Chapter 5

Review of Faking in Personnel Selection

Chris D. Fluckinger, Michael A. McDaniel[1] and Deborah L. Whetzel[2]

The University of Akron,

[1]Virginia Commonwealth University and Work Skills First, Inc.

[2]Human Resources Research Organization (HumRRO)

ABSTRACT

Faking is a potential threat to most forms of noncognitive measurement, even though faking has been defined and analyzed many different ways. This chapter details current understanding of faking in applied settings and is directed toward applied practitioners and researchers using noncognitive measures who might be concerned with faking. Three primary topics are covered: a review of strategies to detect faking, a presentation of the ways that faking can influence test properties, and a critique of the various ways that have been implemented to guard against faking. The main conclusion reached is that although most noncognitive measures can be intentionally distorted and are therefore susceptible to faking, not everyone is expected to fake in motivated settings. The challenge for researchers and practitioners is to find ways to maximize noncognitive test utility, both by encouraging job-relevant responding from applicants (potentially through state-of-the-art test formats and instructions) and by thoughtfully integrating noncognitive measures into the goals of a given selection system.

The issue of faking on noncognitive measures has been the subject of a voluminous amount of applied and experimental research in the fields of selection and human resource management. Although a great deal of knowledge has accumulated regarding the extent of faking and the effectiveness of a number of potential solutions, there remain key areas open to debate concerning best practices in dealing with faking (Hough & Oswald, 2005). This review will focus on three basic aspects of faking on noncognitive measures that will prove useful to researchers and applied practitioners: 1) the usefulness of different strategies to identify faking, 2) potential challenges to test delivery and

noncognitive measurement that faking can present, and 3) a summary and critique of common strategies to combat faking. Unlike many summaries of faking, which provide important theoretical analyses of the construct of faking (Snell, Sydell & Lueke, 1999; McFarland & Ryan, 2000), the goal of this review is to provide practitioners and researchers interested in using noncognitive measures with a solid understanding of the practical issues related to faking in test delivery and assessment.

HOW RELEVANT IS FAKING?

Noncognitive measurement is very important in the selection and placement of employees in organizations. Over a third of corporations in the United States use personality testing, and the industry takes in nearly $500 million in annual revenue (Rothstein & Goffin, 2006). When one factors in other forms of testing that measure noncognitive constructs, including integrity tests, biographical information (or biodata), interviews and assessments, the importance of noncognitive measurement increases. Of course, one of the major limitations of noncognitive measures, or information that is independent of general mental ability (such as numerical ability or short-term memory span) is that people can fake, or intentionally misrepresent the truth about their past or their behavioral tendencies.

The fact that applicants may be highly motivated to fake in order to gain employment has raised many questions as to the usefulness of noncognitive measures in high-stakes settings. Some have even gone far enough to suggest that noncognitive measurement, particularly personality, should not be used for employee selection (Murphy & Dzieweczynski, 2005). The potential problems with personality measurement are so great that Murphy has offered this analogy: "Faking research is like painting a house that is burning down" (personal communication as cited in Hayes, 2007). However, the majority of selection researchers and practitioners would stop short of this drastic assessment. Research strongly supports the value of noncognitive measures, as influential meta-analyses have shown that personality, integrity, and biodata are predictive of important work outcomes, such as job performance and training performance (Barrick, Mount & Judge, 2001; Hough & Furnham, 2003; Schmidt & Hunter, 1998; Bobko, Roth & Potosky, 1999). This review is based on the idea that noncognitive measures are predictive of work outcomes even under conditions where faking is likely, a notion summed up nicely by Rothstein and Goffin that there is "abundant grounds for optimism that the usefulness of personality testing in personnel selection is not *neutralized* by faking" (p. 166).

Unfortunately, even though noncognitive measures often produce moderate predictive validities, there are a number of other ways that faking can wreak havoc on a selection system, including the construct validity of measures and changes in the rank-order of who is selected. Human resource practitioners seem to have mixed views

on the importance of faking as well. On a survey of British personnel staffers, Rees and Metcalfe (2003) found that nearly half (48%) believed that noncognitive tests were difficult to fake (28% were unsure and 20% thought faking was easy). On the other hand, over two-thirds (68%) of Society for Human Resource Managers (SHRM) thought that integrity tests were not useful because they were susceptible to faking (Rynes, Brown & Colbert, 2002). Similarly, 70% of professional assessors believe that faking is a serious obstacle to measurement (Robie, Tuzinski & Bly, 2006). These results suggest that there may be considerable variability in practitioner beliefs about the impact of faking. This review attempts to give practitioners a full appreciation of the challenges of measuring faking, accounting for its presence in selection systems, and attempting to prevent or reduce it.

Evidence of Faking

The concept of faking is relatively straightforward: people engage in impression management and actively try to make themselves appear to have more desirable traits than they actually possess. However, identifying actual faking behaviors in a statistical sense has proven to be exceedingly difficult (Hough & Oswald, 2005). The main problem is that attempts to fake can show up in a number of statistical indicators, including test means, social desirability scales, criterion-related validity, actual or simulated hiring decisions, and construct validity. There is ample evidence that faking likely influences most of these crucial test properties. This section will describe the research evidence on the extent to which faking affects each of these test properties, as well as a few additional measures that may prove useful in the future.

Social Desirability

The construct of social desirability basically stipulates that the tendency to manage the impression one maintains with others is a stable individual difference that can be measured using a traditional, Likert-style self-report survey (Paulhus & John, 1998). Using this logic, Rosse, Stecher, Miller and Levine (1998) found that applicants for a job had higher social desirability scores than incumbents, which was interpreted as evidence that the applicants were faking. The initial view regarding social desirability from an applied perspective was that it could be measured in a selection context and used to correct, or adjust, the noncognitive scores included in the test. For example, someone who has high social desirability scores would be viewed as likely to have inflated his/her conscientiousness scores, leading to the logical decision to adjust the personality scores downward to reflect that person's true standing on the trait.

Perhaps unfortunately, social desirability does not function as frequently theorized. A meta-analysis by Ones, Viswesvaran and Reiss (1996) showed that social desirability does not account for variance in the personality-performance relationship. This means that knowledge of a person's level of social desirability will not improve the measurement

of that person's standing on a noncognitive trait. In addition, there is debate regarding whether people who fake will also have corresponding high scores on a social desirability scale. Viswesvaran and Ones (1999) conducted a meta-analysis showing that those who receive instructions to look as good as possible (fake-good) have higher personality scores than those under honest instructions (d = 1.06). However, this represents a common methodological difficulty of interpreting faking studies, because actual applicants may respond quite differently than those who are directly instructed to fake. A meta-analysis by Hurd, Barrett, Miguel, Tan and Lueke (2001) compared scores from honest, fake-good, applicant and incumbent groups and found that only the fake-good group had significantly higher social desirability scores. This indicates that actual applicants often respond in ways that are not likely to be detected by social desirability scores, a finding supported elsewhere (Alliger, Lilienfeld & Mitchell, 1996; Zickar & Robie, 1999). Given these findings, it is very likely that social desirability is a poor indicator of applicant faking behavior.

Mean Differences

It is an oft-cited finding that noncognitive scores under fake-good instructions lead to higher test means than scores under honest instructions (d H≈.6 across Big 5 personality dimensions; Viswesvaran & Ones, 1999). This pattern is similar when comparing actual applicants and incumbents, with the largest effects found for the traditionally most predictive personality dimensions in personnel selection, conscientiousness (d = .45) and emotional stability (d = .44; Birkeland, Manson, Kisamore, Brannick & Smith, 2006). The effects are not just limited to personality, as a meta-analysis of integrity test means shows the same pattern of increased means in faking conditions (d = .36 to 1.02; Alliger & Dwight, 2000).

To summarize, people have the highest means in experimental, fake-good designs and somewhat lower means in applicant settings, and these means are nearly always higher than honest/incumbent conditions. These are the most consistent findings in faking research, and they are often taken as the most persuasive evidence that faking occurs. However, they are often used to overinterpret the potential danger of faking. The key impact of mean differences in faking involves who is doing the faking and how much. If every applicant in a selection pool fakes the same amount (i.e., raises his/her score by a uniform amount), then the true rank-order of individuals would remain the same, and the criterion-related validity of the test would not be reduced much. However, as detailed later, not everyone fakes, and those who do vary regarding their extent of faking. This means that additional measures of the impact of faking are necessary.

Criterion-related Validity

One of those most important properties for a noncognitive test used in applied settings is the demonstration of criterion-related validity. It is logical to assume that as applicants

fake more, the test will be less able to predict important work outcomes. The situation in reality is much less straightforward, owing partly to the relative lack in the faking literature of appropriate studies (many including poor criteria such as college GPA or other self-report measures) and to the variety of study designs. For example, Douglas, McDaniel and Snell (1996) showed that students' conscientiousness ratings (measured with personality and biodata instruments) were much less predictive of supervisor ratings when they completed the measures under fake-good instructions. The general pattern in applied samples is similar, as predictive validity is highest in incumbent (supposedly honest) samples, slightly lower for applicants, and drastically lower for fake-good directions (see Hough, 1998, for a thorough review). These findings are commonly interpreted as supporting the hypothesis that faking may lower criterion-related validity, but it often does not do so drastically.

There are a number of caveats to this general pattern regarding predictive validity. One is situation strength: when tests are administered in ways that restrict natural variation, criterion-related validity will drop (Beatty, Cleveland & Murphy, 2001). For example, if an organization clearly advertises that it only hires the most conscientious people, then applicants are more likely to fake to appear more conscientious. Another caveat is the number of people who fake. A Monte Carlo simulation found that the best-case scenario for faking is an all-or-nothing proposition: validity is retained with no fakers or many fakers, but if there is a small minority of fakers present, they are likely to be rewarded, thus dragging overall test validity down (Komar, Brown, Komar & Robie, 2008). Finally, the criterion-related validity of the test as a whole may not be sensitive to changes in the rank-ordering of applicants (Komar et al., 2008). Mueller-Hanson, Heggestad and Thornton (2003) tested this assumption by rank-ordering participants from two conditions (honest and fake-good), and then dividing the distribution into thirds. The results indicated that the top third, which included a high percentage of participants who were given faking instructions, had low validity (r = .07), while the bottom third produced high validity (r = .45). These results indicate that noncognitive measures can still produce respectable validities even under conditions where faking is likely to be present, but predictive validity is likely to be somewhat insensitive to changes in rank-ordering of applicants due to faking, and by extension, the quality of hiring decisions.

Selection Decisions

This is the area where faking is likely to have its most detrimental effects in the selection process. It is a common finding that people who fake—identified by higher social desirability scores (Rosse et al., 1998) or by higher proportions of those from a faking condition (Mueller-Hanson et al., 2003)—will rise to the top of the selection distribution and increase their probability of being hired. This situation worsens as the selection ratio is lowered (fewer people are selected, but more of them are likely to be fakers). A

telling example is that of Griffeth, Chmielowski and Yoshita (2005), who obtained applicant personality scores and then honest scores a month later. Out of 60 participants, one individual who was ranked #4 for the applicant test dropped to #52 for the honest test, indicating a large amount of faking. Numerous additional studies have provided similar findings, suggesting that the rank order of applicants will change considerably under different motivational and instructional conditions. This pattern is usually attributed to faking behavior, but it can also be partly explained by random or chance variation. Regardless, these consistent findings demand that users of noncognitive tests cannot simply rely on a test's predictive validity to justify its utility as a selection device.

Construct Validity

The construct validity—or internal structure and reliable relations to other variables—of a test under faking conditions is often somewhat overlooked in applied settings, probably in part because construct validity is not included in analyses of test utility. However, construct validity is crucially important regarding the quality of what is measured, and should be taken into account to support the linkages of a test with the knowledge, skills and abilities specified by the job analysis or competency model. Consistent factor analytic evidence has accumulated indicating that construct validity does indeed drop when faking is likely present. The factor structure of noncognitive tests, especially personality, tends to degrade when applicants are compared with incumbents, as an extra factor often emerges with each item loading on that factor in addition to loading on the hypothesized factors (Zickar & Robie, 1999; Cellar, Miller, Doverspike & Klawsky, 1996). This means that the noncognitive constructs actually change under faking conditions, shedding some doubt as to how similar they remain to the intended, less-biased constructs.

Other types of analyses have shown reduced construct validity as well. Item-response theory (IRT) has been applied to see if certain items are more prone to be faked than others, but these investigations have largely found little evidence that certain items can be identified and altered to increase construct validity (Zickar & Robie, 1999; Griffin, Hesketh & Grayson, 2004). Multi-trait, multi-method analysis has also been applied, resulting in the finding that expected relations between constructs can be severely degraded when faking is present (Douglas, McDaniel & Snell, 1996). Taken together, these results indicate that faking consistently reduces the construct validity of commonly administered measures when faking is likely present.

Additional Indicators of Faking

Of the additional ways to identify faking, response latencies are the most heavily researched. Basically, the underlying hypothesis is that people who are faking will take a different amount of time on average to respond to noncognitive items than those who

are being honest. However, the theory as to whether fakers would take more or less time to respond is still unclear: Should honest responders take longer because they try to find examples to justify their ratings, or should fakers take longer because they need to identify the most desirable answer? The existing results are equally difficult to interpret, as some studies indicate few differences between honest and pseudo-applicants (applying for a hypothetical job; Vasilopoulos, Reilly & Leaman, 2000), and others seem to indicate that fakers take more time (Robie, Brown & Beatty, 2007; Vasilopoulos, Cucina & McElreath, 2005). As computer administration continues to grow and research accumulates, response latencies may be given more weight in determining the extent of faking.

Other technologies have been applied to faking investigations, including polygraphs and brain imaging techniques. Polygraphs, or lie detectors, were banned for use in selection in the United States due to too many false positives and are no longer considered viable faking indicators (Roese & Jamieson, 1993). On the other hand, brain imaging techniques have been increasingly used to investigate the act of conscious deception, which ultimately may allow practitioners to "see" the act of faking. For example, it has been shown that some brain areas are activated during lying (Brodmanns' area), some while telling the truth (cingulate gyrus), and some for both lying and truth-telling (insula; Lengleben, Loughead, Bikler, Ruparel, Childress, Busch and Gur, 2005). Although brain imaging is currently cost prohibitive and legally questionable as a method of detecting faking, it is possible that it could be used for assessment purposes at some point in the future.

Practical Issues in Test Delivery

As should be apparent from the above discussion regarding the different ways to identify faking, the construct of faking is quite complex, and practitioners are left with many aspects to consider when using noncognitive measures for selection purposes. This is in line with the recommended best practice of leading applied personality researchers, as the continuing research agenda has been set in large part to identify boundary conditions where faking becomes too intractable of a problem (Hough & Oswald, 2005). The next section of this review is intended to present a number of practical issues regarding the use of noncognitive measures in high-stakes setting, and by extension the potential impact of faking.

Properties of the Selection System

Two key aspects of selection systems are particularly relevant to the issue of faking: the decision to use multiple-hurdle vs. compensatory systems, and the use and appropriate setting of cut scores. A multiple-hurdle system involves a series of stages that an applicant must pass through to ultimately be hired for the job. This usually involves setting cut scores—a line below which applicants are removed from the pool— at each

step (or for each test in a selection battery). A compensatory system, on the other hand, typically involves an overall score that is computed for each applicant, meaning that a high score for one test can compensate for a low score on another (Bott, O'Connell, Ramakrishnan & Doverspike, 2007). A common validation procedure involves setting cut scores based on incumbent data and then applying that standard to applicants. As Bott et al. (2007) noted, the higher means in applicant groups could result in systematic bias in the cut scores. Basically, since there is faking in applicant samples, using the cut score determined from incumbent data will result in too many applicants passing the cut score.

To test this assumption, Bott et al. (2007) used incumbent data to set cut scores and then applied that score to compare pass rates for both applicant and incumbent samples. They found that applicants had much larger pass rates (sometimes by a factor of two, depending on how the cut score was calculated) than incumbents. They also compared application of a compensatory versus a multiple-hurdle procedure (cognitive ability was also included) and found that the compensatory procedure lead to the smallest differences in pass rates between the applicant and incumbent samples. Although it is not clear why the compensatory approach resulted in the fewest differences, the study provided strong evidence that using incumbent data to set cut scores can drastically lower the utility of the selection system. This insight has not been lost on other observers, as Mueller-Hanson et al. (2003) suggested that personality tests be used from a select-out versus the traditional select-in perspective. This means that the noncognitive measure's primary purpose would be to weed out the very undesirable candidates rather than to identify the applicants with the highest level of the trait. Of course, this practice can reduce a test's effectiveness by likely allowing more applicants to pass the hurdle, thereby increasing the potential cost of the system. On the other hand, it can increase the confidence in the quality of the decisions made in rejecting candidates from the applicant pool (Mueller-Hanson et al., 2003). Select-out may be a reasonable option under conditions of a high selection ratio (with many positions to fill per applicant) or low cost per test administered (such as unproctored internet testing). Practitioners have to carefully consider and justify how the setting of cut scores matches with the goals and constraints of different selection systems.

Who will Fake, and How Much?

This is potentially the most relevant question to be asked in the faking research, and there is a surprising lack of evidence to answer it. This is certainly not for lack of research effort, but it is more a function of the difficulty of capturing specific faking behavior in applicant settings. For example, many researchers have attempted to explain faking based on other individual difference measures (such as self-monitoring, attention to detail, emotional intelligence, etc.; Hough & Furnham, 2003) which can of course be faked as well.

Fortunately, there is some emerging evidence that patterns exists regarding the proportion of fakers in a given sample. Specifically, converging evidence—though tentative— is indicating that approximately 50% of a sample typically will not fake, with most of the rest being slight fakers, and a select few being extreme fakers. Zickar et al. (2004) used IRT to classify response patterns in personality measures and found that the three class model fit best: extreme fakers (13-38% of respondents), slight fakers (50-60%) and honest respondents. In a within-subjects study, Griffeth et al. (2005) found that 30-50% of applicants elevated their scores compared to later honest ratings. There is also self-reported survey evidence that 65% of people say they would not fake an assessment, with 17% unsure and 17% indicating they would fake (Rees & Metcalfe, 2003). Finally, Robie et al. (2007) had participants think aloud while completing a personality measure (with the goal of winning $100), and out of 12 participants, one was labeled an extreme faker, two were labeled slight fakers, and the rest were labeled as honest respondents. Although more research is certainly needed to verify these three groups of respondents, it is an interesting starting point for researchers and represents a meaningful step in understanding people's motivations while taking noncognitive tests. On the other hand, it is not encouraging for practitioners, because the presence of moderate numbers of fakers—particularly small numbers of extreme fakers—presents significant obstacles for the criterion-related validity of noncognitive tests as well as their ability to select the best applicants (Komar, 2008).

Different Constructs and Fake-ability

As has been hypothesized for quite some time, it is a near universal finding that constructs measured using Likert-type response formats have proven fake-able (Podsakoff & Organ, 1986). This means that when instructions are given to appear as socially desirable as possible, respondents can increase their scores over those given instructions to respond honestly, largely independent of the construct measured. Personality, integrity, and biodata are all fake-able, with effect sizes varying by study and construct (*d*s = .3-1.8; Alliger & Dwight, 2000; McFarland & Ryan, 2000). Even constructs using Likert scales that are given titles to sound ability-based, such as emotional intelligence, are equally fake-able (Grubb & McDaniel, 2007). Although little published evidence exists explicitly comparing applicant and incumbent samples for biodata and integrity, there is little reason to believe that the difference varies much from estimates found with personality (applicants scoring somewhat higher than incumbents, but both much lower than those with fake-good instructions). It is important to mention that just because people can increase their scores does not mean that the test will demonstrate a proportional reduction in validity (Rosse et al., 1998; Mueller-Hanson et al., 2003), but it will have a direct effect on the rank-order of those who would be selected in a top-down manner.

It should also be noted that this discussion does accentuate the difficulty of distinguishing between constructs and measurement methods (Arthur & Villado, 2008), and a thorough analysis of different methods to assess noncognitive constructs is presented later. The important point here is that noncognitive test users cannot switch to different, Likert-type items and claim to be reducing faking. For example, practitioners worried about faking on integrity tests are not advised to switch to conscientiousness personality measures to reduce faking. Counter-intuitive as it sounds, a survey of SHRM members found that many were inclined to do exactly this (Rynes et al., 2002).

Maximal Performance Testing

Cronbach (1984) drew a distinction between typical and maximal performance in assessment. Personality, and most all noncognitive tests, request reports of typical performance in that applicants are asked to describe how they typically behave. Maximal performance tests assess how respondents perform when doing their best. A mathematics test of subtraction is an assessment of maximal performance in that one is motivated to subtract numbers as accurately as one is able, even though one is typically less accurate in daily life subtraction tasks (e.g., balancing a checkbook). In high stakes testing, such as employment testing, people are motivated to do their best, that is, to provide their maximal performance. Thus, faking in personality tests may be viewed as an effort to present oneself in the best way, consistent with the maximal performance demands of high stakes testing.

Situational judgment testing researchers argue that personality and other noncognitive traits can be assessed using maximal performance assessments with situational judgment tests (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Nguyen, 2001; Nguyen, Biderman, & McDaniel, 2005). Such tests present applicants with a work situation (e.g., Everyone in your office has received a new computer but you) and a number of possible responses to the situation:

a) ask your supervisor why you have not received a new computer

b) take a new computer from a co-worker's desk

c) find a new job.

One category of response instructions for situational judgment tests is a knowledge instruction. Knowledge response instructions ask respondents to select the best possible response or to rate the effectiveness of each response. A knowledge response instruction is a maximal performance instruction that should be resistant to faking because both honest and faking respondents are seeking to provide the best possible response. Situational judgment tests with knowledge instructions assess variance associated with personality constructs and do so in manner that is resistant to faking (McDaniel et al., 2007; Nguyen et al., 2005). However, the tests also correlate with cognitive ability

(McDaniel et al., 2007) and it is difficult to form homogeneous subsets of items (McDaniel & Whetzel, 2005) as one can easily do with traditional tests of personality. Still, situational judgment tests appear to measure personality in a faking-resistant manner.

Relation to Cognitive Ability

One of the positive aspects of personality and other noncognitive measures that many researchers and practitioners champion is the negligible adverse impact demonstrated by these tests. Most personality and noncognitive dimensions either demonstrate zero or low relationships with key demographic variables such as ethnicity, gender and age, and the addition of noncognitive tests to a selection battery can buffer the adverse impact associated with using cognitive ability measures in isolation (Bobko et al., 1999). However, a burgeoning research field is suggesting that faked scores may be correlated with personality, and by extension increase adverse impact. Some studies indicate that faking is related to cognitive ability (Grubb & McDaniel, 2007), while others do not (Law, Mobley & Wong, 2002). More importantly, some of the strategies introduced to combat faking, such as forced-choice responses and instructions warning not to fake (both to be discussed in detail below) may increase the correlation between personality and cognitive ability (Vasilopoulos, Cucina, Dyomina, Morewitz & Reilly, 2006; Vasilopoulos et al., 2005). Although it is still a tentative assumption that faked scores are related to cognitive ability, users of noncognitive measures can no longer believe in the blanket assumption that these tests are inherently unrelated to cognitive ability.

Cultural Considerations

Many new to the faking discussion share the observation of Hayes (2007) upon reviewing a recent faking book, who was "surprised by the dearth of transnational research on faking" (p. 513). Even though the Big 5 taxonomy of personality traits has replicated relatively well in cultures around the globe (Paunonen, Jackson, Trzebinski & Forsterling, 1992; Saucier & Goldberg, 2001), most of the faking research has been conducted in English-speaking countries, and there is no guarantee that the construct of faking will be experienced the same way in different cultures (Hough & Furnham, 2003). Constructs such as social desirability and self-monitoring, thought to be conceptually related to faking behavior, have been shown to differ between individualistic and collectivistic cultures (Saucier & Goldberg, 2001). The content of self-report in personality may differ as well—for example, Japanese use more verbs to describe others, while Americans use more nouns (Nisbett & Miyamoto, 2005). Even the definition of traits is likely to differ, as collectivistic cultures are more likely predict and exhibit personality change (Ji, Nisbett & Su, 2001). Although there is some preliminary data that collectivists will fake—for example, Chinese pseudo-applicants faked more on a biodata instrument than those under honest instructions (Law et al., 2002)—more replication and extension of faking studies is necessary in cultures outside of the United States.

POTENTIAL SOLUTIONS TO DEAL WITH FAKING

Social Desirability Scales

Given the previous discussion of the often confusing performance of social desirability scales, such as not accounting for variance in the personality-performance relationship, it should not be surprising that using social desirability scales to correct noncognitive predictor scores has in most cases been unsuccessful. For example, Ellingson, Sackett and Hough (1999) used honest personality scores to represent the true rank order of a number of Army candidates. Then they had the candidates attempt to fake the same measures. After correcting the faked scores based on a social desirability scale, they compared the faked and corrected rank order to the honest condition. Corrections did not result in better hiring decisions.

Corrections also seem to have little effect on criterion-related validity, as adjusting predictor scores based on social desirability does not improve predictive validity in actual data or Monte Carlo simulations (Ones et al., 1996; Schmitt & Oswald, 2006). Finally, social desirability does not appear to account for the negative effect of faking on construct validity. Ellingson, Smith and Sackett (2001) used social desirability scores to separate participants into either faking or non-faking groups and found that the factor structure was the same for both. Taken together, these results clearly indicate that social desirability is not a useful proxy of most applicant faking behavior, and correcting scores based on social desirability does not improve test properties that are thought to be degraded by faking. The one caveat to this belief is that social desirability and lie scales may identify extreme fakers, and validity may be improved by removing these few from the selection pool (Schmitt & Oswald, 2006).

Frame-of-Reference

The rationale behind frame-of-reference (FOR) effects in faking is that parts of the test can be designed to keep test-takers responding based on their behavioral tendencies at work. This means that simple additions to standard noncognitive tests, such as putting the phrase "at work" after every item, or placing a reminder at the top of each page to respond based on behavior at work, can partially prevent respondents from allowing work-irrelevant behaviors (such as behavior at home or in social situations) to be included in their answers. Results have been consistent in their support for FOR in noncognitive testing, at least regarding criterion-related validity. For example, an FOR condition predicted supervisor ratings (r = .31) for a sample of incumbents, but a non-FOR condition did not significantly predict ratings (Hunthausen, Truxillo, Bauer & Hammer, 2003). FOR items can also add incremental validity over both non-FOR items and cognitive ability (Bing, Whanger, Davison & VanHook, 2004). An argument can be made that using FOR in noncognitive testing can change the construct being measured, with a shift from general behavioral tendencies to work-related behavioral tendencies;

however, this appears defensible if the goal is to predict work behaviors. A more daunting potential question remains: although FOR generally improves predictive validity, there is no direct evidence that it reduces faking behavior, and as such the impact of FOR on selection decision quality is unknown. More research is needed to address this issue, but FOR should definitely be seen as a useful strategy in noncognitive test delivery.

Test Instructions

Similar to the rationale applied to the use of FOR, it appears logical that if we want people to respond to our tests in a certain way, we can simply tell them via test instructions. Coaching is one kind of instruction, usually in the form of a vignette or example describing how to approach an item in a socially desirable way. Coaching predictably leads to faking behavior (as evidenced by higher test means; Zickar & Robie, 1999) and is certainly a problem as advice to "beat" noncognitive tests circulates around the internet.

Another popular strategy is to warn test-takers that they will be identified if they fake and removed from the selection pool if they fake (known as a warning of identification and consequences). A meta-analysis by Dwight and Donovan (2003) indicated that warnings generally lower test means over standard instructions (d = .23), although there was considerable variability in the direction and magnitude of effects in the studies included. The use of warnings also presents some challenges. There is evidence that the presence of warnings may increase the correlation between personality and cognitive ability (Vasilopoulos et al., 2005). The other difficulty is an ethical one: given the problems associated with social desirability scales in identifying actual fakers, it is somewhat dishonest to warn test-takers that fakers can indeed be identified (Zickar & Robie, 1999). Thus, warnings are admittedly an imperfect method for combating faking, and more research is needed to ultimately determine the extent of their utility. New instructions may be written to alter test-taker motivation or to change the extent to which a noncognitive test measures typical or maximal performance (McDaniel et al., 2007).

Other Reports

Personality and other noncognitive constructs are often evaluated for selection purposes through ratings of others, including interviews and assessment centers. Approximately 35% of interviews explicitly measure noncognitive constructs such as personality and social skills, according to meta-analytic evidence (Huffcutt, Conway, Roth & Stone, 2001). Similarly, many common assessment center dimensions involve noncognitive aspects, including communication and influencing others (Arthur, Day, McNelly & Edens, 2003). Faking and impression management research has traditionally avoided

intentional distortion in these contexts, but it is logical that those who would fake in a personality inventory would also fake in an interview setting. Best practice with these methods involves structuring ratings and dimensions and training interviewers and assessors, which would potentially reduce the impact of applicant faking on ratings; however, more research is needed to determine just how problematic faking in interviews and assessment centers may be.

Situational Judgment Tests with Knowledge Instructions

Although most of the faking research has concentrated on measurement methods that involve traditional Likert scales, many other common methods used in personnel selection also measure personality and noncognitive constructs, including situational judgment tests and forced-choice measures. Situational judgment tests (SJTs) are construct heterogeneous tests that measure both cognitive ability and personality constructs (McDaniel et al., 2007). There is evidence that SJTs with knowledge instructions are resistant to faking (Nguyen et al, 2005) because they are maximal performance assessments (McDaniel et al., 2007). Although replication and additional research are needed, to date, situational judgment tests with knowledge instructions appear to be the sole available method of assessing personality in a manner that is faking resistant.

Forced-choice Tests

Forced-choice measures differ from Likert-type scales because they take equally desirable items (desirability usually determined by independent raters) and force the respondent to choose. Moving to forced-choice requires a cost, which involves abandoning the interval-level scale of measurement and the clearer construct scaling that Likert measures offer. Whether the benefits of forced-choice formats, such as potentially reducing faking, justify these costs is questionable. The effect of forced-choice on test means is unclear, as some studies show higher means of forced-choice compared with Likert measures (Vasilopoulos, Cucina, Dyomina, Morewitz & Reilly, 2006) and others indicate lower means (Heggestad, Morrison, Reeve & McCloy, 2006). Heggestad et al. (2006) also provided a useful analysis of the effect of forced-choice on selection decisions. They used items in both a forced-choice and Likert format under pseudo-applicant instructions (pretend you are applying for a job). They then compared the rank-order produced by both of these tests to an honest condition using a different personality measure. Results showed few differences in the rank-orders between the measures, offering preliminary evidence that forced-choice does not improve selection decisions. Taken together, these results suggest that forced-choice tests do not necessarily reduce faking, and the statistical and conceptual limitations associated with their use probably does not justify replacing traditional noncognitive test formats.

RECOMMENDATIONS FOR PRACTICE

Given the complexity of faking in selection settings, it should not be surprising that there are currently few universal recommendations for best practice. However, research has provided numerous insights toward effectively dealing with (if not preventing) faking, depending on the selection system and the test properties on which there is the most concern about faking having negative effects. This closing section provides a number of considerations that practitioners should take into account when attempting to understand and deal with faking in noncognitive tests.

Avoid Corrections

Little evidence exists that social desirability scales or lie scales can identify faking. Many tests include lie scales with instructions about how to correct scores based on lie scales, with the justification that corrections will improve test validity (Rothstein & Goffin, 2006). Unfortunately, there is no evidence to support this assertion, rendering corrections a largely undefensible strategy. If a few extreme fakers are apparent, based on a social desirability scale or other information, it may be appropriate to remove them from the applicant pool; however, it is necessary to always be on guard for false positives (identifying one as a faker who in reality responded honestly). Use of corrections can vary based on whether the context is for research or high-stakes testing, as 56% of researchers use corrections (Goffin & Christiansen, 2003), but assessors are much less like to use them (M = 1.18 on a 5-point scale; Robie, Tuzinski & Bly, 2006).

Specify How Noncognitive Measures Fit the Goals of the Selection System

Given the consistent effect of faking on test means, faking will affect cut scores and who is selected in both compensatory and multiple hurdle systems (Bott et al., 2007). Cut scores may have to be adjusted upward if they are set based on incumbent scores. The select-out strategy is also an option, particularly if the selection ratio is high. Finally, the quality of measurement desired is also critically important. Because faking does negatively affect construct validity, multiple methods may be needed to triangulate participants' true standing on the trait (Hough & Furnham, 2003). In addition, quality construct measurement depends on careful linkages being made between job requirements and constructs measured, which may often deviate from standard Big 5 personality dimensions (Hogan & Holland, 2003).

Consider Test Means, Criterion-related Validity, and Selection Decisions

Evidence of faking can show up in any of these indicators, and the effect of faking on one is not always proportional to the others. It is important to consider the average increase in scores due to faking (means), the effect that faking has on a test's ability to predict important outcomes (criterion-related validity) and how much of an impact faking

has on who is hired (selection decisions). Practitioners cannot make the mistake that some researchers have made, which is to say that faking does not matter if a test demonstrates criterion-related validity.

Consider Test-taker Motivation

In applied settings, it is assumed that applicants are motivated to score well in order to increase their chances of being hired. However, applicants can be motivated by additional factors as well. Much research is being devoted to organizational fit and finding that applicants are quite motivated to find a job in which they experience high levels of fit (Arthur, Bell, Villado & Doverspike, 2006). This can be tricky to utilize with regard to faking. On one hand, if applicants are given information about the job to which they are applying, they can fake their scores toward that stereotype (Mahar, Cologon & Duck, 1995). On the other hand, if applicants are informed about the potential consequences of poor fit—which faking could realistically lead to during the placement phase—they may be motivated to respond more honestly, and initial research indicates that this may be true (Nordlund & Snell, 2006). Regarding test-taker motivation, research has less insight to offer, and practitioners are encouraged to at least consider the test from the applicant's motivational perspective.

CONCLUSION

Noncognitive tests can be faked, and they are faked in high-stakes settings. Faking does not eliminate these tests' utility, but it does call for more intensive analyses of the extent of the problem than is often deployed. There is no method (easy or difficult) to eliminate faking, but research and practice does offer much insight into understanding and dealing with the phenomenon.

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.

Alliger, G.M, Lilienfeld, S.O. & Mitchell, K.E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7,* 32-39.

Arthur, W., Bell, S.T., Villado, A.J. & Doverspike, D. (2006). The use of person-organization fit on employment decision-making. *Journal of Applied Psychology, 91,*786-801.

Arthur, W., Day, E.A., McNelly, T.L. & Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Journal of Applied Psychology, 56,* 135-154.

Arthur, W. & Villado, A.J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435-442.

Barrick, M.R., Mount, M.K. & Judge, T.A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9-30.

Beaty, J.C., Cleveland, J.N. & Murphy, K.R. (2001). The relation between personality and contextual performance in "strong" versus "weak" situations. *Human Performance, 14,* 125-148.

Bing, M.N., Whanger, J.C., Davison, H.K. & VanHook, J.B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89,* 150-157.

Birkeland, S.A., Manson, T.M., Kisamore, J.L., Brannick, M.T. & Smith, M.A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14,* 317-330.

Bobko, P., Roth, P.L. & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561-589.

Bott, J.P., O'Connell, M.S., Ramakrisnan, M. & Doverspike, D. (2007). Practical limitations in making decisions regarding the distribution of applicant personality test scores based on incumbent data. *Journal of Business Psychology, 22,* 123-134.

Cable, D.M. & Judge, T.A. (1996). Person-organization fit, job choice decisions, and organizational entry. *Organizational Behavior and Human Decision Processes, 67,* 294-311.

Cellar, D.F., Miller, M.L., Doverspike, D.D. & Klawsky, J.D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five-factor model. *Journal of Applied Psychology, 81,* 694-704.

Cronbach, L.J. (1984). *Essentials of Psychological Testing.* (4th ed.) New York: Harper & Row.

Douglas, E.F., McDaniel, M.A. & Snell, A.F. (1996). The validity of non-cognitive measures decays as applicants fake. *Proceedings of the Academy of Management,* August.

Dwight, S.A., & Donovan, J.J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16,* 1-23.

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

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,* 215-224.

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.

Griffin, B., Hesketh, B. & Grayson, D. (2004). Applicants faking good: Evidence of item bias in the NEO PI-R. *Personality and Individual Differences, 36,* 1545-1558.

Griffith, R.L., Chmielowski, T. & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341-355.

Grubb, W.L. & McDaniel, M.A. (2007). The fakability of Bar-On's emotional quotient inventory short form: Catch me if you can. *Human Performance, 20,* 43-59.

Hayes, T.L. (2007). Review of *A Closer Examination of Applicant Faking Behavior. Personnel Psychology, 60,* 511-514.

Heggestad, E.D. Morrison, M. Reeve, C.L. & McCloy (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91,* 9-24.

Hogan, J. & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88,* 100-112.

Hough, L.M. & Furnham, A. (2003). Use of personality variables in work settings. In W.C. Borman, D.R. Ilgen & R.J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology.* Hoboken: John Wiley & Sons, Inc.

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

Hough, L.M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 5,* 139-155.

Huffcutt, A.I., Conway, J.M., Roth, P.L. & Stone, N.J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897-913.

Hunthausen, J.M., Truxillo, D.M., Bauer, T.N. & Hammer, L.B. (2003). A field study of frame-of-reference effect on personality test validity. *Journal of Applied Psychology, 88,* 545-551.

Ji, L., Nisbett, R.E. & Su, Y. (2001). Culture, change and prediction. *Psychological Science, 12,* 450-456.

Komar, S., Brown, D.J., Komar, J.A. & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93,* 140-154.

Langleben, D.D., Loughead, J.W., Bilker, W.B., Ruparel, K, Childress, A.R., Busch, S.I. & Gur, R.C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping, 26,* 262-272.

Law, K.S., Mobley, W.H. & Wong, C. (2002). Impression management and faking in biodata scores among Chinese job-seekers. *Asia Pacific Journal of Management, 19,* 541-556.

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

McDaniel, M.A., Hartman, N.S., Whetzel, D.L. & Grubb, W.L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

McDaniel, M.A. & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

McFarland, L.A. & Ryan, A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812-821.

Mueller-Hanson, R., 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.

Murphy, K.R. & Dzieweczynski, J.L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance, 18,* 343-357.

Nguyen, N.T., Biderman, M.D. & McDaniel, M.A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13,* 250-260.

Nisbett, R.E. & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences, 9,* 467-473.

Nordlund, M. & Snell, A.F. (April, 2006). *Improving criterion-related validities with instructions to applicants: A good idea that actually worked.* Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Dallas, TX.

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

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

Paunonen, S.V., Jackson, D.N., Trzebinski, J., Fosterling, F. (1992). Personality structure across cultures: A multimethod evaluation. *Journal of Personality and Social Psychology, 62,* 447-456.

Peters, H. & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65 (1),* 70-89.

Podsakoff, P.M. & Organ, D.W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12,* 531-544.

Rees, C.J. & Metcalfe, B. (2003). The faking of personality questionnaire results: Who's kidding whom? *Journal of Managerial Psychology, 18,* 156-165.

Robie, C., Brown, D.J. & Beaty, J.C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21,* 489-510.

Robie, C., Tuzinski, K.A. & Bly, P.R. (2006). A survey of assessor beliefs and practices related to faking. *Journal of Managerial Psychology, 21,* 669-681.

Roese, N.J. & Jamieson, D.W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363-375.

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.

Rothstein, M.G. & Goffin, R.D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16,* 155-180.

Rynes, S.L., Colbert, A.E. & Brown, K.G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management, 41,* 149-174.

Saucier, G. & Goldberg, L.R. (2001). Lexical studies of indigenous personality factors: Premises, products and prospects. *Journal of Personality, 69,* 847-879.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research. *Psychological Bulletin, 124,* 262-274.

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

Snell, A.F., Sydell, E.J. & Lueke, S.B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

Vasilopoulos, N.L. Cucina, J.M., Dyomina, N.V. & Morewitz, C.L. (2006). Forced-choice personality tests: A measure personality and cognitive ability. *Human Performance, 19,* 175-199.

Vasilopoulos, N.L., Cucina, J.M. & McElreath, J.M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90,* 306-322.

Vasilopoulos, N.L., Reilly, R.R. & Leaman, J.A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85,* 50-64.

Viswesvaran, C. & Ones, D.S. (1999). Meta-analyses of faking estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197-210.

Zickar, M.J. & Robie, C. (1999). Modeling faking good on personality items. *Journal of Applied Psychology, 84,* 551-563.

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.