15-.18	-.054	.163	.15-.18	-.017	.059	.02-.10	-.121	.160	.14-.20	-.020	.111	.07-.14	-.069
Job satisfaction	.381	.376	.35-.39	-.005	.338	.29-.39	-.043	.301	.25-.35	-.080	.327	.27-.39	-.054	.167	.10-.24	-.214	.309	.28-.34	-.072	.245	.21-.29	-.136
Counterproductive behavior[1]	.543	.532	.52-.55	-.012	.491	.47-.53	-.052	.406	.36-.46	-.137	.477	.45-.52	-.066	.207	.14-.30	-.336	.455	.44-.47	-.088	.421	.39-.46	-.122
Mean across criteria	.307	.303		-.004	.276		-.031	.231		-.076	.268		-.039	.119		-.188	.255		-.052	.220		-.087
Percentage change to OS				-1.3			-10.1			-24.8			-12.7			-61.2			-16.9			-28.3

Note. OS = Original score of integrity test. Otherwise, top row describes distributional properties of simulated scores: .2, .5, 1 SD are standard deviations of the faking score, in SD units of the original score; random: normally distributed and uncorrelated with OS; correlated: normally distributed and negatively correlated with OS; pos./neg. skew: negatively correlated with OS and positively or negatively skewed. R_{mean}, r range, and Δr indicate the mean of validity coefficients across five runs per condition, the range of these validities, and the average change compared to the validity r of the OS. 1 = Criterion recoded such that higher values indicate more desirable performance. Due to listwise deletion of missing values per criterion, N = 253 for training performance, 170 for supervisory ratings of job performance, and 266 to 272 for the remaining criteria.

scores added to the original score, the following data columns show the average validity across the five repeated measurements, their highest and lowest values across repetitions, and the average validity change compared to the original score.

It is evident that adding a random score with a variance as small as .2 *SD* units of the original score had practically no effect on validity (mean r_{diff} = -.004, or 1.3 % of the original validity). The detrimental effect of a random faking score on validities rose sizably as its variance increases to a medium size (mean r_{diff} = -.031, or 10 % of the original validity), and a large size (mean r_{diff} = -.076, or 25 % of the original validity). The fourth and fifth blocks in Table 1 contain the data for faked scores in which the faking component is negatively correlated with the true score to hold the variance of the total score constant. These data are best compared with the respective random variables with the same variance. With a medium-sized variance of faking, relatively small negative correlations between faking and true score are necessary to eliminate the additional variance in the total score (r_{mean} = -.25), and the additional effect on validities compared to randomly faked scores is small (mean r_{diff} = -.039 vs. -.031). By contrast, faking scores with large variances required a substantial negative correlation with the true score to eliminate the additional variance (r_{mean} = -.56), which was accompanied by a very substantial loss of validity (mean r_{diff} = -.188 vs. -.076, or 61 vs. 25 %). In the last two data blocks, positive or negative skewness was added to the medium-sized, correlated faked scores in the fourth block. Somewhat unexpectedly, even a positive skewness further impaired validities across all criteria (mean r_{diff} = -.052 vs. -.039, or 17 vs. 13 %). This detrimental effect of skewness was considerably stronger if faking scores were negatively skewed (mean r_{diff} = -.087, or 28 % of original validities).

Table 2 shows the effects of simulated faking on hiring decisions. For each set of manipulations, the percentage of overlap between hiring decisions based on original scores and faked scores is shown for various selection ratios (.05, .10, .20, and .50, respectively). This is a common indicator of the effect of faking on hiring decisions in the literature, but a rather weak proxy of actual decision *accuracy*. Counts of changed decisions due to faking do not examine the amount of faking involved. For example, it does make a difference whether the true score of an applicant who successfully faked to be among the top ten percent of the distribution is actually at, say, the 88^{th} percentile or the 8^{th} percentile. More importantly, the fact that a decision has changed does not tell anything about the standing of those hired on job-related criteria. Thus, a more adequate measure of decision accuracy may be the average scores of those hired on job-related criteria. Mean *z*-scores of the selected groups on the five criteria measured in the present sample are therefore presented in Table 2.

With regard to the criterion of decision overlap, it is evident that hiring decisions were indeed more sensitive to simulated faking than validity. Even with a small variance of faking, a significant number of decisions changed when selection ratios were low. Further, in most cases decision overlap was smallest with the .05 selection ratio, and it was always highest with .50 selection ratio. However, if one considers that decision overlap by chance is an inverse function of the selection ratio, this effect appears not very pronounced, especially when comparing the more selective ratios of .05, .10, and .20. It is also evident that decision overlap was affected by distributional properties of faking in very similar ways as validity. Regardless of selection ratio, decision overlap decreased sizably as the variance of faking increased. Adding a negative correlation with true scores to the faking score had a modest impact on decision overlap when the variance in faking was of moderate, but substantially increased the proportion of changed decisions when faking variance was large. This is the same pattern of findings as reported for validity above. However, adding either positive or negative skewness to the faking score appeared to have no clear-cut effects on decision overlap vis-à-vis comparable normal distributions.

The lower portion of Table 2 shows the effects of simulated faking on decision accuracy, as measured by mean *z*-scores across criteria. There are several noteworthy aspects of these findings. First, unlike decision overlap, decision accuracy does not appear to be more sensitive to faking than validity. Considering the average percentage of loss compared to true scores as a measure of sensitivity relatively comparable between validity and decision accuracy, results are strikingly similar. Faking scores with small variance had virtually no effect on both outcomes, moderately variant faking decreased decision accuracy by an average of 6.8 % percent across selections ratios and criteria (validity: 10.1 %), and faking scores with large variance by 24.3 % (validity: 24.8 %). Corresponding values for the remaining four simulated faking conditions, ordered from the fifth to last data column in Table 2, are: 8.7 % (validity: 12.7 %), 66.2 % (validity: 61.2 %), 11.6 % (validity: 16.9 %), and 13.7 % (validity: 28.3 %). The correlation between effects of faking on validity and decision accuracy across conditions is .97, and the average distance in percent is 3.6, indicating a slightly higher sensitivity of validity to faking. Moreover, effects on decision accuracy did not systematically increase with lower selection ratios. The mean percentage of loss in decision accuracy across faking conditions was 15.5 % at a selection ratio of .05, 15.0 % at .10, 24.5 % at .20, and 19.8 % at .50. If anything, these findings appear to point to a *lower* sensitivity of decision accuracy at more competitive selection ratios. Finally, the most notable specific differences between outcomes occurred for the condition of negatively skewed faking scores. Validity was considerably more sensitive to faking scores distributed this way than was decision accuracy. By far the largest effects on all outcomes, however, were observed for the combination of large faking variance and a negative correlation with true scores.

General discussion

Since a number of meta-analyses consistently revealed that taking personality tests and other fakable selection devices in applicant settings does not diminish criterion-related validities, there has been some debate as to whether this finding may have masked effects of faking on other outcomes of personnel selection, like hiring decisions and decision accuracy. The present paper was aimed at clarifying some of the issues involved in that debate by means of conceptual and empirical analyses of the effects of distributional properties of faking on various outcomes. For this purpose, "faking" was not defined by its content or psychological meaning, but in purely technical terms as a variable with varying psychometric properties, which is added to the true score. Thus, the present paper does not directly address the nature of faking, but results may also be used to infer features of this nature indirectly in light of previous research in this area.

First, a conceptual analysis using Taylor and Russell's (1939) classic framework of utility showed that, in the absence of any assumptions on faking, the likelihood of being hired on the basis of fakable test scores is expected to be a positive function of the true score on the trait measured by this test, and, if true scores relate positively to criteria, also a positive function of these criteria. It was further shown that maximum test scores, and therefore the

Table 2:
Effects of Simulated Faking Scores on Decision Overlap and Accuracy

Predictor	OS	(1) .2 SD, random	(2) .5 SD, random	(3) 1 SD, random	(4) .5 SD, correlated	(5) 1 SD, correlated	(6) .5 SD, pos. skew	(7) .5 SD, neg. skew
Selection Ratio		Percentage of overlap in hiring decisions with decisions based on OS						
.05		81.4	68.6	41.4	62.9	15.7	61.4	67.2
.10		85.7	72.1	53.6	67.9	25.0	65.7	61.4
.20		85.5	73.1	57.8	70.5	34.9	68.7	66.9
.50		94.2	84.6	77.8	83.7	62.4	83.6	79.6
		Mean z-scores of hired groups on criteria						
Training performance at SR								
.05	.455	.578	.573	.502	.562	.223	.421	.479
.10	.499	.488	.521	.424	.477	.219	.424	.273
.20	.270	.304	.323	.328	.319	.171	.333	.265
.50	.258	.229	.189	.144	.183	.084	.175	.167
Job performance (supervisor) at SR								
.05	.445	.319	.292	.313	.322	.208	.481	.508
.10	.204	.321	.276	.088	.304	-.037	.258	.392
.20	.328	.299	.256	.229	.256	.015	.242	.237
.50	.102	.117	.100	.127	.092	.048	.113	.094
Job performance (self) at SR								
.05	.138	.166	.224	.128	.195	.064	.281	.147
.10	.161	.204	.290	.157	.300	-.011	.214	.266

.20	.344	.308	.235	.152	.235	.031	.225	.237
.50	.103	.122	.121	.140	.127	.060	.139	.101
Job satisfaction at SR								
.05	1.147	1.140	1.080	.798	1.006	.353	1.012	.905
.10	.861	.824	.744	.637	.723	.242	.724	.663
.20	.687	.641	.504	.420	.499	.223	.497	.480
.50	.284	.298	.264	.243	.263	.140	.251	.222
Counterprod. behavior at SR								
.05	.984	.984	.895	.653	.860	.345	.839	.883
.10	.806	.808	.806	.599	.778	.248	.697	.713
.20	.684	.678	.617	.484	.617	.257	.583	.611
.50	.456	.437	.389	.334	.383	.166	.349	.349
Mean across criteria at SR								
.05	.634	.637	.613	.479	.589	.239	.607	.584
.10	.506	.529	.527	.381	.516	.132	.463	.461
.20	.463	.446	.387	.323	.385	.139	.376	.391
.50	.241	.241	.213	.198	.210	.100	.205	.187
Mean percent difference to OS at SR								
.05		.5	-3.3	-24.4	-7.1	-62.3	-4.3	-7.8
.10		4.5	4.2	-24.7	2.0	-73.9	-8.5	-8.9
.20		-3.6	-16.4	-30.2	-16.8	-70.0	-18.8	-15.5
.50		0.0	-11.6	-17.8	-12.9	-58.5	-14.9	-22.4

Note: SR = selection ratio. Otherwise, see Table 1 for descriptions.

opportunity to fake, are irrelevant for hiring decisions in the presence of faking. These conclusions may not be entirely novel, but it seems necessary to make the points, because some arguments used in discussions of this issue appear to implicitly convey the impression that high true scorers were actually disadvantaged by the use of fakable tests originally designed to measure these true scores. On the contrary, high true scorers on a job-relevant personality trait would, in general, be disadvantaged by *not* attempting to measure this trait, because this would mean that the probability of being hired had no relationship at all to the trait in question.

The actual relationship between faking and several outcomes, however, can be expected to depend in large parts on distributional properties of faking. Namely, application of general statistical principles suggested that the variance and skewness of faking, along with its covariance with the true score would affect validity as well as hiring decisions and the accuracy of these decisions, whereas the mean does not. While these considerations may seem purely technical at first sight, they are important in order to judge the soundness of the argument that detrimental effects of faking on decision accuracy may have been overlooked in validation studies.

These conceptual thoughts were tested empirically based on data collected under anonymous conditions, which contained a broad range of job-related criteria. Adding simulated faking scores with systematically varied distributional properties to the predictor allowed for an examination of the effects of these properties. In general, empirical findings confirmed conceptual considerations and shed new light on the mechanics of personnel selection in the presence of faking. For example, it was shown that even under extremely undesirable assumptions on faking (large variance combined with high negative correlations), true scores were positively related to the probability of being hired. Decision overlap with original scores exceeded chance levels at all selection ratios under this condition (see data block 5 in the upper portion of Table 2), and it was much more likely for true high scorers to be hired with less extreme faking distributions.

With respect to the single distributional properties investigated, it is obvious that the variance of faking plays a crucial role, regardless of which outcome is considered. At small, moderate, and large levels of faking variance, effects on all outcomes also varied from small (or negligible) to substantial. If faking were unrelated to true scores, we would generally expect that the outcomes of faking were largely a direct function of its variance. Finally, it is noteworthy that these results occurred regardless of whether criteria were measured by self-reports (and thus may be subject to method bias and even shared components of social desirability), supervisory ratings, or the more objective criterion of grades in school exams. The only things that seemed to make a substantial difference in the present data were the distributional properties of faking.

As we currently know little about the actual variance of faking in selection settings, it may be preliminarily assumed that the variance of faking is moderate, but the assumption of faking as a normally distributed random variable may be unrealistic. As there is no consistent evidence of range enhancement in applicant scores on personality tests, the assumption that these scores contain a faking component requires that faking is negatively correlated with true scores. The size of this correlation depends on the variance of faking if the variance of the total score is held constant. Thus, variance affects the outcomes of faking both directly and indirectly via correlations with true scores. As a result, the combination of large variance with high negative correlations led to a dramatic decline in all outcome indicators observed in the present study. Adding skewness to moderate levels of the other parameters further affected validity and decision accuracy negatively, especially when data were negatively skewed, but had no clear effect on decision overlap. Hence, contrary to some beliefs, hiring decisions appear to be more robust than validity against violations of normality in test scores. This result occurred regardless of what selection ratio was chosen. Taken together, the present findings suggest that the effects of various distributional properties of faking can be sizeable and interdependent, which coincides with a paucity of field research on precisely these properties. Clearly, this is a promising avenue for future research.

The data also permit to address the crucial question of sensitivity of different outcomes of selection to faking. The present findings suggest that the tenability of the assumption of faking effects on decision criteria in the absence of effects on validity largely depends on which type of decision criterion is of interest. In the present study, decision accuracy was measured as the average criterion score of the group selected. This measure resembles a major parameter in psychometric utility analysis (e.g., Cronbach & Gleser, 1965), which defines utility in terms of the gain in performance expected from using a selection device compared to not using it. I used this parameter for comparing the loss in performance (or other criteria) expected from faking a selection device compared to not faking it. If decision accuracy is measured this way, the present results clearly point to the conclusion that, if an effect of selection setting on the criterion-related validity of a test is not found, it is extremely unlikely that there is any hidden effect on decision accuracy using the same criterion. Effects on both outcomes were virtually identical, almost regardless of the distribution of faking scores as well as selection ratios.

If overlap between decisions based on faked and unfaked scores is examined, the picture slightly changes. Effects on this outcome are still very similar to those on validity or decision accuracy, but decision overlap was indeed more sensitive to faking scores with small variance and changes of the selection ratio. The crucial aspect about regarding changed decisions as false is that this requires defining changes in test scores that lead to altered decisions as an inherently immoral activity, independent of its relation to job-related criteria. This would lead into basically philosophical discussions, which are beyond the scope of the present paper and what I/O psychology is mostly concerned with.

Staying closer to the usual issues of applied psychology, the present results may well be interpreted as indirect evidence of the likely nature of what is used to be called faking. The substantial evidence of a lack of any detrimental effect of faking on validity in real applicant settings is consistent with a range of scenarios. For instance, there may be no faking at all in the real world, or the variance and other distributional properties of faking may be so trivial that their effects were not revealed in meta-analyses. These scenarios appear unlikely, however, given the fact that validity generalization studies comparing applicants with incumbents were partially based on hundreds of thousands of participants (for the case of integrity tests, the type of predictor used herein, see Ones et al., 1993). If faking varies at least moderately, and especially if this is combined with undesirable distributional properties, we would expect to find a detriment. As mentioned earlier, a large detriment in validity is consistently found in experimental research on induced faking (Hough, 1998). These results are most consistent with the present condition (5) of highly variant faking scores combined with large negative correlations with the true score (e.g., McFarland & Ryan, 2000, report standard deviations and true score correlations of their induced faking scores that were consistently even larger than in the present condition 5). Findings from induced faking studies, however, tend to be inconsistent with what is found with real applicants, though to varying degrees.

Under distributional assumptions appearing realistic in light of earlier research, the only explanation for failure to find detrimental effects on validity in the field is that what we tend to think of as the "faking" component in applicant scores actually is more complex, containing at least one facet, which itself is a valid predictor of job-related criteria. A theoretical interpretation of the content of this construct is beyond the scope of the present paper (see, e.g., Hogan & Hogan, 1998; Marcus, 2005; Mount & Barrick, 1995, for various proposals), but that there is some component in apparently "faked" applicant scores that predicts performance beyond true scores appears to be highly likely in light of the present and earlier research.

The present study was limited in that it lacked the means to address the psychological meaning of deliberately changing scores on tests used in applicant settings. It was also limited in that the faking scores investigated in this study were simulated, not actually observed. The latter limitation means that it was not possible to decide which of the simulated scenarios approximates reality most closely, or if reality is actually more complex (e.g., involving nonlinear relationships). However, the present research tested a wide range of plausible scenarios, and the major finding that effects on decision accuracy in the absence of effects on validity are unlikely occurred with all variations of distributional properties. Still, the issues just mentioned are clearly in need for more research. The present paper was set out to clarify a range of possible effects (and lack thereof) of faking on outcomes of most interest to scholars and practitioners of personnel selection. In conclusion, there appears to be little reason to abandon the use of selection instruments simply because they can be faked, as long as there is evidence of validity.

References

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

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

Birkeland, S., Manson, T., Kisamore, J., Brannick, M., & Liu, Y. (2003, April). *A meta-analytic investigation of job applicant faking on personality measures.* Paper presented at the 18th Annual Conference of the Society of Industrial and Organizational Psychology, Orlando, FL.

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6*, 199-214.

Bozeman, D. P., & Kacmar, K. M. (1997). A cybernetic model of impression management processes in organizations. *Organizational Behavior and Human Decision Processes, 69*, 9-30.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16,* 81-106.

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology, 69*, 498-508.
Hays, W. L. (1977). *Statistics for the social sciences* (2nd ed.). London, UK: Holt, Rinehart & Winston.
Hogan, R. T., & Hogan, J. (1998). Theoretical frameworks for assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 27-53). San Francisco, CA: Jossey-Bass.
Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209-244.
Iwanowa, A. (2004, September). *Kurzer Fragebogen zu Formen der Arbeitszufriedenheit (FAZ): Eine empirische Überprüfung* [A short questionnaire of forms of job satisfaction (FAZ): An empirical assessment]. Paper presented at the 44th Kongress der Deutschen Gesellschaft für Psychologie, Göttingen, Germany.
Kluger, A. N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology, 46*, 761-780.
Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes, M. D. Mumford & W. A. Owens (Eds.), *Biodata handbook* (pp. 391-419). Palo Alto, CA: Consulting Psychologists Press.
Marcus, B. (2003). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte“ Antworten wirklich nicht wünschenswert? [Personality testing in personnel selection: Is “socially desirable” responding really undesirable?] *Zeitschrift für Psychologie, 211,* 138-148.
Marcus, B. (2005). *Self-Presentation as Adaptive Behavior in Personnel Selection Settings: A Theory and an Application*. Unpublished manuscript. London, ON: University of Western Ontario.
Marcus, B., (2006). *Inventar berufsbezogener Einstellungen und Selbsteinschätzungen* [Job-Related Attitudes and Self-Evaluations Inventory]. Test manual. Göttingen, Germany: Hogrefe.
Marcus, B., & Wagner, U. (2005). *Combining Personality Traits and Perceptions of Vocation and Job Situation to Predict Counterproductive Work Behavior*. Unpublished manuscript. London, ON: University of Western Ontario.
McFarland, L. A. & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.
Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65*, 131-149.
Mount, M. K., & Barrick, M. R. (1995). The big five personality dimensions: Implications for research and practice in human resource management. *Research in Personnel and Human Resources Management, 13,* 153-200.
Mueller-Hanson, M., Heggestad, E. D., & Thornton, G. C. (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,* 245-269.
Ones, D. S. & Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection. *Journal of Applied Psychology, 81*, 660-679.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Robie, C., Born, M. P., & Schmit, M. J. (2001). Personal and situational determinants of personality responses: A partial reanalysis and reinterpretation of the Schmit et al. (1995) data. *Journal of Business and Psychology, 16*, 101-117.

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

Schlenker, B. R., Britt, T. W., & Pennington, J. (1996). Impression regulation and management: Highlights of a theory of self-identification. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (Vol. 3, pp. 118-147). New York: Guilford Press.

Smith, D. B., Hanges, P. J., & Dickson, M. W. (2001). Personnel selection and the five-factor model: Reexaming the effects of applicant's frame of reference. *Journal of Applied Psychology, 86,* 304-315.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

Viswesvaran, C., Ones, D. S. & Hough, L. M. (2001). Do impression management scales in personality assessment predict managerial job performance ratings? *International Journal of Selection and Assessment, 9*, 277-289.

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