

The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution?

Douglas N. Jackson and Victor R. Wroblewski

*Department of Psychology
The University of Western Ontario, Canada*

Michael C. Ashton

*Department of Psychology
Brock University, Ontario, Canada*

We evaluated the effects of faking on mean scores and correlations with self-reported counterproductive behavior of integrity-related personality items administered in single-stimulus and forced-choice formats. In laboratory studies, we found that respondents instructed to respond as if applying for a job scored higher than when given standard or “straight-take” instructions. The size of the mean shift was nearly a full standard deviation for the single-stimulus integrity measure, but less than one third of a standard deviation for the same items presented in a forced-choice format. The correlation between the personality questionnaire administered in the single-stimulus condition and self-reported workplace delinquency was much lower in the job applicant condition than in the straight-take condition, whereas the same items administered in the forced-choice condition maintained their substantial correlations with workplace delinquency.

Paper-and-pencil integrity testing was considered in the early part of the 20th century when Hartshorne and May (1928) investigated the use of such tests as predictors of dishonesty in children. Although studied since World War II (Betts, 1947; Houtchens & Betts, 1947), the first commercially published integrity test for job applicants did not appear until the work of Ash (1971) and his development of the

Reid Report. Integrity tests have been classified into two types: *overt tests*, in which respondents answer direct questions regarding attitudes toward theft and past transgressions, and *personality-oriented tests*, in which respondents answer questions to assess personality traits associated empirically with integrity and good job performance (Sackett, Burris, & Callaghan, 1989).

Recent literature reviews have described extensive use of integrity testing in personnel selection (Sackett et al., 1989; Sackett & Harris, 1984; Sackett & Wanek, 1996). According to some reviews, currently available paper-and-pencil questionnaires of integrity have demonstrated only modest validity in predicting the delinquency of potential employees (Sackett et al., 1989; Sackett & Harris, 1984), but in a recent meta-analysis, Ones, Viswesvaran, and Schmidt (1993) reported relatively high corrected validities, averaging .41 for personality-oriented integrity tests predicting job performance.

FAKING ON INTEGRITY TESTS

Hartshorne and May (1928) mentioned the prevalence of faking and the type of distortion engaged in by children in test-taking situations. Sackett et al. (1989) described three types of faking in job application testing: faking good, faking bad, and faking a specific job role. *Faking good* involves responding to the questionnaire so as to portray oneself in a generally positive light, whereas *faking bad* represents attempting to present a negative impression. The third type of faking is *faking a specific job role*, such as an air force captain or a psychiatric nurse (Mahar, Cologon, & Duck, 1995). Faking good has particular relevance to personnel selection and is the focus of this study. It has long been recognized that faking occurs when job applicants are motivated to make a good impression, but its impact on validity remains controversial.

The susceptibility of workplace measures to faking good has been widely researched (Dunnette, McCartney, Carlson, & Kirchner, 1962; Furnham, 1990; see Sackett et al., 1989; Sackett & Harris, 1984; Sackett & Wanek, 1996, for reviews). Individuals have been shown generally to be able to fake good if the situation motivates them to give a favorable impression (Furnham, 1986). Faking good has been demonstrated to occur with a wide variety of employment selection tests (see Furnham, 1986, 1990).

DETECTING AND CORRECTING FOR FAKING

Dealing with respondents who distort responses on personality tests has traditionally involved one of two methods (Sackett & Decker, 1979): disqualification of respondents by using lie scales, or correction of scores for faking. Lie scales have

been used to detect those who are motivated to present an unrealistically favorable impression, but the problem with disqualification is that it will eliminate without further consideration some potentially qualified candidates who might not, in fact, be consciously lying. Correction for faking has been attempted for some published personality tests with limited success (Christiansen, Goffin, Johnston, & Rothstein, 1994; Ellingson, Sackett, & Hough, 1999; Hough, 1998; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). Attempts to correct for faking are based on a suppressor variable rationale, but because suppressor variables can have only negligible effects even under ideal conditions. At low to moderate levels of validity, attempts to correct scores for faking can have little impact on validity and are, in fact, misguided and futile (Conger & Jackson, 1972).

Other approaches to the problem of motivated distortion have been proposed. Nederhof (1985) discussed a comprehensive list of possible solutions, including their perceived benefits and drawbacks. One approach that is of particular interest in this research is the use of a forced-choice question format. The recommended method for creating forced-choice items involves matching items on statistical properties, such as the social desirability of items. Notwithstanding the number of published studies on the forced-choice technique (e.g., Christiansen, Edelstein, & Fleming, 1998; Dunnette et al., 1962; Graham, 1958; Hicks, 1970; Villanova, Benardin, Johnson, & Dahmus, 1994; Waters, 1965; Zavala, 1965), its relative lack of popularity might have stemmed from the technical problems associated with scale construction, including the difficulties of finding items to pair that appear to be equally valid but in fact are not.

Even though the approaches summarized by Nederhof (1985) are more than a decade old, research on the detection and correction of faking on job selection questionnaires continues. Some might argue that the whole issue is a red herring with no need for a solution (Barrick & Mount, 1996; Ones, Viswesvaran, & Reiss, 1996). Others would disagree. We take up this issue from the standpoints of some technical issues and empirical data. Given that no conclusive solution to the problem of faking has yet been demonstrated, the purpose of this study was to investigate a test format that might deter or reduce faking.

FORCED-CHOICE QUESTIONNAIRES

Forced-choice questionnaires represent one possible method of minimizing the problem of faking in employment testing by making the task of responding desirably much more difficult. If respondents are motivated to make the best possible impression, being forced to choose between items similar in perceived relevance to the job tends to reduce this type of impression management.

The use of forced-choice methodology in personality assessment has been known since the 1940s through the unpublished work of Horst and Wherry (see

Zavala, 1965). It was partially through Horst's influence that Edwards (1954) developed the Edwards Personal Preference Schedule, a standardized forced-choice personality questionnaire.

Christiansen et al. (1998) recently argued that criticisms against forced-choice formats (see Nederhof, 1985) might be overstated. They administered personality-adjective scales in single-stimulus (Likert scale) format and in binary forced-choice format to 350 participants. Half of the participants completed the scales under normal instructions, and the other half under instructions to respond as if applying for a job in sales. Christiansen et al. found that the difference between the means of the normal instruction group and the applicant instruction group was smaller for the forced-choice scales than for the single-stimulus scales. They also investigated the correlations of the scales with a personality questionnaire administered under normal instructions. For the job applicant conditions, the correlation of the single-stimulus scales with the criterion personality scales was much lower than the correlations using the normal instructions ("straight-take") sample, but the correlation for the forced-choice scales was almost as high for the applicant condition as it was for the straight-take condition, even though the item formats differed between the forced-choice condition and the comparison personality scales.

In this research we took steps to minimize the problems associated with forced-choice test formats. First, we matched personality items with unscored items and those from several different trait scales that were considered irrelevant to our counterproductive behavior criterion, so that the issue of the loss of degrees of freedom due to ipsative total scores was eliminated. Also, many of the items within the other scales included apparently job-relevant content, which made those items attractive distractors for job applicants who try to fake. Finally, we used the tetrad (Dunnette et al., 1962) or dichotomous quartet method of combining items. The dichotomous quartet format matches two socially desirable items with two undesirable items. Respondents are instructed to select, from each quartet, the item that is most characteristic of their behavior and the item that is least characteristic of their behavior. Dunnette et al. (1962) and others (Jackson & Payne, 1963) recommended such a procedure to avoid forcing respondents to endorse one of two undesirable alternatives.

VALIDATING PERSONALITY-BASED MEASURES OF INTEGRITY

Sackett et al. (1989) studied many examples of possible methods for validating integrity tests and subsequently grouped them into five categories: polygraph comparisons, admissions, future behavior, shrinkage reduction, and contrasted groups. Of the five, the one of primary interest in this study is correlating the test with

self-reported behavior, which includes past theft and other counterproductive activities. Accordingly, in this study, scores on a personality-based dependability scale are correlated with self-reported measures of on-the-job delinquency as an indication of the respective correlations with different item formats.

The major purpose of this study is thus to determine whether or not forced-choice items equated on psychometric properties reduce faking under job applicant conditions, and whether or not such items maintain correlations with reported counterproductive behavior under straight-take conditions. We addressed these questions by administering a personality-based measure of integrity—the Dependability scale of the Employee Selection Questionnaire (ESQ; Jackson, in press)—under standard or straight-take instructions and under job application instructions. The ESQ is a personality questionnaire measuring six factors, derived from an original pool of more than 3,000 items. Only the dependability factor was of interest for this study. One group of respondents completed single-stimulus versions of the ESQ items under both instructional conditions, and the other group completed a forced-choice version of the same items, also under both instructional conditions. We assessed the performance of the dependability items of both formats under both conditions to evaluate the relative merits of the forced-choice versus single-stimulus formats with regard to confronting the problem of faking. We hypothesized that the forced-choice scale would be less susceptible to faking, in terms of both mean score shifts and decrements in correlation with self-reported counterproductive behavior, than would the single-stimulus scale.

STUDY 1: DOES FAKING REDUCE PERSONALITY SCALE CORRELATIONS WITH SELF-REPORTED COUNTERPRODUCTIVE BEHAVIOR?

The first study involved the administration of dependability items to research participants in a standard single-stimulus format. This was intended to provide a basis for comparison with the same items administered under forced-choice instructions in the second study.

Method

Participants

Participants for this study were 84 undergraduate students (39 men and 45 women) with a median age of 21 years, all of whom reported that they had an employment history.

Materials

ESQ: Single-Stimulus Version. The traits measured by the ESQ include dependability (integrity or probity, prudence, and responsibility), industriousness (a concern for excellence, and seriousness, and perseverance at work), methodicalness (careful attention to detail), agreeableness (desire to please others), extraversion (preference for social activity), and independence (willingness to make decisions without assistance). The Dependability scale was developed using items drawn from the Responsibility and Risk Taking scales of the Jackson Personality Inventory –Revised (JPI–R; Jackson, 1994). The JPI–R was developed using a large item pool (more than 2,000 items) and elaborate statistical procedures. The Responsibility and Risk Taking scales have been found to be substantially correlated with other integrity measures (Mikulay & Goffin, 1998), with illegal substance use, relevant peer ratings, and experimental and self-reported measures of counterproductive behavior (Ashton, 1998; Mikulay & Goffin, 1998). A set of 120 additional items from the integrity domain were written and selected for the ESQ Dependability scale on the basis of empirical data bearing on internal consistency, self-reported counterproductive work behavior, and the differentiation of incarcerated and nonincarcerated persons. The Dependability scale used in this study contained 17 items administered with a 7-point response scale, ranging from 1(*strongly disagree*) to 4(*neutral*) to 7(*strongly agree*).

Workplace Behavior Questionnaire. The criterion measure used in this study is the Workplace Behavior Questionnaire (Ashton, 1998), which contains eight items related to workplace delinquency. Response options for items from this questionnaire used a 6-point scale, with the higher numbers representing a greater frequency of delinquent acts. Ashton (1998) suggested that response style variance in delinquency scales may be reduced by the use of response categories that contain specific numerical estimates rather than subjective judgments of frequency, (e.g., rarely, sometimes, often).

Participants were asked to answer the questionnaire frankly, giving their best estimate of the total number of various delinquent behaviors they had committed. The types of work-related delinquent behaviors measured in this questionnaire included “absenteeism, lateness, alcohol use or influence, safety violations, goldbricking (i.e., avoiding work during paid time), theft, giving ... free goods or services to friends or relatives, and vandalism or sabotage” (Ashton, 1998, p. 292). To ensure frank and accurate responses, participants were assured of complete confidentiality and instructed not to place their names on the questionnaire.

Procedure

Respondents each completed the personality questionnaire under two instructional conditions—first, the straight-take condition, and then the job application condi-

tion. In the straight-take condition, participants were instructed to answer as frankly as possible; this condition was intended to serve as a basis for comparison with the job applicant trial. In the subsequent job applicant trial they were asked to answer the items as if they were trying to make a good impression to better their chances of gaining a job they really wanted. Both the straight-take and the job applicant trials preceded the modified version of the Workplace Behavior Questionnaire. Norman (1967) recommended that, in repeated-measures studies using personality tests, the experimental condition should follow the control condition to limit carryover effects.

Results

Work Experience

The median number of hours worked per week by respondents was 10 to 14 hr during the school year and 30 to 39 hr during the summer.

Dependability Scale

Respondents produced higher mean scores for the single-stimulus Dependability items under the job application condition than under the straight-take condition ($M = 82.5$, $SD = 13.9$ vs. $M = 70.9$, $SD = 12.4$), $t(83) = 8.30$, $p < .001$. In terms of the straight-take distribution, participants scored .95 SD higher under the job application condition than under the straight-take condition. This result indicated that scores were substantially affected by instructing the participants to respond as if applying for a job. In terms of the direction of score shifts from straight-take to job application conditions, 83% of participants scored higher under the job application condition than under the straight-take condition, 5% of the participants had no difference in scores, and only 12% scored higher under the straight-take condition than under the job application condition.

Internal consistency reliability coefficients (Cronbach's α) for the Dependability scale were .70 and .79 under the straight-take and job application conditions, respectively. The correlation between the Dependability scores of the two conditions was .53.

Delinquency Criteria

The eight-item Workplace Behavior Questionnaire had a mean score of 16.7 ($SD = 6.6$). Scores ranged from 8 to 40 on this scale, on which the possible range was 8 to 48. The reliability coefficient was .75, aggregating across all types of counterproductive work behavior.

TABLE 1
Reliability and Correlations With Self-Reported Counterproductive Behavior of
Single-Stimulus Dependability Scale

<i>Condition</i>	<i>Reliability</i>	<i>Correlation</i>
Straight-take	.70	.48*
Job-application	.79	.18

Note. $N = 84$. Correlations between the dependability and counterproductive behavior have been reflected.

* $p < .001$.

Correlations

Table 1 shows the correlation of the Dependability scale scores with the Workplace Behavior Questionnaire. Lack of dependability was significantly correlated with self-reported past workplace counterproductivity for the straight-take condition ($r = .48, p < .001$), but not for the job-application condition ($r = .18, ns$). The difference in the correlations between the two conditions was significant ($z = 3.00, p < .001$).

STUDY 2: DOES A FORCED-CHOICE FORMAT REDUCE FAKING?

In the second study, the same Dependability items used in the first study were administered, but in a forced-choice format using quartets of items. This was intended to provide a comparison of the fakeability of the single-stimulus and forced-choice item formats.

Method

Participants

Participants for this study were 106 undergraduate students (36 men and 70 women) with a median age of 19 years, all of whom reported that they had an employment history.

Materials

ESQ: Forced-Choice. Personality measures consisted of the same item stems as in Study 1. The method employed to develop forced-choice items was first

to identify, for each item, statistical indexes of attractiveness or desirability. Two such indexes were employed: (a) the actual mean item endorsement proportions from a sample of 3,760 job applicants, and (b) desirability scale values based on the mean rating of 20 judges who estimated, using a 7-point scale, the desirability of the item in the context of a job applicant motivated to make a good impression. Items were placed in dyads using a computer program designed to match items from different scales with very similar endorsement frequencies and desirability scale values. High-desirability item dyads were then combined randomly with the item dyads of low desirability to form quartets of items. These item quartets were then randomly distributed throughout the test with the restriction that no two quartets representing the same combination of traits would follow each other.

Each respondent was instructed to choose from each quartet the item member most characteristic and the item member least characteristic of him or her. Item members were scored +1 if the participant answered that it was most characteristic and -1 if the participant answered that it was least characteristic. Item members not selected as being most or least characteristic were scored as 0 for their scale. Thus, each quartet would contain one +1 score, one -1 score, and two 0 scores, with reverse-keyed items being appropriately reflected.

Workplace Behavior Questionnaire. The criterion items and the instructions were the same as those in Study 1. The participants were assured of complete confidentiality and instructed not to place their names on the questionnaires.

Procedure

Respondents completed the ESQ under straight-take and job application sets consecutively, and then completed the Workplace Behavior Questionnaire. The instructions and order of the questionnaires were the same as in Study 1.

Results

Work Experience

The median number of hours worked per week by respondents was 10 to 14 hr during the school year and 30 to 39 hr during the summer.

Dependability Scale

Although mean scores for Dependability were higher under job applicant conditions for the forced-choice format ($M = 2.9, SD = 3.6$ vs. $M = 1.5, SD = 4.4$), $t(105) =$

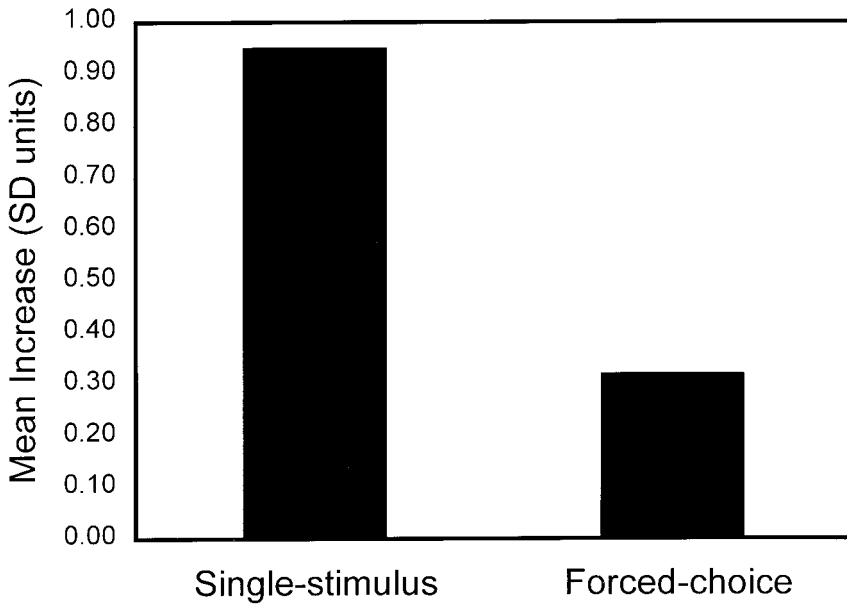


FIGURE 1 Mean increases in single-stimulus and forced-choice Dependability scale scores from straight-take to job applicant conditions.

$-3.39, p < .01$, these differences were much smaller than those obtained for the single-stimulus format used in Study 1. In terms of the straight-take distribution, the average participant scored $.32 SD$ higher under the job application condition than under the straight-take condition, a value only about one third as large as that found for the single-stimulus scales in Study 1 (Figure 1). In terms of the direction of score shifts for individual respondents, 62% of participants scored higher on the Dependability scale under the job application condition than under the straight-take condition. Ten percent of the participants had no difference in scores, and 28% scored higher under the straight-take condition than under the job application condition.

Reliability coefficients (Cronbach's α) were $.69$ and $.55$ for the straight-take and job application conditions, respectively. Despite the moderate reliabilities, the correlation between the Dependability scale scores in the two conditions was $.56$. (Of course, lengthening the 17-item forced-choice Dependability scale would increase reliability.)

Delinquency Criteria

The Workplace Behavior Questionnaire had a mean score of $15.9 (SD = 6.9)$. The reliability coefficient was $.76$.

Correlations of Forced-Choice Dependability Scale

Table 2 presents correlations between low dependability and self-reported past counterproductive behaviors in the workplace. Lack of dependability correlated with self-reported past counterproductive behaviors in the workplace (see Table 2). Workplace Behavior Questionnaire scores were significantly negatively correlated with both the straight-take ($r = .41, p < .001$) and job application ($r = .36, p < .001$) measures of (reflected) dependability. The difference in the correlations between the two conditions was not significant ($z = 0.60, ns$).

Thus, the forced-choice Dependability scale's correlations were not appreciably different for the job application and straight-take conditions. The criterion cor-

TABLE 2
Reliability and Correlations With Self-Reported Counterproductive Behaviors of Forced-Choice Dependability Scale

Condition	Reliability	Correlation
Straight-take	.69	.41*
Job-application	.55	.36*

Note. $N = 106$. Correlations between the dependability and counterproductive behavior have been reflected.

* $p < .001$.

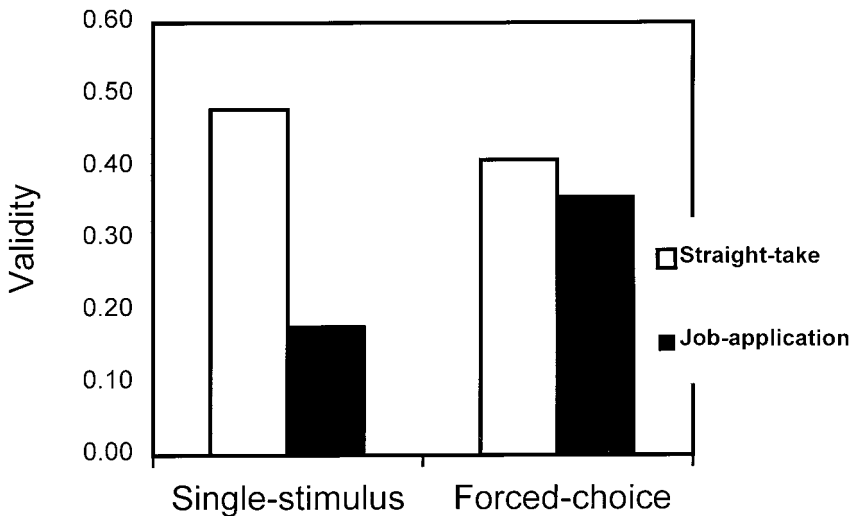


FIGURE 2 Correlations of single-stimulus and forced-choice Dependability scales with a workplace delinquency criterion under straight-take and job applicant conditions.

relations remained statistically significant and moderately strong even under job application instructions. This is a markedly different result from that obtained in Study 1, in which the single-stimulus Dependability scale showed much lower correlations in the job application condition than in the straight-take condition (see Figure 2).

DISCUSSION

Two important differences were found between the single-stimulus and the forced-choice administrations of the Dependability items. First, single-stimulus Dependability scores showed a considerably larger mean difference between the straight-take and the job application conditions than did the forced-choice Dependability measure. Second, for the single-stimulus administration of the items, the Dependability scale had considerably lower criterion correlations in the job application condition than in the straight-take condition, but under forced-choice administration, the correlations were similar across the two conditions. Both of these results—the smaller difference in the mean shifts and the minimal decrease in criterion correlations—indicate a lower susceptibility to faking for the forced-choice format than for the single-stimulus format. The results of our study thus support the findings of Christiansen et al. (1998), and even show a somewhat greater superiority of forced-choice scales over single-stimulus scales than found in their study. Our results extend the findings of those of Christiansen et al. to a quartet version of the forced-choice format, to personality-based integrity tests, and, most important, to workplace delinquency.

Faking poses major problems for the use of personality tests in personnel selection, and there are accordingly important benefits to reducing faking. As we have demonstrated, respondents under simulated job applicant conditions using traditional single-stimulus items yield scores that are unrealistically high and are poorly correlated with reported workplace delinquency. Far from representing a red herring (Ones et al., 1996), faking can have dramatic effects on hiring decisions (see also Rossé, Stecher, Miller, & Levin, 1998) and on correlations with reported counterproductive behavior.

Our results are consistent with those of Douglas, McDaniel, and Snell (1996), who administered personality and biodata questionnaires under faking and straight-take conditions to randomly assigned samples of university students, for whom job performance data from previous employers was also obtained. Douglas et al. found that respondents who faked showed substantially higher personality and biodata scores than did nonfakers. Consistent with our findings, scale reliabilities were higher in the faking condition. Furthermore, Douglas et al. found that noncognitive measures administered under faking conditions showed substantial decrement in validity when compared with those administered under

straight-take conditions. Unlike the instructions in the study reported here, which asked respondents to answer the personality questionnaire as if they were applying for a job that they really wanted, Douglas et al. instructed respondents to “make as favorable impression as possible ... we are not interested in your true opinion, rather we would like you to identify the most socially desirable response for each question” (p. 3). From the substantial body of research on social desirability judgments of personality items, it is well known that such instructions yield a strong group consensus and level individual differences. Accordingly, it would be useful to replicate the Douglas et al. study with less extreme instructions more typical of those given to job candidates. Nevertheless, the Douglas et al. results are important because they bear on the most important reason for administering noncognitive measures, namely, their impact on the validity of job selection decisions.

In addition to improved correlations, a second benefit to reducing the role of faking in job selection testing is an ethical one. Many potential users of employee tests decline to utilize them because they believe that it is manifestly unfair to make job offers to fakers and deny job opportunities to more honest respondents. Rossé et al. (1998) found statistical support that faking among applicants can affect who is hired.

A third benefit to reducing the role of faking, one not addressed directly in our studies but important, is that it permits a more differentiated basis for the use of personality scales in hiring (Sackett & Wanek, 1996). Thus, if a particular personality-oriented job analysis revealed that a certain job required persons with a pattern of distinct traits, a measure that contained only one factor could not provide all of the relevant information for a decision. The multiple correlation is at a maximum when all valid predictors are minimally intercorrelated. A one-factor solution, like that obtained by Ellingson et al. (1999) when respondents faked, will not yield the same level of predictive accuracy as when there are a number of distinct valid predictors (Nunnally & Bernstein, 1994; Paunonen, Rothstein, & Jackson, 1999).

Our results are based on a single measure of personality as it relates to self-reported measures of counterproductive behavior. Sometimes, however, profile measures of personality are used for personnel screening. Usually particular scales on a profile measure are aggregated in a weighted or unweighted composite based on previous criterion validity studies and a cut score is applied. In such a circumstance the profile measure is susceptible to faking to the extent that the selected scales individually are susceptible to faking. If a linear composite is employed, one might view the linear composite as a single predictor scale and evaluate the effects of faking on score levels and on validity. Given this interpretation, the differences between single measures and profile measures from an analytical perspective are minimal; they are very similar in terms of their implications for faking. There is, however, a widespread belief among practitioners that the use of heterogeneous item and scale content in a personality measure makes faking more difficult. This

is plausible, but we know of no studies in selection contexts that have investigated this issue.

Another potential approach to selection, one that departs from the traditional linear regression model, is a *profile-matching* procedure. This involves identifying the personality profile of a set of successful or all incumbents and matching the profile of an applicant to that of the incumbents. Although this procedure is used widely in the area of vocational interest measurement, with substantial evidence for validity in terms of classification studies, it has rarely been used in personnel selection, and its validity has been even more rarely reported in the published literature. The extent to which such an alternative use of personality profiles might be less susceptible to faking, and whether or not such a measure might retain its validity under job selection administration conditions, remain open questions. There are data, however, indicating that judges are quite accurate in identifying the profile of different occupational groups in vocational interest measurement (Jackson, 2000, pp. 114–118). This raises the possibility that applicants could simulate the profile of successful incumbents, but the issue is not settled.

It should be recognized that persons simulating the role of job applicants will not necessarily yield the same results as will real job applicants. Real job applicants might fake more or less than simulators and might produce higher or lower correlations with counterproductive behavior. Whether the motivational conditions for selection testing differ greatly from those of research testing on incumbents also is an open question. Indeed, the conditions under which incumbent research takes place vary widely, from, for example, questionnaires administered by external researchers with anonymity for testees, to those administered by organizational personnel in which respondents are identified. A delineation of the impact on scores and validity of these varying conditions would be useful in deciding on how much relevance meta-analyses of incumbent personality test validity have for inferring the validity of tests used for personnel selection.

Our results offer a perspective on findings from a series of meta-analyses of personality and job performance (e.g., Barrick & Mount, 1991; Salgado, 1997; Tett, Jackson, & Rothstein, 1991). These studies, in general, did not distinguish between the conditions of testing, whether under applicant or nonapplicant motivational conditions. Because of the large numbers of respondents in nonapplicant studies, these data carried considerable weight in meta-analyses, which have been interpreted as having an important bearing on the use of personality measures in selection. However, our data suggest that testing data gathered from respondents completing personality measures for reasons other than personnel selection have limited bearing on their use for job applicants, in the absence of convincing data to the contrary.

The susceptibility to faking found for the measures used in this study is especially remarkable when one considers the extraordinary methods used in scale construction designed to suppress the role of item desirability in responses to the

items. Jackson (1970, 1994) outlined these procedures in detail; they include the development of large item pools and the use of multivariate statistical procedures designed to maximize item–scale correlations, minimizing relations with desirability and extraneous content. For measures not subjected to similar methods of scale development, one might reasonably expect more extreme dissimulation effects. Their emergence with scales constructed using methods designed to suppress desirability attests to the strong motivation that job applicants have to manage impressions.

It is well known that rational job selection procedures can have considerable economic benefits (Schmidt & Hunter, 1981) and that, in particular, the reduction of such counterproductive behavior as theft, loafing, and job-related substance abuse impacts substantially on organizational effectiveness and profitability. It is also well known that the utility of selection decisions is a function among other things of the effectiveness of the selection procedures used. Recently, much attention in the study of personality and job performance has focused on the validity of various traditional single-stimulus types of personality measures. Our results suggest an alternative: the value of studying and identifying the means for improving testing measures by recognizing and controlling for the motivated distortion that job applicants frequently demonstrate.

It would be desirable to investigate further the impact of faking on the prediction of counterproductive workplace behavior, particularly using other criterion measures in addition to those based on self-report. However, this is not easy. First, any assessment of personality sponsored by an organization, even if presented as nonevaluative, is likely to elicit some degree of bias in self-presentation. Second, it is widely recognized that only a small percentage of instances of serious counterproductive behavior, such as theft, are observed; when they are, incumbents are usually terminated, consequently becoming unavailable for further personality assessment.

CONCLUSIONS

1. A forced-choice personality measure of integrity yields only about one third of the mean shift in the desirable direction under job applicant instructions than does the same measure administered using a single-stimulus format. A single-stimulus personality questionnaire shows a substantial mean shift, nearly a full standard deviation, from straight-take to job application conditions.

2. A traditional single-stimulus personality-based measure designed to predict potential workplace counterproductive behavior is less effective at identifying self-reported workplace delinquency when administered with instructions for the participant to respond as if applying for a job. When the same items are organized into a properly developed forced-choice format, the forced-choice format

maintains its correlation with reported workplace delinquency under job applicant conditions.

ACKNOWLEDGMENTS

We thank Richard D. Goffin and Sampo V. Paunonen for their helpful advice.

REFERENCES

- Ash, P. (1971). Screening employment applicants for attitudes toward theft. *Journal of Applied Psychology, 55*, 161–164.
- Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19*, 289–303.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.
- Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261–272.
- Betts, G. L. (1947). The detection of incipient army criminals. *Science, 103*, 93–96.
- Christiansen, N. D., Edelstein, S., & Fleming, B. (1998, April). *Reconsidering forced-choice formats for applicant personality assessment*. Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*, 847–860.
- Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement, 32*, 579–599.
- Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996, August). *The validity of non-cognitive measures decays when applicants fake*. Paper presented at the annual conference of the Academy of Management, Cincinnati, OH.
- Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on forced-choice self-description checklist. *Personnel Psychology, 15*, 13–24.
- Edwards, A. L. (1954). *Edwards Personal Preference Schedule*. New York: Psychological Corporation.
- Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166.
- Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385–400.
- Furnham, A. (1990). The fakeability of the 16PF, Myers–Briggs and FIRO–B personality measures. *Personality and Individual Differences, 11*, 711–716.
- Graham, W. R. (1958). Social desirability and the forced-choice method. *Educational and Psychological Measurement, 18*, 387–401.
- Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: Vol. 1. Studies in deceit*. New York: Macmillan.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.
- Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209–244.

- Houtchens, H. M., & Betts, G. L. (1947). Word portraits of Army prisoners. *American Psychologist*, *2*, 327.
- Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spelberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 61–96). New York: Academic.
- Jackson, D. N. (1994). *Jackson Personality Inventory Research manual—Revised manual*. Port Huron, MI: Sigma Assessment Systems.
- Jackson, D. N. (in press). *Employee Selection Questionnaire manual*. Port Huron, MI: Sigma Assessment Systems.
- Jackson, D. N. (2000). *Jackson Vocational Interest Survey manual*. Port Huron, MI: Sigma Assessment Systems.
- Jackson, D. N., & Payne, I. R. (1963). Personality scale for shallow affect. *Psychological Reports*, *13*, 687–698.
- Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational setting. *Personality and Individual Differences*, *18*, 605–609.
- McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, *43*, 335–354.
- Mikulay, S. M., & Goffin, R. D. (1998). Measuring and predicting counterproductivity in the laboratory using integrity and personality testing. *Educational and Psychological Measurement*, *58*, 768–790.
- Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*, 263–280.
- Norman, W. T. (1967). Personality measurement, faking and detection: An assessment method for use in personnel selection. In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 374–391). New York: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- 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.
- Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.
- Paunonen, S. V., Rothstein, M. G., & Jackson, D. N. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior*, *20*, 389–405.
- Rossé, 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.
- Sackett, P. R., Burris, L. R., & Callaghan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology*, *42*, 491–526.
- Sackett, P. R., & Decker, P. J. (1979). Detection of deception in the employment context: A review and critical analysis. *Personnel Psychology*, *32*, 487–506.
- Sackett, P. R., & Harris, M. M. (1984). Honesty testing for personnel selection: A review and critique. *Personnel Psychology*, *42*, 491–529.
- Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty, integrity, conscientiousness, dependability, trustworthiness and reliability for personnel selection. *Personnel Psychology*, *49*, 787–829.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology*, *82*, 30–43.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, *36*, 1128–1137.

- Tett, R. P., Jackson, D. N., & Rothstein, M. G. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742.
- Villanova, P. V., Bernardin, J. H., Johnson, D. L., & Dahmus, S. A. (1994). The validity of a measure of job compatibility in the prediction of job performance and turnover of motion picture theatre personnel. *Personnel Psychology, 47*, 73–90.
- Waters, L. K. (1965). A note of the “fakability” of forced-choice scales. *Personnel Psychology, 16*, 187–191.
- Zavala, A. (1965). The development of the forced-choice rating technique. *Psychological Bulletin, 63*, 117–124.