HUMAN PERFORMANCE, *18*(3), 267–307

Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment

Neil D. Christiansen, Gary N. Burns,
and George E. Montgomery
Department of Psychology
Central Michigan University

The effects of motivated distortion on forced-choice (FC) and normative inventories were examined in three studies. Study 1 examined the effects of distortion on the construct validity of the two item formats in terms of convergent and discriminant validity. The results showed that both types of measures were susceptible to motivated distortion, however the FC items were better indicators of personality and less related to socially desirable responding when participants were asked to respond as if applying for a job. Study 2 considered the criterion-related validity of the inventories in terms of predicting supervisors' ratings of job performance, finding that distortion had a more deleterious effect on the validity of the normative inventory with some enhancement of the validity of the FC inventory being observed. Study 3 investigated whether additional constructs are introduced into the measurement process when motivated respondents attempt to increase scores on FC items. Results of Study 3 indicated that individuals higher in cognitive ability tend to have more accurate theories about which traits are job-related and therefore are more successful at improving scores on FC inventories. Implications for using personality inventories in personnel selection are discussed.

Despite optimism that motivated distortion does not represent a serious threat to personality tests used to aid organizational decision making (Barrick & Mount, 1996; Hogan, Hogan, & Roberts, 1996; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ones & Viswesvaran, 1998; Ones, Viswesvaran, & Reiss, 1996), evidence continues to emerge that attempts to improve scores on self-report per-

Requests for reprints should be sent to Neil D. Christiansen, Department of Psychology, Central Michigan University, Mt. Pleasant, MI 48859. E-mail: chris1nd@cmich.edu

sonality inventories may be a problem in applicant samples. For example, when surveyed confidentially a substantial proportion of applicants admit to intentionally misrepresenting themselves in self-reports collected as part of the hiring process (Donovan, Dwight, & Hurtz, 2003; McDaniel, Douglas, & Snell, 1997). Applicant scores on personality tests have also been shown to be inflated compared to nonapplicants and to correlate more highly with measures of socially desirable responding (Hough, 1998; Rosse, Stecher, Miller, & Levin, 1998). Perhaps of greater concern, Hough (1997) analyzed 764 validity coefficients and found that personality test scores from applicant samples did not predict performance as well as those from incumbents. Thus, data from applicant samples are generally consistent with the findings of simulations where the validity of distorted scores has been shown to differ from that of scores from individuals less motivated to distort responses (e.g., Douglas, McDaniel, & Snell, 1996; Mueller-Hanson, Heggestad, & Thornton, 2003).

Although the effects of applicant distortion continue to be debated in academic circles, substantial skepticism exists in industry regarding the use of self-report measures to facilitate hiring decisions. Concern that motivated applicants can easily distort personality measures remains the most widespread criticism organizational decision makers have of personality testing (Cook, 1993; Hogan & Hogan, 1992; Hogan et al., 1996). Disregarding such admonitions, many personality inventories popular in industry have been developed without concern over socially desirable responding due to the belief that self-enhancement represents valid trait variance and, therefore, is not a problem (see e.g., Costa & McCrae, 1992; Hogan & Hogan, 1992). By better addressing these concerns, applied psychologists may engender more confidence in the professionals our practice serves. Indeed, a recent survey of practitioners who work in the area of selection and assessment found that approximately 70% expressed preference for using a personality inventory that includes a method to deal with applicant distortion (Goffin & Christiansen, 2003).

One of the earliest methodologies proposed for dealing with motivated distortion in personality assessment was to employ forced-choice (FC) item formats. These formats result in response options equated in terms of perceived attractiveness so that respondents cannot simply describe themselves more favorably in an effort to create a positive impression (e.g., Berkshire, 1958; Edwards, 1959). In spite of some early promise (Zavala, 1965), the popularity of the FC response format declined throughout the 1970s to the point that few professionals advocated their use. For example, contemporary texts on psychological testing tend to discount FC formats for the control of response bias and recommend against their use in inventory construction (e.g., Anastasi & Urbina, 1997; see also Paulhus, 1991). In addition, use of FC items in commercial personality inventories is relatively uncommon. For example, Goffin and Christiansen (2003) recently reviewed the strategies used to combat motivated distortion in 14 of the personality inventories most commonly used in applied settings. Of these, only one test relied on FC items exclusively (the Occupational Personality Questionnaire 4.2; SHL, 1998) and one for approximately 10% of the items (the PDI Employment Inventory; Paajanen, Hansen, & McLellan, 1993). The remaining inventories all use normative, single-stimulus items. This research takes a closer look at FC formats and their use in contexts where attempts to dissimulate may be prevalent.

FC METHODOLOGY

FC methods direct respondents to choose among two or more options that appear equally acceptable but differ in terms of their validity. The matched responses are intended to be roughly equal in terms of their desirability with each choice generally representing a different trait. As such, FC formats can be seen as a special case of more general strategies to develop items along a narrow band of social or job desirability. However, an important distinction exists between FC methodology and the way normative items are typically developed to minimize socially desirable responding. In the latter cases, extremely desirable or undesirable options are generally excluded in favor of "neutral" responses, raising questions regarding content sampling. With the FC methodology, very desirable or undesirable options can be included provided that equally acceptable alternatives can be found.

The use of FC methodology to control socially desirable responding requires two types of information regarding each potential response alternative: an attractiveness index and a validity index. Attractiveness is often derived from ratings of desirability for response options, correlations of the endorsement of options with some measure of desirability, or the overall frequency of endorsement in some normative group. Validity indexes are most commonly based on the factor loadings of response options or their theoretical relevance to the predictor construct, although some have suggested a criterion-keyed approach (Waters, 1965).

FC questions were initially popularized for use in self-descriptive personality inventories but this methodology quickly spread to other employment contexts where enhancement is probable. For example, Caroll and Nash (1972) used such an approach in designing rating forms to be included with letters of recommendation, where ratings (and the letters themselves) tend to be uniformly positive and not very discriminating. Similarly, in the 1940s the U.S. Army developed an FC methodology for performance appraisals where leniency errors often result in consistently favorable ratings (Travers, 1951). Finally, research in the assessment of job compatibility by Bernardin and colleagues (e.g., Bernardin, 1989; Villanova, Bernardin, Johnson, & Dahmus, 1994) has used FC formats to ensure that respondents cannot simply improve the assessment of their congruence to a job by distorting reports of their proclivities and inclinations. In many such contexts, FC techniques have demonstrated some degree of validity for accurately assessing worker characteristics by minimizing tendencies to give strictly favorable responses.

CRITICISMS OF FC FORMATS

Historically, two general criticisms of the use of FC items have carried the most weight. First, early research suggested that FC tests were still susceptible to response distortion in spite of test developers' attempts to equate response choices on attractiveness. Second, critics identified psychometric problems associated with the relative nature of the scoring system typically found with FC items.

Early research suggested that the FC technique described previously was successful at reducing socially desirable responding in normal assessment circumstances. However, subsequent research questioned the effectiveness of the method when situational demands were more pronounced. Studies of the fakability of FC inventories conducted in the late 1950s and early 1960s clearly showed that when respondents were provided with a specific job-oriented response set, distortion was present (Borislow, 1958; Corah, Feldman, Cohen, Meadow, & Ringwall, 1958; Dicken, 1959; Dunnette, McCartney, Carlson, & Kirchner, 1962; Feldman & Corah, 1960; French, 1958; Graham, 1958; Izard & Rosenberg, 1958; Krug, 1958; Maher, 1959; Norman, 1963). Comparisons of applicants to job-holders were less consistent in the amount of distortion found, although all studies reported some differences (Bass, 1957; Dunnette et al., 1962; Gordon, 1963; Kirchner, 1962). Understandably, the emergence of so many studies in a brief period showing that FC inventories were still susceptible to distortion shook the faith of the psychological community in the merit of this methodology.

Importantly, finding motivated individuals are able to elevate scores in an absolute sense may not be the most relevant standard for determining the usefulness of the format for applicant personality assessment. A more relevant question is whether scores obtained using an FC methodology are better indicators of applicants' actual personality than those obtained using common alternatives. A recent meta-analysis by Stanush (1997) of faking studies found that the standardized mean difference between normal and faking conditions was significantly smaller for FC inventories than that typically observed for inventories not employing this format. If FC inventories were less susceptible to distortion, scores from these assessments may be better estimates of applicants' actual personality than more normative alternatives (Baron, 1996).

The other criticism involving the relative nature of FC responses has received more attention in the recent literature. With FC formats, judgments are typically made about which option is most true relative to the others. As a result, some have argued that the resulting scores do not reveal anything about the absolute trait elevations and are thus inappropriate for comparisons across individuals (e.g., Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988). Because response choices for each item are generally statements from different traits, choosing one response means that one of another trait is not chosen. When a small number of traits are assessed, being high in elevation on one trait necessitates lower scores on others. This results in mathematical dependency among the trait scales that can create numerous problems for multivariate analyses, such as increased collinearity in regression analyses and improper solutions using factor analytic techniques. Cornwell and Dunlap (1994) present a convincing summary of such criticisms.

However, when evaluating the severity of these condemnations two important points need to be kept in mind. First, whether comparisons across individuals are appropriate is largely an empirical question involving evidence of construct-related validity. Research that has directly compared FC and traditional item formats has found considerable convergence between normative and FC methods. For example, Lanyon (1966) found that scores from items presented in a free-choice format (yes/no) correlated quite highly with FC format with a median correlation of .73 between scores from scales using the two item types. Jackson, Neill, and Bevan (1973) used a similar strategy and reported that responses to the FC versions of the Personality Research Form (Jackson, 1984) were positively correlated with responses to the traditional true/false items. Moreover, although internal consistency estimates were slightly smaller for the FC inventories, they showed similar levels of convergence with alternate measures of the traits (Jackson et al., 1973). Thus, individuals high in trait elevation in a relative sense clearly tend to be high in an absolute sense as well and the constructs being measured by FC measures are strongly related to the constructs measured by their normative counterparts.

Second, most of the debate unfortunately has presumed FC methodology and ipsative scoring to be synonymous. With ipsative scoring, the sum of all trait scores is constant across individuals. This result, if desired, may be realized not only through the use of FC methodology, but also from data gathered with normative inventories by subtracting each respondent's mean score across measures, generating deviation scores that sum to zero (e.g., Baron, 1996; Hicks, 1970). On the other hand, FC inventories vary considerably in the degree to which their scores are ipsatized (Bartram, 1996) with some constraining the overall sum of scores to a much greater extent than others. Other factors, such as inventories with differing numbers of items or inclusion of unscored response options, may result in partial ipsatization. Another variant involves the choice between two responses where both are related to the same trait, but in a different direction. This format was used in the construction of the Meyers-Briggs Type Indicator and results in more normative scoring, although developers were unable to equate successfully response options on desirability (DeVito, 1985).

Underlying the criticisms surrounding the relative nature of FC responses is that the dependency introduced will compromise measurement the most when different traits are considered in tandem. However, not all traits are of interest for the majority of organizational activities that use personality measures. For most jobs only a subset of the potential traits are judged by experts to be job related (e.g., Raymark, Schmit, & Guion, 1997) or are found to correlate empirically with performance criteria (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Once those traits have been identified that are job related, constructing FC items without including response choices from more than one trait of interest may be possible. Filler choices for the items could be gleaned from those traits that have been demonstrated to be unrelated to the job and some of the problems associated with ipsative measures potentially avoided—each item will contain at least as many responses that are unscored and the sum of the scores should be relatively unconstrained. However, before this method could be used with applicants, those traits that are job related would have to be known in advance. This might be done based on data from the cumulative literature, through the use of a job-analysis method sensitive to the personality-based job requirements (e.g., Raymark et al., 1997) or based on the results of a concurrent validity study using a normative measure that includes a broad range of traits.

Thus, the extent that FC formats introduce difficulties into the measurement process may depend on the ultimate purpose for which scores from the personality assessment will be used. From a practical standpoint, some of the statistical anomalies associated with ipsative measurement might be avoided. However, because the alternative response options in the FC items may involve the relative elevation of individuals' other traits, these items may no longer be strictly univocal even if the other constructs are not of interest for a particular job. Additionally, the response process involved when applicants attempt to increase scores on FC personality inventories may possibly introduce other constructs into the measurement process that would be absent when responding to normative items. In particular if other individual differences result in only some applicants being motivated or capable of improving their scores on FC items, changes to the construct validity of inventories that use this format will occur. This may in turn affect criterion-related validity depending on the relation of the particular criterion to any additional individual differences involved. Taken together this argues for careful examination of how motivated distortion may affect aspects of the validity of FC formats.

In summary, critics have often confused the issue of ipsative scoring schemes with the use of multiple-response options equated on desirability. FC measures need not be scored ipsatively and have been shown to correlate with normative measures of the same trait even when purely ipsative scoring is used. Furthermore, FC formats provide better control over some common response biases that might threaten the validity of single-stimulus (SS) items. When motivation to distort is absent, this control might afford little advantage over normative measures and the additional time and effort required to develop FC items may not be cost effective. On the other hand, in applicant situations where distortion is more prevalent, FC instruments may provide more accurate information about the trait of interest. Attempts to distort FC items may possibly introduce other constructs into the measurement process, an issue that may complicate understanding of the validity of this format in applicant settings.

RECENT RESEARCH ON FC ITEM FORMATS

Despite the criticisms levied against FC methodologies, recent research has shown that FC instruments offer resistance to intentional attempts to distort responses than traditional assessment procedures. For example, Jackson, Wroblewski, and Ashton (2000) administered integrity tests using SS and FC response options to participants under two conditions: a normal ("straight-take") condition where test-takers were told to answer as truthfully as possible and a job applicant instructions condition where test-takers were told to answer the items to make a good impression. Results showed that scores on the normative measure were almost a full standard deviation higher in the applicant instructions condition than in the normal condition. Test-takers in the applicant instructions condition still scored higher with the FC questionnaire, but scores were only one third of a standard deviation higher than those in the normal condition. Jackson et al. also examined test scores with self-report admissions of past counterproductive behavior, finding that the correlation between normative test scores and admissions dropped from .48 in the normal condition to .18 in the applicant instructions condition whereas with FC inventories this relation only decreased from .41 to .36. Although suggestive of the potential benefits of FC formats, this research did not investigate the effects on construct validity and is limited by the use of a self-report criterion.

Other research has begun to explore explanations for the differences between SS and FC personality inventories when test-takers are motivated to distort their responses. Dyomina, Vasilopoulos, Cucina, and Reilly (2003) administered SS and FC personality items to students in a normal instructions condition and in a condition asking them to respond as if applying for admission to a university. Similar to past research, scores on the normative measures were almost three fourths of a standard deviation higher in the applicant than in the normal instructions condition, whereas the corresponding FC scores were less than one fourth of a standard deviation higher than the normative scores. In addition to the personality inventories, participants also completed the Wonderlic Personnel Test (WPT; Wonderlic, 1999), a measure of general cognitive ability. The results indicated that the SS personality measures had little to no relation with cognitive ability in either the normal ($r = –.09$ to .06) or applicant ($r = –.01$ to –.00) instructions conditions. However, although little relation was found between FC scores and WPT scores in the normal instructions condition ($r = –.03$ to .10), significant correlations were found in the applicant instructions condition ($r = .32$ to .40).

This research investigated the effects of motivated distortion intended to improve the chances of being selected for a job on the validity of personality inventories that use an FC methodology. Study 1 evaluated the impact of motivated distortion on the construct validity of FC items by examining evidence of convergent and discriminant validity. In Study 2 employed students completed FC and normative measures of Conscientiousness under normal and applicant instructions conditions. Scores were then correlated with supervisors' ratings of job performance to evaluate the effect on criterion-related validity using an external criterion. Finally, Study 3 investigated whether additional cognitive processes are involved when motivated respondents attempt increase scores on FC items, introducing additional constructs into the measurement process.

STUDY 1: EFFECTS OF MOTIVATED DISTORTION ON THE CONSTRUCT VALIDITY OF FC MEASURES

Study 1 evaluated whether FC personality tests successfully reduce distortion and maintain better construct validity than a normative measure when respondents are motivated to distort. To do this effectively, two conditions must be met that are difficult to realize in a field sample. First, determining unequivocally which individuals have distorted responses on the personality measure and which have not must be possible. Second, independent assessments of trait elevation must be available outside of the situational demands that resulted in score inflation. In such a situation, mean trait elevations of normal and distorted responses can be compared as is done in laboratory faking studies or investigations of differences in applicant and incumbent score distributions. More important, convergence with the independent assessment can be estimated to provide evidence of construct validity and how much trait variance is deteriorating due to distortion. A simulation was therefore conducted where these conditions could be satisfied.

In this simulation, participants completed FC and SS personality items along with a social desirability (SD) measure. Half completed these measures under instructions to describe themselves honestly (normal instructions condition), whereas the other half were given instructions to fill out the remaining measures as if applying for a job they desired very much (applicant instructions condition). To provide participants in the applicant instructions condition with more direction, they were also given a job description of the desired sales position. To obtain estimates of true scores untainted by the demand characteristics of the applicant situation, all participants completed alternate measures of the same traits prior to being exposed to the manipulation. Thus, directly evaluating two important questions that need to be addressed for the FC format to be a viable alternative in applicant contexts was possible: (a) Does the FC method reduce distortion (as indicated by a mean shift) caused by a job applicant response set? and

(b) Do scores from FC instruments provide more accurate trait information than a normative format when distortion is present?

Based on the research outlined previously, the following hypotheses were developed:

H1: Personality test scores will be more favorable for the job in the applicant instructions condition than in the normal instructions condition, resulting in higher mean scores for Extraversion and Conscientiousness. These differences will be more pronounced for responses to normative items than those using a FC format.

H2: Scores from the FC and SS inventories will be positively correlated, indicating the formats are measuring related constructs. Strong evidence of convergence across methods is expected if the rank order of individuals within each method is similar and both are providing information about the absolute trait elevations. This should be particularly true in the normal instructions condition where scores are expected to reflect true trait elevations more faithfully.

H3: The correlation between the FC scores and SD scores will be smaller than those observed for the SS inventory scores.

H4: In the applicant instructions condition, FC scores will explain incremental variance in the premanipulation measures beyond that explained by the SS inventory scores. This would not be expected in the normal instructions condition; only when situational demands to distort scores are strong would FC scores be expected to be better predictors of true trait elevation

Method

Participants and Procedure

Participants in the study were 350 undergraduate students (64% female and 36% male) with a median age of 19.8 years. Participants were instructed that they would be answering several personality inventories to compare the statistical properties of the different measures. The measures were completed in groups of 10 to 15 individuals with all participants in each session being randomly assigned to either the normal or applicant instructions conditions. In both conditions, all participants completed an initial personality assessment with standard instructions to answer all questions honestly and that there were no correct or incorrect responses. In the normal instructions condition, participants then completed FC personality inventories, a popular measure of SD, and an SS personality inventory (completed in counterbalanced order). Instructions for the postmanipulation assessment in the normal condition were similar to the initial assessment.

The other half of the participants received written and oral instructions explaining that personality inventories such as the ones they were completing are sometimes used in industry to help select the best persons for a particular job. They were then provided with a job description of a sales position (patterned after one used by a local organization) and instructed to imagine that they had been asked to complete the remaining inventories as part of the application process for that job. The participants in the applicant instructions condition were then given the same measures as described previously.

Measures

Premanipulation personality assessment. All participants completed the Extraversion and Conscientiousness measures of the Revised NEO Personality Inventory (Costa & McCrae, 1992) prior to the manipulation. These traits were chosen because they have been shown to be valid predictors of performance across sales jobs (Barrick & Mount, 1991), results that are consistent with the traits hiring managers considered important for such positions (Dunn, Mount, Barrick, & Ones, 1995). Each inventory consists of 48 items recorded on a 5-point scale (anchored from *strongly disagree* to *strongly agree*) with higher scores indicating more extraverted and conscientious dispositions. The internal consistency reliabilities for this measure were .90 for Conscientiousness and .86 for Extraversion.

Social desirability. The measure of social desirability used was the impression management measure of the Balanced Inventory of Socially Desirable Responding (BIDR; Paulhus, 1988). The BIDR has been commonly used in past research to make inferences about response distortions (e.g., Barrick & Mount, 1996; Rosse et al., 1998). This measure yields scores for two dimensions, self-deception and impression management; because the concern of this research lies principally in intentional distortion, only the 20 items of the impression management dimension were used. Consistent with Paulhus (1988), scores were calculated by scoring only the extremely desirable responses as contributing a point to the total score.

SS personality inventory. The personality measures used in the post-manipulation were based on the adjectives used in the Extraversion and Conscientiousness inventories of the factor markers Goldberg (1992) developed. Items on these instruments were developed to be univocal indicators of their respective Big Five factors. Goldberg reports correlations between scores on the factor markers and those of the NEO Personality Inventory (Costa & McCrae, 1985) of .69 for Extraversion and .65 for Conscientiousness. The 20 adjectives in each measure were administered in their unipolar format with responses recorded on a 9-point Likert scale (anchored from *not-at-all true of me* to *very true of me*). Scores were computed as the sum of the 20 items keyed such that higher scores indicated a more favorable trait elevation for sales selection.

FC personality inventory. An FC inventory was developed to measure the traits of Conscientiousness and Extraversion based on the adjectives used in Goldberg's (1992) factor markers. The job relatedness of Conscientiousness and Extraversion were based on meta-analytic studies of sales position where these traits of the Big Five have been estimated to have true validity greater than .10, whereas Agreeableness and Openness have been shown to be virtually uncorrelated with successful sales performance (e.g., Barrick & Mount, 1991). The FC personality inventory was then developed using Extraversion and Conscientiousness markers to represent sales-related behavior and Openness and Agreeableness markers to represent nonsales-related behavior.

The initial step in the development of the prototype FC scales was to obtain attractiveness indexes for the adjectives comprising the Extraversion and Conscientiousness inventories. Desirability ratings were obtained from 36 undergraduates who read the job description and rated how desirable the trait would be in applicants for the sales position described. Ratings were made on a 9-point scale (anchored from *not-at-all desirable* to *very desirable*). Attractiveness indexes for each adjective were then developed based on the average desirability ratings across respondents with mean desirability ratings for the adjectives as follows: *withdrawn* (2.6) to *active* (7.6) for Extraversion; *inefficient* (1.4) to *prompt* (8.5) for Conscientiousness; *rude* (1.6) to *cooperative* (8.3) for Agreeableness; and *unintelligent* (2.0) to *bright* (7.9) for Openness to Experience.

Eighty trial items were constructed by pairing adjectives from each inventory with an adjective from the Openness to Experience or Agreeableness list that had a comparable desirability rating (within .5 standard deviations of the target adjective). These trial items were then piloted on 40 undergraduates to identify the most promising items based on two criteria: (a) the proportion of the sample endorsing the target adjective was considered with trial items having endorsements closer to 50–50 being favored; (b) item-total correlations were considered such that trial items with higher item-total correlations were favored.

The final inventories each consisted of 20 items in dyadic format with respondents being asked to choose which of the two adjectives was most true of them. For each item, only one choice indicated a higher elevation on a job-related trait. For items with pro-trait adjectives, choosing the job-related trait added 1 point toward the inventory total. For items with con-trait adjectives, choosing the negative job-related choice indicates a lower trait elevation. Because of this, 1 point was added to the total score when the alternative was selected (which is the same as adding a constant to all scores equal to the number of negative items and subtracting one for each negative job-related trait chosen). Scores for the FC Extraversion and Conscientiousness inventories could thus range from 0 to 20.

The following is an example of instructions, items, and keyed responses (shown with an asterisk) from each inventory:

Which of the following adjectives is MOST TRUE or MOST DESCRIPTIVE of you? Are you more:

(A) QUIET or DEMANDING*
(B) PRACTICAL* or IMAGINATIVE
(C) UNKIND* or CARELESS
(D) SYMPATHETIC or BOLD*

Results

Mean Differences Between Conditions

The initial step in analyzing the data was to compare mean scores in the postmanipulation measure to mean scores obtained from the normal and applicant instructions conditions. Table 1 shows descriptive statistics and statistical comparisons. As might be expected, the means for all of the measures were significantly higher when they were completed with an applicant response set rather than when under an induction to answer honestly. This clearly shows that response distortion occurred under this condition with the difference in SD means serving as a manipulation check confirming that those responding as applicants were attempting to choose responses to create a more favorable impression, [$d = .49$, $F(1, 348) = 21.28$, $p < .01$].

TABLE 1
Descriptive Statistics and Comparisons of Postmanipulation Scores From Study 1

	Instructions						
	Normal		*Applicant*		*Comparison*		
Measure	*M*	*SD*	*M*	*SD*	*d*	*F*	R^2
Single-stimulus							.17*
Conscientiousness	132.67	18.14	145.03	18.29	0.68	40.24*	
Extraversion	115.82	22.32	131.77	20.57	0.74	48.35*	
Forced-choice							.07*
Conscientiousness	8.42	4.41	10.22	4.68	0.40	13.72*	
Extraversion	6.42	4.41	8.53	4.49	0.47	19.62*	
Social Desirability	5.29	2.52	6.82	3.57	0.49	21.28*	.06*

Note. $n = 175$ in normal condition and $n = 175$ in applicant instructions condition. *d* indicates the standardized mean difference obtained by subtracting the mean of the normal condition from that of the applicant and dividing by the pooled standard deviation; R^2 is the overall percentage of variance explained by the manipulation in the two trait scales of each format and the desirability scale.

*$p < .05$.

The differences in the personality inventory scores support the first part of H1, predicting score inflation for measures of both the FC and SS formats when participants were asked to respond as if applying for the sales job. Specifically, across traits the FC inventory scores were higher in the applicant instructions condition than those observed in the normal instructions condition [$d = .43$, $F(1, 348) = 21.30$, $p < .01$]. This was also true for the SS inventories with the standardized mean difference being even larger [$d = .71$, $F(1, 348) = 68.43$, $p < .01$]. A comparison between the (dependent) effect sizes indicated that the difference was significantly greater for the FC format than the SS inventories [$F(1, 348) = 6.89$, $p < .01$]. This supports the second expectation of H1, that the FC inventories would be less susceptible to distortion than those constructed using SS items.

Relations Between FC and SS Measures

The relations between the postmanipulation trait scores were next examined to provide evidence of convergence between the FC and SS methods. This was done by observing the correlations in each condition between scores measuring the same trait (see Table 2), and then computing Cohen's set correlation to assess the overall rela-

TABLE 2
Correlation Between Study Variables in Normal and Applicant Instructions Conditions

	Conscientiousness			*Extraversion*			
Measure	*SS*	*FC*	*NEO*	*SS*	*FC*	*NEO*	*SD*
Normal Instructions Condition							
SS Conscientiousness	(.84)						
FC Conscientiousness	.54*	(.81)					
NEO Conscientiousness	**.61***	**.59***	(.92)				
SS Extraversion	.16	–.05	.08	(.88)			
FC Extraversion	-.10	.02	.00	.68*	(.84)		
NEO Extraversion	.11	–.05	.10	**.68***	**.52***	(.89)	
Social desirability	.44	.13	.44*	.14	.01	.05	(.81)
Applicant Instructions Condition							
SS Conscientiousness	(.87)						
FC Conscientiousness	.33*	(.83)					
NEO Conscientiousness	**.28***	**.43***	(.93)				
SS Extraversion	.44*	.18	.04	(.87)			
FC Extraversion	.12	.39*	.11	.65*	(.81)		
NEO Extraversion	–.01	–.04	.22*	**.34***	**.42***	(.88)	
Social desirability	.54*	.11	.16	.46*	.08	.07	(.73)

Note. n = 175 in normal condition and n = 175 in applicant instructions condition. FC = forced-choice scales; SS = single-stimulus scales; NEO = NEO-PIR scale. Estimates of internal consistency are provided in parentheses on the diagonal. Correlations between premanipulation and postmanipulation trait scores measuring the same trait are bold to facilitate comparison.

*$p < .05$.

tion between the two sets of measures using Rao's F ratio as a test of significance (Cohen & Cohen, 1983). The results of these analyses showed considerable convergence between the FC and the SS measures. In the normal condition the two measures of Extraversion correlated .68 and those of Conscientiousness .54 with an overall set correlation between the inventories of $R = .83$. In the applicant instructions condition these correlations were .65 and .33 with $R = .74$. All of these relations were significant at the .01 level. The level of convergence is consistent with the idea that both formats are providing information about the absolute trait elevations and that the rank order of individuals is similar across methods. This supports the H2 prediction that the FC scores would provide information about absolute trait elevations and therefore should correlate with the normative measures.

Relations With Socially Desirable Responding

The third step in analyzing the data from Study 1 was to examine the relations of the trait inventories with the SD measure to corroborate that FC scores were less related to attempts at impression management. Table 2 reports the correlations between the study variables, including trait scores from the FC and SS formats. As Table 2 shows, across conditions the FC scores correlated between .01 and .13 with SD scores with none of the relations being significant. In contrast, the same estimates of the SS scores ranged from .14 to .54, showing considerable dependency between these scores and an index of how much respondents were intentionally trying to present a favorable impression. Regressing the SD scores from the applicant instructions condition on the two trait scores obtained indicated that the SS scores explained 35% of the variance [$R = .59$, $F(2, 172) = 46.80$, $p < .01$], whereas the FC scores only explained approximately 1% of the variance in desirability scores [$R = .12$, $F(2, 172) = 1.28$, *ns*]. A comparison between the (dependent) effect sizes indicated that the multiple correlation was significantly greater for the SS inventories than for the FC inventories [$F(1, 348) = 46.24$, $p < .01$], supporting H3.

Relations With Premanipulation Trait Measures

Finally, the relations between the premanipulation and postmanipulation traits score were examined to gauge how well the prior estimates of trait elevation could be explained by the items of each format. In the normal condition, the FC Extraversion measure correlated .52 with the premanipulation measure; the same correlation for Conscientiousness was .59. For the SS inventories completed in the normal instructions condition, these correlations were higher, ($r = .68$ for Extraversion and .61 for Conscientiousness).

In the applicant instructions condition, the correlations between the premanipulation and postmanipulation measures for the FC and SS inventories were notably smaller than their counterparts in the normal instructions condition. However, in the applicant instructions condition the FC inventories correlated slightly higher with the premanipulation trait scores ($r = .42$ for Extraversion and $r = .43$ for Conscientiousness) than did the SS inventories ($r = .34$ and $r = .28$, respectively). On average, the correlation for the FC measures went down by .21, whereas the applicant responses set caused a decline in the SS correlations by approximately .34.

To estimate the overall amount of variance in the premanipulation trait measures explained by the different formats and to test the hypothesis that the FC inventories would explain incremental validity in the applicant instructions condition, Cohen's set correlation was again used. Table 3 offers the results of a series of set correlation analyses with the two premanipulation trait scores as dependent variables. The first column (*R*) is the overall set correlation between the FC and SS inventories with the premanipulation trait measures. Following the pattern of univariate correlations discussed previously, in the normal condition the SS inventories explained slightly more variance ($R = .81$) than did the FC inventories ($R = .74$), whereas in the applicant instructions condition the FC inventories explained more trait variance ($R = .64$) than did the SS inventories ($R = .50$).

The second column of Table 3 (ΔR^2) indicates the incremental change in the squared set correlation after removing the variance attributable to the other format's scores with the third column ($F\Delta$) corresponding to the test of significance for the incremental variance using Rao's *F* ratio (at *df*'s of 4 and 338). In the normal condition, the instruments of both formats had significant unique contributions to explaining variance in the initial personality assessment. Specifically, the two FC inventories explained an additional 12% of the premanipulation trait variance beyond the SS inventories; the SS inventories explained an additional 26% after controlling for that explained by the FC scores. However, in the applicant in-

TABLE 3
Set Correlation Analysis of the Incremental Trait Variance
Explained by Single-Stimulus and Forced-Choice Conditions

Scale Format	*R*	ΔR^2	$F\Delta$
Normal Instructions			
Single-stimulus	.81*	.26*	13.66*
Forced-choice	.74*	.12*	5.80*
Applicant Instructions			
Single-stimulus	.50*	.04	1.68
Forced-choice	.64*	.20*	9.91*

Note. $n = 175$ in normal condition and $n = 175$ in applicant instructions condition. *R* refers to the set correlation obtained when the 2 trait measures of that format are correlated with the premanipulation trait measures; ΔR^2 indicates the incremental change in the squared set correlation after removing the variance attributable to the other format's scale scores; ΔF corresponds to the test of significance for the incremental variance explained at 4 and 338 *df*s.

*$p < .01$.

structions condition only the FC inventories contributed significant incremental variance ($\Delta R^2 = .20$) and the SS inventories did not ($\Delta R^2 = .04$). This provides strong support for H4, indicating that the FC scores were providing more information about actual trait elevations in the applicant instructions condition than did the scores derived from the SS ratings.

Discussion

Originally conceived of as a method of minimizing response distortion, FC inventories are widely believed to be flawed for most types of personality assessment. However, some have also argued that the control over response bias might still be advantageous in situations where motivation to distort is acute. Study 1 evaluated whether the FC method reduces distortion caused by an applicant response set and if they provide additional information about the actual trait elevation when responses are distorted. Based on the results of the laboratory simulation, it appears that criticisms of FC formats for personality assessment may be overstated when distortion is prevalent. The data demonstrate that FC instruments that avoid including more than one scored alternative provide information about absolute trait elevations, are not negatively correlated, and may be better indicators of applicant personality than normative methodologies.

Several conclusions can be drawn from comparisons of scores from instruments constructed with FC and SS items when half of the participants had been instructed to respond as if applying for a sales position. First, although the applicant response set caused scores from both formats to shift, the change was markedly smaller in the FC assessment. Second, given the convergence of the FC scores with the scores of two other SS measures of the same trait, information about individuals' elevation on the trait of interest clearly is provided by both assessment methodologies. Third, relations with a measure of socially desirable responding indicated more of this bias was present in the SS inventories than in the inventory constructed using the FC methodology. Also related to discriminant validity, the correlations among traits measured by the same format were stronger for the SS inventories. Fourth, evidence indicated that whereas scores from the SS inventories were better predictors of prior estimates of trait elevations in a normal assessment situation, scores from the FC inventories were better predictors when participants were asked to respond as if applying for a job.

Taken together the evidence suggests that the constructs the FC inventories measured were related to those the normative inventories measured. However, even in the normal instructions condition the correlations (.52 to .68) may not be high enough to suggest that the constructs measured by the two inventories are identical. Given that the FC items involve response options known to be linked to other traits, other individual differences were most likely involved in responses to the FC items that were not involved in responses to the SS items. Finally, when instructed to respond as if applying for the sales job even less convergence was noted, suggesting that additional factors other than the trait of interest may play an even larger role to FC responses when motivation to distort is high.

STUDY 2: EFFECTS OF MOTIVATED DISTORTION ON THE CRITERION-RELATED VALIDITY OF FC MEASURES

Although the results from Study 1 shed some light on some of the issues surrounding the construct validity of personality measures constructed using FC methodology, several limitations of the design could be addressed by additional research. First, because the effects of motivation to distort FC items were examined in the context of only one job, generalizability to other jobs may be limited. Patterns of distortion should therefore be examined when individuals are motivated to improve their scores where the position requirements are different than sales. Second, although the format of the normative items were developed to use the same adjectives as those used in the FC scales, most normative personality items use statements that are more rich than single adjectives (e.g., “I often forget to put things back in their place.”). Third, the FC items themselves used a dyadic response format, which may facilitate guessing the keyed response. Fourth, whereas Study 1 addressed construct validity in terms of evidence of convergent and discriminant validity, it did not examine the criterion-related validity of FC personality scores as related to important work outcomes.

Study 2 replicated and extended the results of Study 1 to address these issues. Employed students were to complete two measures of Conscientiousness, an SS-based measure using more traditional item stems and an FC measure that included three response options. Half of the participants completed both measures honestly. The other half were given a job description for a customer service position and asked to complete the measures as if part of the application process for a very desirable job. Supervisors of each student were then contacted and asked to rate the job performance of the participants.

Thus, an important goal of Study 2 was to examine how motivated distortion affects the validity of FC and normative formats using job performance as the criterion. No research to date has compared how well FC and normative formats predict job performance under conditions where motivated distortion is probable, and the exact pattern of relations that would be expected was unclear. In Study 1, although the instructions to respond such as job applicants resulted in a decay in trait variance for both the FC and SS measures, the FC format maintained construct validity better when motivation to distort was high. Consistent with this, examining the validity of FC and normative inventories, Jackson et al. (2000) found that although validity for the prediction of self-reported counterproductive behaviors was worse in the applicant instructions condition, this attenuation was only significant for the SS inventory. In addition, some theorists have suggested that the impression management involved in altering responses to self-reports in applicant contexts may be related to social competence and contribute positively to the prediction of performance (Hogan et al., 1996). To the extent this is the case and FC items eliminate this opportunity, criterion validity might also be reduced when this format is used with the applicant instructions condition.

On the other hand, one could argue that because FC formats reduce but do not eliminate distortion that only the most competent respondents are successful at increasing their scores. For example, Dyomina et al. (2003) found that when students were asked to respond as if applying for college admission, FC scores were better predictors of college grade point average than were SS scores, largely because the FC scores were more correlated with cognitive ability in that condition. This is particularly relevant to this study given that job performance has been shown to be related to cognitive ability across jobs (Schmidt & Hunter, 2004). Because of this, even if trait variance might decay for both formats, part or all of the negative impact on the relation between FC scores and job performance might be ameliorated. In this case, the relation with job performance might be relatively unaffected. Conceivably, if test score variance attributable to other job-related constructs increased at a greater rate than trait variance decayed, enhanced validity might even be observed.

Thus, existing research suggests a probable decay in the criterion-related validity of a normative personality measure as a result of the applicant instructions. In the case of FC measures, one could reasonably argue that criterion-related validity might also decline but at a lower rate. However, if additional constructs are involved, criterion validity of FC scores might be maintained or even increased. In any of these scenarios the relation between the FC scores and performance would be predicted to be superior to that of the SS scores.

> H5: Scores on the FC Conscientiousness measure will be better predictors of job performance in the applicant instructions condition than scores on the SS measure of Conscientiousness.

Method

Participants and Procedure

Participants in the study were 220 undergraduate students (69% female and 31% male) who were currently or recently employed (within the past 3 months). Of these, 122 (about 56%) had performance appraisal forms returned by their supervisors; those without completed forms were not included in subsequent analyses. The procedure in Study 2 was very similar to that used in Study 1. Participants completed multiple questionnaires in groups of 5 to 15. Half of the participants completed the personality measures with instructions to respond as honestly as possible, and the other half of the participants received a job description for a customer service position and were asked to complete the personality measures as if part of the application process for the customer service job. Participants were asked for permission to contact their work supervisors, who were then mailed performance appraisal forms along with a copy of each participant's consent form.

Measures

SS personality inventory. A 20-item personality inventory was constructed to measure Conscientiousness using items from the International Personality Item Pool (IPIP) database (Goldberg, 1997). Inventories constructed from the IPIP database have been shown to correlate highly with other established measures of personality, such as the NEO-PIR (Costa & McCrae, 1992) and the Hogan Personality Inventory (Hogan & Hogan, 1992). For example, a 20-item IPIP conscientious inventory was estimated to correlate .92 with the NEO-PIR inventory of the same name once measurement errors had been accounted for (Goldberg, 1997). Sample items include: "I make plans and stick to them," "I often forget to put things back in their place," and "I am exacting in my work." Participants responded to the 20 items using a 5-point scale (anchored from *very inaccurate* to *very accurate*) to assess the statements. The internal consistency of the SS inventory was .85 in the normal condition and .87 in the applicant instructions condition.

FC personality inventory. The FC inventory developed for Study 2 targeted the trait of Conscientiousness because the jobs of the employed students were diverse and that trait has been found to predict performance across jobs (e.g., Barrick & Mount, 1991). The same methodology as that described in Study 1 was followed to develop the FC measure with two notable differences. First, each item was designed to include three response options rather than two. Second, attractiveness indexes were based on ratings from 30 students on how desirable each adjective would be of someone applying for any job rather than a specific position. Again, for a given item all of the distracters had mean desirability ratings within .5 standard deviations of the target trait.

The final inventory of Conscientiousness consisted of 20 items in triadic format with respondents being asked to choose which of the three adjectives were most true of them. Twelve of the items were pro-trait and used adjectives positively related to Conscientiousness (e.g., *dependable, efficient, organized*). The other 8 were con-trait items involving adjectives that negatively related to Conscientiousness (e.g., *careless, unsystematic, haphazard*). Each time participants chose a pro-trait adjective related to Conscientiousness, they were awarded 1 point, whereas each time they chose a con-trait adjective related to that trait they were not. Sample items (with keyed responses shown by asterisks) include:

(A) PRACTICAL* or KIND or GENEROUS
(B) TRUSTFUL or RELAXED or THOROUGH*
(C) DISORGANIZED or JEALOUS* or RUDE*
(D) DISTRUSTFUL* or UNCOOPERATIVE* or UNDEPENDABLE

The internal consistency of the 20-item inventory was .76 in the normal instructions condition and .74 in the applicant instructions condition.

Supervisor ratings of job performance. Table 5 shows the five items supervisors completed on a performance appraisal form. Four were intended to reflect general job activities important across jobs and the last to gauge the supervisor's overall evaluation of performance. Employees were rated on a 9-point scale (anchored from *does not meet requirements* to *far exceeds requirements*). The internal consistency of the performance composite was .92.

Results

Mean Differences Between Conditions

Mean differences were examined to determine the extent that FC and SS scores were successfully distorted in the applicant instructions condition. Table 4 shows means and standard deviations for the SS and FC Conscientiousness scores. Scores were more favorable for both measures of personality in the applicant instructions condition, suggesting that participants were successfully presenting themselves as

TABLE 4
Descriptive Statistics and Comparisons of Single-Stimulus and Forced-Choice Scores From Study 2

	Instructions					
	Normal		*Applicant*		*Comparison*	
Measure	*M*	*SD*	*MN*	*SD*	*d*	*F*
Single-stimulus	71.40	11.42	84.18	10.52	.96*	36.23
Forced-choice	13.35	5.84	14.95	6.87	.24	1.91

Note. $n = 62$ in normal condition and $n = 60$ in applicant instructions condition. *d* refers to the standardized mean difference obtained by subtracting the mean of the normal condition from that of the applicant instructions condition and dividing by the pooled standard deviation; *F* indicates the test statistic at 1 and 121 *df*s.

*$p < .01$.

more conscientious. Specifically, for the SS inventory the mean score in the applicant instructions condition was approximately 1 standard deviation higher than that from the normal instructions condition [d = .96, F(1, 120) = 36.23, p < .01]. Scores on the FC measure were also higher but only by about one quarter of a standard deviation [d = .24, F(1, 120) = 1.91, *ns*] with the difference between the effect sizes being both practically and statistically significant [F(1, 119) = 16.65, p < .01]. Thus, we again observed that measures using each format were susceptible to attempts to improve scores, but that the items on the SS inventory were distorted significantly more than were the FC items.

Relation Between SS and FC Measures

Consistent with the results of Study 1, scores on the SS and FC measures were positively correlated in the normal instructions condition (r = .44, p < .01) as well as the applicant instructions condition (r = .36, p < .01). However, one should note that these correlations were smaller than the corresponding convergence estimates from Study 1. In Study 1 the FC and SS measures were both constructed using the same adjectives as stimuli; in Study 2 the FC items again used adjectives as stimuli whereas the SS inventories used declarative statements.

Relations With Job Performance

Table 5 provides the correlations between the personality measures and the performance ratings for the two conditions. As can be seen in the table, the correlations within the normal instructions condition were modest: the criterion-related validity of the SS measure was observed to be .21 using the performance compos-

TABLE 5
Correlations Between Scores on Conscientiousness Measures
and Supervisor Ratings of Job Performance

	Instructions			
	Normal		*Applicant*	
Performance Ratings	*SS*	*FC*	*SS*	*FC*
Dependably performs job duties	.20†	.14	.09	.46**
Completes tasks without mistakes	–.02	.09	–.03	.33**
Takes initiative in completing duties	.22*	.17†	.06	.34**
Does high-quality work	.22*	.21*	.11	.40**
Overall performance rating	.20†	.12	.11	.42**
Composite (α = .92)	.21*	.17†	.08	.46**

Note. n = 62 in normal condition and n = 60 in applicant instructions condition. SS = single-stimulus; FC = forced-choice item formats.

†p < .10. *p < .05. **p < .01.

ite, whereas that of the FC inventory was .17. Although not large, they are comparable to what is often obtained for measures of Conscientiousness using validity coefficients uncorrected for criterion errors (Barrick & Mount, 1991; Tett et al., 1991). Although consistent with the prediction that the SS inventory would be a better predictor of performance when strong situational demands are absent, the difference was slight [$t(59) = .61$, *ns*].

On the other hand, in the application instructions conditions the correlation between the SS measure and the performance composite was .08. Thus, compared to the normal instructions condition, the SS scores in the applicant instructions condition explained less than one fourth as much variance in performance (i.e., introducing instructions to respond as if applying for a job degraded the variance in performance explained by the SS scores from about 4% to less than 1%). In contrast, the FC scores correlated with the performance ratings to a surprising extent in the applicant instructions condition with an observed correlation of .46 between the FC Conscientiousness scores and performance. The difference between the validity coefficients associated with the FC- and SS-based measures in the applicant instructions condition was quite large and is unlikely to be attributable to sampling error [$t(57) = 3.01, p < .05$]. Overall, in the applicant instructions condition the FC scores explained more than 26 times as much performance variance as did the SS scores, explaining an additional 18% of the variance when SS scores were partialled out using linear regression [$\Delta F(1, 57) = 4.95, p < .05$]. Thus, the pattern of correlations was consistent with H5.

In summary, the personality test using the SS format was a slightly better predictor than the FC inventory in normal instructions condition but worse in the applicant instructions condition. The data were consistent with the prediction that in the applicant instructions condition FC scores would correlate more strongly with performance than did the SS scores. Interestingly, an increase in the amount of performance variance explained by FC scores in the applicant instructions condition was observed relative to the normal instructions condition.

Discussion

This study extended past research by comparing how well scores from an FC and anormative measure predict actual job performance under similar conditions. Consistent with the Study 1, the results clearly indicated that the FC items were less susceptible to distortion and that scores from that method were better predictors of performance when test-takers had been asked to respond as if applying for a job than were scores from the SS inventory. Thus, when motivation to distort responses is high, using FC methodology to construct test items will most likely maintain validity better than using more transparent normative items.

Equally interesting was the suggestion that in some circumstances criterion validity may have been enhanced by using FC methodology. This was somewhat surprising given that past research has shown trait variance is reduced by the same manipulation used in this study. Indeed, one can argue that any evidence of score inflation suggests decay in trait variance because it is implausible that a constant value has been added to all respondents' scores. Reconciliation of the results showing that motivated distortion results in a decrease in validity (Study 1; Jackson et al., 2000) versus those demonstrating an increase (Study 2; Dyomina et al., 2003) can be found by considering differences in the criteria used.

Specifically, the criterion in Study 1 was another measure of Conscientiousness and Jackson et al. (2000) used self-reported admissions of delinquency. Because neither of these criteria would be expected to correlate with cognitive ability, when distortion resulted in a decrease in the trait variance in the FC scores, the correlations with criteria showed a corresponding decrease. However, cognitive ability is a well-known correlate of both the job performance criterion in Study 1 and the grade point criterion Dyomina et al. (2003) used. In these instances no such corresponding decrease in the correlations with these criteria was found and stronger criterion relations were actually observed as a result of applicant instructions. Therefore additional cognitive processes appear to be involved when individuals attempt to distort scores on FC measures and the effects on validity may depend on whether the outcome of interest is a correlate of cognitive ability.

STUDY 3: COGNITIVE PROCESSES IN THE MOTIVATED DISTORTION OF FC MEASURES

The finding that applicant instructions may enhance the criterion-related validity of FC measures at the same time as the validity of SS measures deteriorates raises a number of questions concerning the process involved when individuals attempt to distort responses to items using these formats. Unfortunately, little is known about the role of individual differences in the ability to fake personality measures (McFarland & Ryan, 2000), and this is especially true of FC measures. Thus, Study 3 examined the role of cognitive ability in the decision process involved when attempting to distort FC and SS personality items to understand which individuals are most successful in this context.

This research approaches this issue from the perspective that FC formats represent a more cognitively demanding task because they necessitate a more accurate representation of the personality-based requirements of the position in question. With an SS item format, all that is necessary to identify the responses that earn higher scores is to infer the direction of the relation between a single trait and the performance-based decision (e.g., "Would someone who likes to take care of other people be a better real estate agent?"). Depending on how transparent the item is, this might be a relatively easy task. However, FC items demand that respondents choose between two or more equally desirable options representing different traits. To distort their responses successfully, applicants must identify both the direction as well as the relative importance of different traits, resulting in the additional demand that the belief systems accurately reflect the rank order of all the traits concerned. For simplicity, these job stereotypes associated with individuals' beliefs about the behaviors involved in performing a job successfully and the trait-based requirements are referred to as implicit job theories (IJTs).

The conceptual basis for IJTs is rooted in research showing that the pattern of intentional distortion depends on the position when individuals are asked to respond as if applying for a job (Furnham, 1990). More specifically, research on the decision-making process involved has indicated individuals distort in a way that mirrors the ideal profile they hold for the position in question (Martin, Bowen, & Hunt, 2002) and that the occupational stereotypes relied on may not always be accurate (Mahar, Cologon, & Duck, 1995). This suggests that people have a theory as to which personality traits are most important for the job and adjust their answers accordingly.

Research on the more general topic of the implicit personality theories has found correspondence among such cognitive structures involving beliefs about traits and the results of empirical research, although differences exist among individuals with regard to the extent that their beliefs are consistent with empirical findings (Jackson, Chan, & Stricker, 1979). That is, individuals with valid implicit personality theories tend to make more accurate inferences about the cooccurrence of traits in others even when provided with small amounts of data (Lay & Jackson, 1969). Recent research has also shown that individual differences in the validity of implicit personality theories tend to be more related to cognitive ability than personality (Christiansen, Wolcott-Burnam, Janovics, Quirk, & Burns, 2005).

Figure 1 illustrates the conceptual framework that guided the research, based on the premise that those individuals with higher cognitive ability will tend to have a better idea of which traits are important for the job and therefore will be more successful at elevating their scores. In this model, IJTs guide the decision about how an individual will distort responses to personality inventories. The extent that a

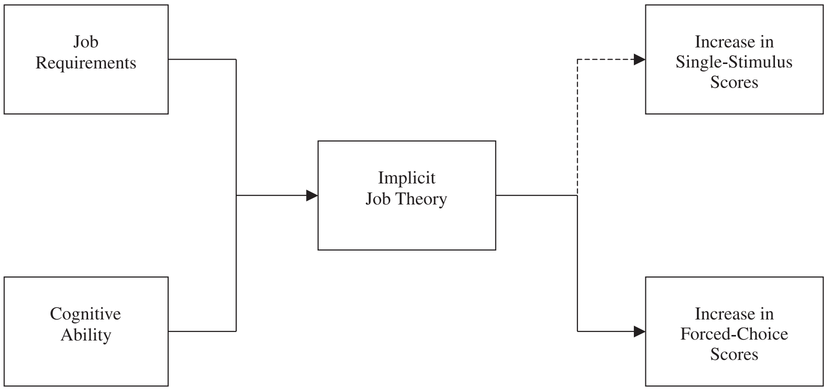
Job Requirements
Cognitive Ability
Implicit Job Theory
Increase in Single-Stimulus Scores
Increase in Forced-Choice Scores

FIGURE 1 Conceptual model of effects of applicant instructions on single-stimulus and forced-choice personality measures.

trait is believed to be important in an IJT is posited to be a function of both the position and the cognitive ability of the individual who is motivated to distort responses to the personality inventory to obtain that position.

A stronger link is thought to exist between IJTs and how much increase is expected in FC scores compared to SS scores because the former demands that IJTs be accurate in terms of both the direction a trait is related to decisions as well as the relative importance of a trait.

We expected that different IJT profiles would emerge as a result of manipulating the job description provided to participants because the position requirements of a job should elicit different stereotypes about job success. To test this, one job description was provided for the customer service job from Study 2 and one that described an assembler position in a manufacturing plant. Research has shown that both Conscientiousness and Agreeableness tend to be important indicators of success for service jobs (e.g., Frei & McDaniel, 1998; Mount, Barrick, & Stewart, 1998), whereas for skilled and semiskilled jobs without interpersonal requirements Conscientiousness emerges as the strongest correlate of performance (Barrick & Mount, 1991).

H6: The requirements of the job will affect the traits considered important for success with: (a) the job description for the customer service position activating an IJT emphasizing Conscientiousness and Agreeableness, and (b) the job description for the assembler position activating an IJT emphasizing Conscientiousness.

H7: Cognitive ability will be related to the belief that Conscientiousness is important for success across jobs.

H8: Cognitive ability and IJT scores for Conscientiousness will be related to the distortion of FC scores more than the distortion of SS scores, such that: (a) the relations between the cognitive predictors and scores on the personality test will be stronger with the applicant instructions than normal instructions; and (b) when completed under applicant instructions conditions, the relations between the cognitive predictors and scores on the personality test will be stronger with the FC format than with the SS format.

Method

Participants and Procedure

Participants in the study were 518 undergraduate students (77% female and 23% male) with an average age of 19.5 years. First, participants completed FC and SS personality measures with instructions to respond as honestly as possible. Following this, participants were randomly provided with a job description for either an assembler or a customer service representative position. Next, participants completed an IJT measure for the described position and completed a measure of cognitive ability. Finally, participants were asked to complete the personality measures again as if part of the applicant process for the described job.

Measures

Cognitive ability. General cognitive ability was measured using the WPT (Wonderlic, 1999), a frequently used commercial test consisting of 50 items with a time limit of 12 min. Reliability estimates reported in the manual range from .82 to .94 for test–retest reliability and from .88 to .94 for internal consistency. The WPT has also been shown to correlate with other well-established measures of cognitive ability; for example, scores on the WPT correlate with overall scores on the Wechsler Adult Intelligence Scale in the range of .75 to .96 (Wonderlic, 1999).

Implicit job theory. A measure of IJTs was developed and adapted to customer service and assembler positions. Participants read the relevant job description along with a list of 60 adjectives and were asked to identify the 30 that were the most important for a person employed in the described position to possess. This list of adjectives was based on 12 positive trait terms from each dimension of the Five-factor-model of personality (FFM). Scores were computed by summing the number of times adjectives were chosen that were related to each personality construct. The format of this measure did not allow for internal consistency estimates to be computed.

SS personality measure. The 20-item personality inventory used in Study 2 from IPIP (Goldberg, 1997) was shortened to 15 items due to time constraints. Participants responded to the 15 items using the same 5-point scale (anchored from *very inaccurate* to *very accurate*). The internal consistency of the SS inventory was .86 when completed under instructions to answer honestly and .86 when completed under instructions to answer as if applying for the described job.

FC personality inventory. Five items were added to the 20-item FC measure of conscientious used in Study 2 in an effort to improve internal consistency. The resulting internal consistency of the FC inventory was .78 when completed under instructions to answer honestly and .74 when completed under instructions to answer as if applying for the described job.

Results

Differences in Implicit Job Theories by Position

The initial step in analyzing the results of Study 3 was to compare mean IJT profiles obtained from the two job descriptions to determine whether different traits were believed to be important depending on the position. Table 6 shows descriptive statistics and statistical comparisons. A 2 (Job Description) × 5 (FFM Trait) mixed-model ANOVA was conducted on the mean number of times adjectives from each trait were chosen as being associated with a successful employee for a position. The main effect for Job Descriptions was significant [F(1, 516) = 4.56, *p* < .01], as was the main effect associated with the FFM Trait that the adjectives chosen belonged to [F(4, 2064) = 585.28, *p* < .01]. Consistent with the idea that the relative importance of FFM Traits would vary by Job Description (i.e., the two job descriptions elicit different implicit job theory structures), the interaction was also significant [F(4, 2064) = 155.69, *p* < .01].

Two sets of planned comparisons were conducted to test H6a and H6b, one set within each position and one set across positions. Consistent with predictions, within the assembler position conditions Conscientiousness was believed to be more important for the assembler job than the other four traits [F(1, 259) = 3511.18, *p* < .01], and within the customer service position condition Conscientiousness and Agreeableness were judged as more important than the other three traits [F(1, 257) = 1290.32, *p* < .01]. Across conditions Conscientiousness was believed to be more important for the assembler job than the customer service position [F(1, 517) = 117.38, *p* < .01], whereas Agreeableness was considered more important for the customer service job than the assembler position [F(1, 517) = 546.9, *p* < .01]. Taken together, the results are consistent with H6 that predicted the job descriptions would elicit different trait profiles associated with success in the two positions.

TABLE 6
Mean Implicit Job Theory Scores by Trait and Job Description

	Description					
	Assembler		*Customer Service*		*Comparison*	
Implicit Job Theory Trait	*M*	*SD*	*M*	*SD*	*d*	*F*
Conscientiousness	9.91	1.52	8.34	1.77	.95	117.38*
Agreeableness	3.12	1.85	7.02	1.95	–2.05	546.92*
Emotional Stability	6.69	1.68	6.16	1.75	.31	12.30*
Openness	3.93	2.55	2.06	1.94	.83	87.98*
Extraversion	5.11	1.89	5.34	1.91	–.12	1.87

Note. *n* = 260 in the assembler condition and *n* = 258 in the customer service condition. *d* refers to the standardized mean difference obtained by subtracting the mean of the customer service condition from that of the assembler condition and dividing by the pooled standard deviation; *F* indicates the test statistic at 1 and 516 *df*s.

**p* < .01.

Cognitive Ability and Implicit Job Theories

After demonstrating that the provided job description changed IJT profiles, the next step was to examine the relation between cognitive processes and IJTs. H7 predicted that individuals with higher cognitive ability would be more likely to identify Conscientiousness as important for each job. Results indicated that a relation exists between cognitive ability and perceived importance of Conscientiousness ($r = .20$) collapsed across jobs, with a slightly higher correlations observed for the assembler job (.23) than the customer service position (.20).

Predictors of SS and FC Personality Scores

The relations between the cognitive predictors (cognitive ability and IJT) and personality test scores completed under applicant instructions were hypothesized to be a function of the format of the personality measure. Specifically, we predicted that those individuals with higher cognitive ability and who believed Conscientiousness was important would be more successful at inflating scores, but that this would be particularly true for the FC scores. Table 7 details the relations of cognitive ability and IJT scores with personality scores in from the normal and applicant instructions.

A review of Table 7 indicates that both cognitive ability and IJT Conscientiousness scores had stronger relations with personality scores when they were completed under applicant instructions rather than under the normal instructions with

TABLE 7
Correlations Between Cognitive Predictors of Applicant Distortion and Conscientious Measures

					Instructions			
					Normal		*Applicant*	
Measures	*M*	*SD*	*CA*	*IJT-C*	*SS*	*FC*	*SS*	*FC*
Cognitive ability	25.63	4.93	1.00					
IJT-Conscientiousness	9.13	1.82	.20*	1.00				
Conscientiousness measures								
Normal single-stimulus	54.56	8.69	.01	.03	1.00			
Normal forced-choice	12.12	4.70	.02	.06	.61*	1.00		
Applicant single-stimulus	67.54	7.31	.15*	.25*	.18*	.14*	1.00	
Applicant forced-choice	17.65	3.75	.25*	.52*	.19*	.31*	.33*	1.00

Note. $N = 518$. CA = cognitive ability; IJT-C = implicit job theory and refers to scores based on individuals' belief about how important conscientiousness is for job success; SS = single-stimulus measure of Conscientiousness; FC = forced-choice measure of Conscientiousness.

*$p < .05$.

all of the former correlations being significant and none of the latter. For the SS items, the correlation between cognitive ability and personality scores was stronger with the applicant instructions ($r = .15$) than with the normal instructions (.01) [t(515) = 2.73, $p < .05$], as was the correlation between IJT and personality scores with applicant instructions (.25) compared to normal instructions (.03), [t(515) = 4.27, $p < .05$]. A similar but stronger pattern was observed for FC items where the correlation between cognitive ability and personality scores was stronger with applicant instructions ($r = .25$) than with honest instructions (.02), [t(515) = 4.11, $p <$ .05], as was the correlation between IJT and personality scores in applicant instructions (.52) compared to honest instructions (.06), [t(515) = 9.38, $p < .05$], thereby supporting H8a.

Within the applicant instructions condition the correlation between cognitive ability and personality scores was stronger with the FC items (.25) than with the SS items (.15), [t(515) = 2.05, $p < .05$]. Similarly, with applicant instructions the correlation between IJT and personality scores was stronger with FC items (.52) than with SS items (.25), [t(515) = 6.13, $p < .05$]. This was consistent with H8(b). The observation that the difference in relations was stronger for IJT than cognitive ability suggests it may be a more proximal mechanism responsible for inflation of scores.

Also instructive is considering the correlations between the Conscientiousness measures across instructions and item formats. Three observations are noteworthy. First, a stronger level of convergence between the FC and SS measures was observed with normal instructions ($r = .61$) than with applicant instructions ($r = .33$). Second, the correlation between the applicant instructions FC scores and the normal instructions FC scores ($r = .31$) was greater than the correlation of the applicant SS scores with either the SS ($r = .18$) or FC ($r = .14$) scores with normal instructions. This is consistent with Study 1 results where we found that the FC measures contained more trait-relevant variance than the SS measures in the applicant instructions condition. Third, the correlation between the applicant instructions FC scores with the normal instructions FC scores (.31) was slightly larger than the correlation between the applicant FC scores and cognitive ability (.25). This suggests that with applicant instructions the FC scores contain variance attributable both to the intended trait as well as to cognitive ability (and that more of the former may exist).

Mediating Role of Implicit Job Theories

A path analysis was conducted to investigate the intervening role of IJTs as a mechanism whereby cognitive ability is related to the increase in scores that result from the applicant instructions. In these analyses, the relations between cognitive ability scores, IJT scores for Conscientiousness, increase in SS scores, and increase in FC scores were decomposed using casual modeling. Because of well-known problems associated with the use of raw difference scores, increase in scores on the personality measures were computed using regression-adjusted difference scores (cf. Pedhazur & Schmelkin, 1991). As a preliminary step, this model was fit to the participants from the assembler and customer service conditions using multisample analysis. Figure 2 shows the paths from the common metric solution for this model, which are shown in parentheses. As can be seen, all of the paths within each sample were positive and in a consistent direction. Constraining each path to be equivalent across the samples did not result in any significant decrements to fit, nor did imposing equality constraints on all three direct paths simultaneously, $\Delta\chi^2(3) = 2.58$, *ns*.

Figure 2 also shows the paths from the total sample. Fit of the model without direct paths from cognitive ability to the increases in personality scores was favorable [$\chi^2(2) = 18.01$, Comparative Fit Index = .94, Goodness of Fit Index = .98, Adjusted Goodness of Fit Index = .92]. Consistent with what would be expected based on H8, the relation between IJT and increases in FC scores (.51) was considerably stronger than was observed for increases in SS scores (.25). Adding direct paths from cognitive ability to the amount of increase in the SS and FC scores resulted in coefficients of .11 and .15, respectively. Given that these values are less than the zero-order correlations of .15 and .25, the mediat-

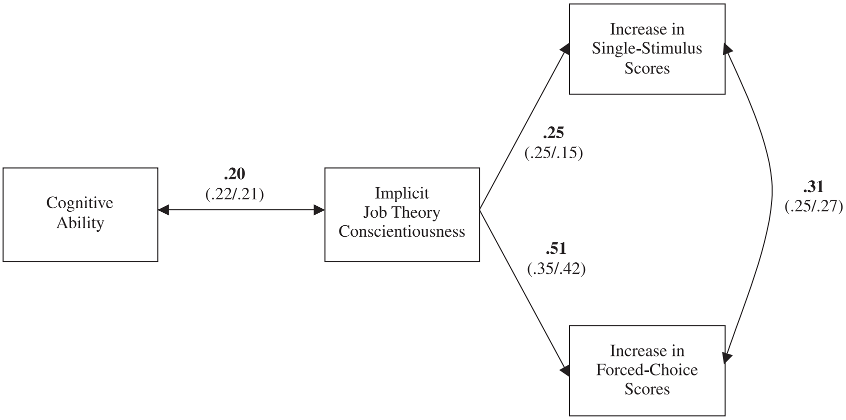
Cognitive Ability
.20
(.22/.21)
Implicit Job Theory Conscientiousness
.25
(.25/.15)
Increase in Single-Stimulus Scores
.51
(.35/.42)
Increase in Forced-Choice Scores
.31
(.25/.27)

FIGURE 2 Path analysis of effects of cognitive ability on increase in scores on single-stimulus and forced-choice measures of Conscientiousness. Path coefficients shown in parentheses are from the common metric solution of the assembler ($n = 260$) and customer service ($n = 258$) conditions, those outside parentheses are from the total sample solution ($N = 518$). All paths are significant at the .05 level. Constraining the paths across conditions to be equivalent in multisample analysis did not result in a significant change in fit, $\Delta\chi^2(3) = 2.58$, *ns*. Model fit for the total sample: $\chi^2(2) = 18.01$, Comparative Fit Index = .94, Goodness of Fit Index = .98, Adjusted Goodness of Fit Index = .92.

ing role of IJT Conscientiousness scores is supported. In summary: (a) the model without direct paths fits reasonably well; (b) the paths in the model linking cognitive ability, IJT, and personality score increases were sizeable; (c) accounting for IJT scores resulted in a decrease in the relation between cognitive ability and the amount of increase in scores on the personality measures; and (d) after controlling for IJT the remaining relations between cognitive ability and the increases in personality scores were small but potentially meaningful. Taken together, the role of IJTs would be best described as consistent with partial mediation; there may be additional reasons cognitive ability is related to the amount of score increase that results from applicant instructions.

Discussion

Study 3 was conducted to clarify the role of cognitive processes in the distortion of FC measures when individuals respond in a way intended to improve their chances of getting a job offer. The results showed that cognitive ability was related to both SS and FC scores when participants were instructed to respond as if applying for the jobs described, but that the relations were stronger for FC items. Although cognitive ability explained less than one-tenth of 1% of the variance in the FC Conscientiousness scores when completed with instructions to respond honestly, cognitive ability explained more than 6% of the variance in scores when participants were instructed to respond as job applicants. Thus, the results of Study 3 support the idea that distortion of FC measures to improve the chances of being hired is a more cognitively demanding task and helps explain why enhanced validity due to applicant instructions has been found with some criteria but not others.

The results were also instructive with regard to why cognitive ability is required for successful distortion of FC measures in applicant contexts. Specifically, the belief that the keyed trait in the FC inventory was important for job success was shown to be cognitively driven. Those low in cognitive ability were less likely to consider Conscientiousness to be the most important trait and as a result did not elevate their scores as much on the FC inventory. On the other hand, those higher in cognitive ability were more likely to believe Conscientiousness was critical for job success and consequently were more capable of elevating their scores when motivated to do so. Although individuals' IJT about the importance of Conscientiousness was related to their ability to distort the SS items, when instructed to respond as applicants, the IJT Conscientiousness scores explained more than four times as much variance in FC scores than they did in SS scores. This is consistent with the proposition that successful distortion of FC items is a more cognitively demanding task because it requires a more accurate theory about the personality-based requirements of the job.

GENERAL DISCUSSION

The criticisms leveled at FC methods have largely resulted in the discrediting of these inventories for personality assessment. This research evaluated the recommendation that personnel psychologists reconsider the use of FC inventories when motivation to distort is high. Consistent with past research, all three studies demonstrated that SS scores are more susceptible to inflation associated with an applicant response set than were the inventories fashioned using FC methodology. All three studies also showed some convergence between the two methods at a level suggesting that they measure related (but not necessarily the same) constructs. Finally, it was demonstrated that the ability of a SS measure to predict job performance was attenuated by distortion whereas the FC measure's predictive validity actually increased. Therefore, the FC methodology appears to offer some advantages over normative alternatives in selection contexts.

However, we do have some reasons for caution. We found that motivation to distort responses resulted in complex changes to the construct validity of the FC personality measures. Consistent with Dyomina et al.'s (2003) recent findings, when participants were instructed to complete an FC inventory as if applying for a job, the scores became increasingly related to cognitive ability. This is not to say that scores were no longer related to Conscientiousness when participants were instructed to respond as applicants, but rather that the resulting scores were a function both of the intended trait as well as cognitive ability. Given this complexity, caution is particularly important when inferring the validity of applicant scores from the results of concurrent validation studies using incumbents who presumably have less motivation to distort.

Implications for Practice

One implication of the relation between applicant FC scores and cognitive ability is that although personality measures are often considered to be relatively free of adverse impact based on ethnic group membership, our research suggests this may not hold when an FC format is employed. Research that has examined effects of combining personality and cognitive measures has shown that when even a modest amount of variance attributable to cognitive ability is present, the result is minorities being underselected at an alarming rate (Ryan, Ployhart, & Friedel, 1998). If confirmed in actual job applicants, the relation between cognitive ability and FC scores is most likely to result in adverse impact for such personality measures at lower selection ratios.

Another implication is that incremental validity associated with adding a personality measure will most likely be reduced to the extent that other cognitive tests are included as part of the battery. One of the advantages to personality tests is that they are relatively uncorrelated with cognitive tests (Ackerman & Heggestad,

1997) and therefore have potential to explain additional criterion variance (e.g., Schmidt & Hunter, 1998). As the correlation between the FC scores and cognitive ability increases, less incremental validity is likely to result (and therefore less utility) when combined with other assessment that are also cognitively loaded such as structured interviews (Cortina, Goldstein, Payne, & Gilliland, 2000). However, given that the FC scores obtained under applicant instructions correlated better with the assessment of Conscientiousness with instructions to respond honestly than the SS scores did in both Study 1 and Study 3, one might expect that some incremental validity would be associated with the FC scores beyond what would be expected with the normative alternatives.

Finally, the results underscore that the effects of distortion on criterion-related validity will most likely depend on the outcome of interest. When the outcome is one that is a known correlate of cognitive ability (e.g., training proficiency), FC measures of personality may maintain or enhance validity. However, many outcomes of interest when validating personality measures are not cognitively loaded, such as extrarole or organizational citizenship behavior. Taken together, the results of the three studies suggest that with such criteria the validity coefficients of FC measures will probably be attenuated by motivated distortion but at a rate less than what would be observed for SS measures.

Distortion of FC Measures

From these studies, two general observations can be made regarding the susceptibility of FC and SS format inventories to applicant distortion that place the results in context. First, when the applicant response set was absent, the SS inventories were better indicators of the intended trait, displaying both superior reliability and validity. In many research and practical situations where motivation to distort is less, normative assessments should still be preferred. This is particularly salient given that FC inventories involve additional resources to develop. However, in situations where sufficient motivation for distortion exists, FC inventories should be considered given that they offer greater control over response bias. Second, although the scores from the FC method were better construct indicators in the applicant instructions condition in Study 1, the effects of distortion on convergent validity were still appreciable.

Therefore, room for improvement on control of response bias for FC measures used in applicant contexts exists. Study 2 attempted to reduce applicant distortion further by using three response options rather than the dyads employed in Study 1, making identifying the keyed response more difficult and hence improving validity with more motivated test-takers (Zavala, 1965). Such a finding is suggested by comparing the mean shift of Conscientiousness scores from the normal and applicant instructions conditions across the two studies. Specifically, in Study 2 scores of Conscientiousness only increased by one quarter of a standard deviation whereas they increased by almost one half of a standard deviation in Study 1. Investigating the use of four or even five response options may be worthwhile if the goal is to minimize distortion.

It is also possible to view the failure of the FC to eliminate mean shift due to applicant instructions as a breakdown in the attempt to match response options on attractiveness. Using stricter matching criteria or using different data for attractiveness might reduce the amount of distortion observed. For example, rather than ratings of desirability, matching could be based on item means from SS formats collected from the personality assessment of simulated or real applicants. Another possibility is to make the keyed response less desirable rather than equally desirable, using whichever index of attractiveness is used for matching. Consider that matching is generally done based on mean attractiveness ratings to make response options similar. But because some individuals are invariably above the mean and therefore will believe that an option is more desirable, some mean shift is probably inevitable when these individuals become motivated to try to discern the keyed response. Therefore, reducing successful distortion by making sure the keyed response is always lower in desirability than the other options may be possible. Given the complexity of the effects on construct validity of asking participants to respond as applicants, to the extent such strategies are effective at minimizing successful distortion, they might also change construct validity and would need to be evaluated carefully.

Effects of Distortion on Validity

The results from this research that demonstrate validity decay when participants are asked to respond as if applying for a job were at odds with the assertions of some researchers that applicant distortion does not meaningfully affect criterion-related validity (e.g., Ones & Viswesvaran, 1998; Ones et al., 1996). Importantly, considering the methodology used in the research that these assertions are based on, namely studies that have found little effect when SD scores are considered as a suppressor (e.g., Barrick & Mount, 1996; Christiansen, Goffin, Johnston, & Rothstein, 1994; Ones et al., 1996) or a moderator (Hough, 1997) of the personality–performance relation. The studies that have used this strategy have several limitations, which taken together suggest the conclusion that applicant distortion does not affect validity may be premature.

First, the inference rests on the use of SD scores as a reliable and valid measure of applicant distortion. Careful examination of the research base suggests that although somewhat sensitive to distortion, the relation between SD scores and applicant distortion may not be very strong. Consider that the mean shift in SD scores that results from instructing a group to fake good answers compared to one instructed to respond honestly is approximately 1 standard deviation (Viswesvaran

& Ones, 1999). Results of laboratory faking studies have been shown to be an upperbound or overestimate of what takes place in real applicant samples (e.g., Smith & Ellingson, 2002). In other words, at most 20% of the variance in SD scores (based on $d = 1.0$) is explained by applicant distortion, amounting to a correlation of approximately .45.

Second, it is important to consider that the vast majority of the studies that base their inferences on SD scores (such as most included in the Ones et al., 1996 meta-analysis) have not used personality scores from actual job applicants or even individuals asked to respond as if applying for a job. Whether significant amounts of motivated distortion exist in either the SD scores or the personality test scores of participants (typically students or incumbents) who have no real incentive to dissimulate is questionable. This in turn begs the question of whether in these contexts any compromise to validity had occurred to be uncovered by partialing SD scores or by considering them as a moderator.

Finally, for a suppressor variable to improve prediction appreciably, the respective zero-order relation must be stronger than the predictor-criterion relations that are typically observed between personnel selection tests and job performance (Conger & Jackson, 1972). Thus, even if SD measures were more strongly related to applicant distortion and if personality had been assessed in a context where motivated distortion was likely, finding little or no improvement to validity as a result of partialing was a foregone conclusion. Asserting the null that applicant distortion does not affect criterion-related validity based on these results therefore would seem to represent the sort of hypothesis testing that the scientific method long ago rejected because it cannot be falsified.

In contrast, research using methodology that compares individuals that differ in their motivation to distort responses to personality tests typically has found differences in the validity of their scores. For example, deterioration in the criterion-related validity of personality test scores is observed when an incentive to do well on the test is provided to one group but not another (Douglas et al., 1996; Mueller-Hanson et al., 2003). Even among those with the same incentive to do well on the personality test, criterion-related validity of more motivated test-takers was worse than those with less motivation with the exact opposite pattern being observed for a cognitive test (Schmit & Ryan, 1992). Finally, a comparison of validity studies using applicants and incumbents reveals that the validity of applicants' scores tends to be worse (Hough, 1997), and other research confirms that applicants are much more motivated test-takers than nonapplicants (Arvey, Strickland, Drauden, & Martin, 1990). Given a general agreement that applicant distortion occurs (e.g., Donovan et al., 2003), we conclude that much more research is needed before the conclusion is embraced that applicant distortion does not affect criterion-related validity. In the meantime, developing strategies to combat the effects of motivated distortion would seem prudent.

Limitations and Future Directions

The most obvious limitation to this research is that all three studies were simulations using student samples rather than actual job applicants. Field research that examines the construct validity of FC measures using applicants and nonapplicants is needed to determine whether the results generalize. Given that research suggests that directed faking studies generally overestimate the effects of distortion compared to actual selection contexts (Smith & Ellingson, 2002), the advantages of FC in such contexts might be expected to be less than that reported herein. In addition, only a limited number of fairly broad traits were considered and future research should focus on whether the effects on construct validity observed here extend to other traits and in particular those of more narrow bandwidth than are found in the FFM.

Future research should also consider response process models more specifically tailored to the task involved in making FC judgments. Item response theory and classical test theory approaches generally assume that responses are a function of one (and only one) latent trait factor; this is probably not the case with FC formats. Response options may trigger an evaluation of whether the latent true score associated with one response is higher than the others. On the other hand, choosing a given response is a function of the discrepancy between each true score of the individual and the placement of the responses on those same measures (e.g., θ in terms of item response theory). However, this assumes that an item being "too true" has the same value as being "not true enough," which may be unrealistic given what is known of self-serving biases. Although models of responses to FC items will be necessarily complicated, the results of studies such as those presented here suggest that a better understanding of this format is warranted.

Finally, the results of Study 3 suggest that occupational stereotypes play an important role in the intentional distortion of personality measures that is not included in current models of faking (e.g., McFarland & Ryan, 2000; Snell, Sydell, & Lueke, 1999). This is similar to the findings of research that have shown "job desirability" may be more important than "social desirability" when respondents attempt to distort biodata items to tailor responses to the specific job being applied for (Kluger & Colella, 1993). One should note that this perspective suggests applicant distortion may be conceptually more similar to guessing on a multiple-choice test than an active attempt to manage impressions as is done in social interactions. A final limitation to note is that the method of assessing IJT used in this study, as well as past approaches to measuring the ideal job profiles used in the past (e.g., Martin et al., 2002), may not be optimal. Rather than a direct focus on traits of successful employees, a strategy for future research would be to assess different occupations systematically based on how frequently behaviors with known trait linkages are performed. Comparing responses of job experts to those of laypeople to map out the normative beliefs about occupations and their accuracy would then be possible. This could be done using existing personality-based job analytic instruments such as the Personality-Related Position Requirements Form (Raymark et al., 1997).

CONCLUSION

A goal when assessing the personality of job applicants might be to obtain the same responses as would have been obtained in a less consequential assessment without the situational press found in selection contexts. Ideally, this would prevent anything but the construct of interest from affecting scores and only those with more favorable trait elevations would choose the keyed response. However, similar to many ideals, this may not be practical using self-report methodology. A more realistic goal may therefore be to prevent any factors that are not job-related from affecting trait scores, such that only those individuals with more favorable trait elevations or who possess some other important work requirement (such as cognitive ability or job knowledge) would be likely to choose the keyed response. Although less than the ideal, FC measures may be better indicators of the applicant's personality profile than would have been obtained under more honest conditions than could be obtained using their normative counterparts.

ACKNOWLEDGMENTS

Portions of this article were presented at the 13th (April 1998, Dallas, TX) and 15th (April 2001, New Orleans, LA) annual conferences of the Society for Industrial and Organizational Psychology.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 12,* 219–245.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* New York: MacMillan.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69,* 49–56.

Barrick, M. B., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Barrick, M. B., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 42,* 261–272.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational & Organizational Psychology, 69,* 25–39.

Bass, B. M. (1957). Faking by sales applicants of a forced-choice personality inventory. *Personnel Psychology, 41,* 403–404.

Berkshire, J. R. (1958). Comparison of five forced-choice reference check. *Educational and Psychological Measurement, 18,* 553–561.

Bernardin, H. J. (1989). Innovative approaches to personnel selection and performance appraisal. *Journal of Management Systems, 1,* 25–36.

Borislow, B. (1958). The Edwards Personal Preference Schedule and fakibility. *Journal of Applied Psychology, 42,* 22–27.

Caroll, S. J., & Nash, A. N. (1972). Effectiveness of a forced-choice reference check. *Personnel Administration, 35,* 42–146.

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.

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J., Quirk, S., & Burns, G. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance, 18,* 123–149.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement, 32,* 579–599.

Cook, M. (1993). *Personnel selection and productivity* (Rev. ed.). New York: Wiley.

Corah, M. L., Feldman, M. J., Cohen, I. S., Meadow, A., & Ringwall, E. A. (1958). Social desirability as a variable in the Edwards Personal Preference Schedule. *Journal of Consulting and Clinical Psychology, 22,* 70–72.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville and Willson (1992). *Journal of Occupational and Organizational Psychology, 67,* 89–100.

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53,* 325–351.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). *The Revised NEO-PI/NEO-FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

DeVito, A. J. (1985). Review of the Meyers-Briggs Type Indicator. *Ninth Mental Measurements Yearbook, 2,* 1030–1032.

Dicken, C. F. (1959). Simulated patterns of the Edwards Personal Preference Schedule. *Journal of Applied Psychology, 43,* 372–378.

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

Douglas, E. F., McDaniel, M. A., & Snell, E. F. (1996, August). The validity of non-cognitive measures decays when applicants fake. *Proceedings of the Academy of Management,* 127–131.

Dunn, W. S., Mount, M. K., Barrick, M. R., & Ones, D. S. (1995). Relative importance of personality and general mental ability in managers' judgments of applicant qualifications. *Journal of Applied Psychology, 80,* 500–509.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15,* 13–24.

Dyomina, N. V., Vasilopoulos, N. L., Cucina, J. M., & Reilly, R. R. (2003, April). *Forced-choice personality tests: A measure of personality or "g"?* Poster session presented at eighteenth annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Edwards, A. L. (1959). *Edwards Personal Preference Schedule manual.* New York: Psychological Corporation.

Feldman, M. J., & Corah, M. L. (1960). Social desirability and the forced-choice method. *Journal of Consulting Psychology, 24,* 480–482.

Frei, R. L., & McDaniel, M. A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11,* 1–27.

French, E. G. (1958). A note on the Edwards Personal Preference Schedule for use with basic airman. *Educational and Psychological Measurement, 18,* 109–115.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology: Research & Reviews, 9,* 46–55.

Graham, W. R. (1958). Social desirability and forced-choice methods. *Educational and Psychological Measurement, 18,* 387–401.

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

Goldberg, L. R. (1992). The development of markers for the Big-Five structure. *Psychological Assessment, 4,* 26–42.

Goldberg, L. R. (1997). International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences. Retrieved August 1999 from http://ipip.ori.org/ipip/

Gordon, L. V. (1963). *Revised manual for the Gordon Personal Profile.* New York: Harcourt.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167–184.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist, 51,* 469–477.

Hough, L. M. (1997). Personality at work: Issue and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131–166). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities (Monograph). *Journal of Applied Psychology, 75,* 581–585.

Izard, C. E., & Rosenberg, N. (1958). Effectiveness of a forced-choice leadership test under varied experimental conditions. *Educational and Psychological Measurement, 18,* 57–62.

Jackson, D. N. (1984). *Personality Research Form manual*. Port Huron, MI: Research Psychologists.

Jackson, D. N., Chan, D. W., & Stricker, L. J. (1979). Implicit personality theory: Is it illusory? *Journal of Personality, 47,* 1–10.

Jackson, D. N., Neill, J. A., & Bevan, A. R. (1973). An evaluation of forced-choice and true–false item formats in personality assessment. *Journal of Research in Personality, 7,* 21–30.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13,* 371–388.

Johnson, C. E., Wood, R., & Blinkhorn, S. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61,* 153–162.

Kirchner, W. K. (1962). Real-life faking on the Edwards Personal Preference Schedule by sales applicants. *Journal of Applied Psychology, 46,* 128–130.

Kluger, A. N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology, 46,* 763–780.

Krug, R. W. (1958). A selection set preference index. *Journal of Applied Psychology, 32,* 89–92.

Lanyon, R. I. (1966). A free-choice version for the EPPS. *Journal of Clinical Psychology, 22,* 202–205.

Lay, C. H., & Jackson, D. N. (1969). Analysis of the generality of trait-inferential relationships. *Journal of Personality & Social Psychology, 12,* 12–21.

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

Maher, H. (1959). Studies of transparency in forced-choice scales: I. Evidence of transparency. *Journal of Applied Psychology, 63,* 275–278.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality & Individual Differences, 32,* 247–256.

McDaniel, M. A., Douglas, E. F., & Snell, A. F. (1997, April). *A survey of deception among job seekers.* Paper presented at the twelfth annual conference of the Society of Industrial and Organizational Psychology, St. Louis, MO.

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

Mount, M. K., Barrick, M. B., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11,* 145–165.

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.

Norman, W. T. (1963). Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology, 47,* 225–241.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11,* 245–269.

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

Paajanen, G. E., Hansen, T. L., & McLellan, R. A. (1993). *PDI Employment Inventory and PDI Customer Service Inventory manual.* Minneapolis, MN: Personnel Decisions.

Paulhus, D. L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding.* Unpublished manual, University of British Columbia, Vancouver, Canada.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes*. San Diego, CA: Academic.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach* (student ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50,* 723–736.

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

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83,* 298–307.

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 findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality & Social Psychology, 86,* 162–173.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77,* 629–637.

SHL. (1998). *OPQ 4.2 concept*. Boston, MA: Author.

Smith, B. D., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87,* 211–219.

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

Stanush, P. L. (1997). *Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation*. Unpublished doctoral dissertation, Texas A&M University, College Station, TX.

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.

Travers, R. M. (1951). A critical review of the forced-choice technique. *Psychological Bulletin, 48,* 62–70.

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 theater personnel. *Personnel Psychology, 47,* 73–90.

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

Waters, L. K. (1965). A note of the “fakability” of forced-choice scales. *Personnel Psychology, 16,* 187–191.

Wonderlic. (1999). *Wonderlic personnel test and scholastic level exam*. Libertyville, IL: Author.

Zavala, A. (1965). The development of the forced-choice rating technique. *Psychological Bulletin, 63,* 117–124.