Journal of Applied Psychology
2007, Vol. 92, No. 5, 1270–1285

0021-9010/07/$12.00 DOI: 10.1037/0021-9010.92.5.1270

Personality Measurement, Faking, and Employment Selection

Joyce Hogan, Paul Barrett, and Robert Hogan
Hogan Assessment Systems

Real job applicants completed a 5-factor model personality measure as part of the job application process. They were rejected; 6 months later they (n = 5,266) reapplied for the same job and completed the same personality measure. Results indicated that 5.2% or fewer improved their scores on any scale on the 2nd occasion; moreover, scale scores were as likely to change in the negative direction as the positive. Only 3 applicants changed scores on all 5 scales beyond a 95% confidence threshold. Construct validity of the personality scales remained intact across the 2 administrations, and the same structural model provided an acceptable fit to the scale score matrix on both occasions. For the small number of applicants whose scores changed beyond the standard error of measurement, the authors found the changes were systematic and predictable using measures of social skill, social desirability, and integrity. Results suggest that faking on personality measures is not a significant problem in real-world selection settings.

Keywords: personality measurement, faking, impression management, personnel selection

There are two major criticisms of the use of personality measures for employee selection, both of which are, in principle, amenable to empirical resolution. The first is that personality measures are poor predictors of job performance (Murphy & Dzieweczynski, 2005). This criticism persists despite evidence showing that well-constructed measures of personality reliably predict job performance, but with no adverse impact (J. Hogan & Holland, 2003; R. Hogan, 2005). The second criticism is that job applicants distort their scores by faking. Beginning with Kelly, Miles, and Terman (1936), the vast literature on this topic refers to faking with a variety of terms. We use the term *impression management* to refer to the process of controlling one's behavior during any form of social interaction, including responding to inventory items.

There are two views regarding how impression management affects personality measures. One view is that people engage in impression management on specific occasions—for example, when applying for a job—and doing so inevitably degrades test validity. The second view is that, during social interaction, most people behave in ways that are intended to convey a positive impression of themselves (Schlenker & Weigold, 1992). They do this whether reacting to questions in an employment interview, to assessment center exercises, or to items on a personality inventory—and this impression management has minimal consequences for predictive validity.

Hough and Furnham (2003) and Smith and Robie (2004) have carefully reviewed the vast and complex faking literature; their reviews can be summarized in terms of four points. First, when instructed, some people can alter their personality scores as compared with their scores when not so instructed (Barrick & Mount, 1996; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Mersman & Shultz, 1998). In addition, mean score differences are larger in laboratory faking studies than in applicant studies (Hough et al., 1990). Hough and Furnham concluded that impression management has minimal impact on employment outcomes, although Mueller-Hanson, Heggestad, and Thornton (2003) and Rosse, Stecher, Miller, and Levin (1998) disagreed.

Second, in several studies researchers have concluded that the base rate of faking in the job application process is minimal (Dunnette, McCartney, Carlson, & Kirchner, 1962; Hough, 1998; Hough & Ones, 2001). Unfortunately, it is hard to know how to assess the base rate of faking—a problem that may be logically intractable. One potential solution would be to compare a person's score on the same measure twice, the first time when applying for a job and the second time having failed the measure on the first occasion.

Third, impression management seems not to affect criterion-related validity. Ones, Viswesvaran, and Reiss (1996) conducted a meta-analysis of correlations between personality measures and job performance, after partialing out social desirability from the predictors. They concluded that social desirability does not moderate the validities of personality measures in real-world settings, and they recommended against correcting for impression management in personnel selection. Ellingson, Smith, and Sackett (2001) and Schmitt and Oswald (2006) have echoed these conclusions. Similarly, Piedmont, McCrae, Riemann, and Angleitner (2000) noted that using validity scales to correct possible biases in personality scores ignores the extent to which high scores may actually be valid. They argued that individual differences in socially desirable responding is a substantive personality variable; hence, correcting for these differences reduces valid interindividual variability.

Another way to evaluate the effects of impression management on personality measurement is to compare the factor structure of a measure completed by a "normal" sample with the factor structure of that measure completed by a sample asked to fake. Ellingson,

Joyce Hogan, Paul Barrett, and Robert Hogan, Hogan Assessment Systems, Tulsa, Oklahoma.

We thank Scott Davies, Jeff Facteau, and Lewis Goldberg for their valuable suggestions on earlier versions of this article.

Correspondence concerning this article should be addressed to Joyce Hogan, Hogan Assessment Systems, 2622 East 21st Street, Tulsa, OK 74114. E-mail: jhogan@hoganassessments.com

Sackett, and Hough (1999) showed that when this is done, the factor structure of the personality measure for the faking sample collapses. On the other hand, when samples are compared whose known level of social desirability is different, the factor structure of their scores stays the same (cf. Ellingson et al., 2001; Marshall, DeFruyt, Rolland, & Bagby, 2005; Smith & Ellingson, 2002; Smith, Hanges, & Dickson, 2001). Unlike laboratory studies of faking, ordinary everyday impression management has little influence on the factor structure of measures of normal personality.

The fourth point concerns how to reduce faking on personality inventories. Many methods have been proposed, including instructional warnings (Dwight & Donovan, 2003; Smith & Robie, 2004), forced-choice item formats (Christiansen, Edelstein, & Flemming, 1998; Heggestad, Morrison, Reeve, & McCloy, 2006; Jackson, Wrobleski, & Ashton, 2000), subtle versus overt content items (Worthington & Schlottmann, 1986), social desirability corrections (Ellingson et al., 1999), and applicant replacement (Schmitt & Oswald, 2006). None of these solutions appears to work very well. Dwight and Donovan (1998) found that warnings against "faking" reduced distortion by about 0.23 standard deviations; they concluded that warnings have small effects and that different warnings are differentially effective. In fact, Arkin (1981) cautioned that warning participants may introduce systematic biases rather than reduce response distortion. Snell (2006) suggested that disingenuous warnings are unethical. A. L. Edwards (1957) recommended pairing response alternatives based on similar social desirability weighting. The success of this method is inconclusive, perhaps because of the ipsative scoring of most such measures. Some people still believe that forced-choice formats control impression management (Christiansen et al., 1998; Jackson et al., 2000; White & Young, 1998); others are more skeptical (Bartram, 1996; Heggestad et al., 2006).

We want to add one more generalization to this list. At no point in the history of faking research has a study used a research design that is fully appropriate to the problem. We need data from actual job applicants, in a repeated measures design, where applicants are encouraged to improve their scores on the second occasion. Earlier research consists primarily of (a) laboratory studies, artificial conditions, and student research participants—see Smith and Robie (2004) for a critique; (b) between-subjects designs with no retest data to evaluate score change; and (c) studies that mix real-world and artificial instructions to create honest versus faking conditions. These designs compromise the inferences that can be drawn from the results. Abrahams, Neumann, and Githens (1971) concluded that "simulated faking designs do not provide a particularly appropriate estimate of what occurs in selection, instead they provide only an indication of how much a test *can* be faked" (p. 12), and we agree.

The Present Study

No research has evaluated personality data collected in real employment settings over two occasions, where respondents are naturally motivated to improve their scores on the second occasion. Ellingson et al. (1999) recommended this strategy: "Future research should consider collecting data in settings where respondents will be naturally motivated to respond in a socially desirable manner" (p. 165).

From a measurement perspective, faking can only be understood as a motivated and significant change from a natural baseline condition of responding. The present study used a repeated measures design to evaluate changes in scale scores on a personality inventory on two occasions. Applicants for a customer service job who were rejected because they did not pass the employment tests provided the data. After a minimum of 6 months, they applied for the same job in the same company and completed the same test battery a second time. It seems reasonable to assume that failing the test the first time will create an incentive to change scores during the second testing—regardless of the degree to which applicants faked on the first occasion. Moreover, the goal-setting literature (cf. Austin & Vancouver, 1996) suggests that applicants who fail a selection battery and retake it are motivated to improve their scores the second time.

We defined faking in terms of score changes from an initial baseline of responses in the first job application process. We used no experimental instructions or manipulations; we applied no corrections for social desirability, faking, or response distortion; and we made no estimates of "honest scores." Rather, we used the results from the first employment test administration as a benchmark against which to compare the results from the second employment test administration.

The research presented here consists of three studies using the Hogan Personality Inventory (HPI; R. Hogan & Hogan, 1995). In the first study we compared the personality scale scores of job applicants on two occasions: (a) when applying for a job and (b) when reapplying for the same job after having been denied employment the first time. This study showed no meaningful changes between the first and second occasions, and applicants' personality scale scores went down as often as they went up on the second occasion. The second study showed that such personality scale score changes as did occur (in the positive and negative directions) were systematic and predictable using measures of social skill, social desirability, and integrity. In the third study we compared the personality scale scores of a sample of job applicants who completed the personality inventory for research purposes with the scores of the job applicants from Study 1 and found no meaningful score differences.

Study 1

Hypotheses

> *Hypothesis 1:* The scores for job applicants who fail a personality assessment battery at Time 1 (T1) will not improve on retesting at Time 2 (T2).

Previous research with cognitive ability measures supports the hypothesis that applicants will try to change their scores to improve their test results. In a meta-analysis of practice effects, J. A. Kulik, Kulik, and Bangert (1984) found that cognitive test scores increased on second administration, using a common form, by 0.42 standard deviations. Hausknecht, Trevor, and Farr (2002) found that law enforcement officers who were retested for promotion three or more times increased their cognitive test scores by 0.69 standard deviations and their oral communication test scores by 0.85 standard deviations. Lievens, Buyse, and Sackett (2005) examined test–retest practice effect sizes for cognitive ability,

knowledge, and situational judgment tests and, after correcting for unreliability, reported effects of .46, .30, and .40, respectively. These results show that, given an opportunity, applicants will try to change their scores in a positive direction. Specifying what applicants do to improve their scores is beyond the scope of this article, but it could include remembering item content from T1 (i.e., practice effects), seeking advice (i.e., coaching), and attempting to change self-presentational style. Nonetheless, the degree to which applicants can improve their scores is an empirical question. Rosse et al. (1998) found considerable variation in people's scores on faking scales—which suggests that the ability to fake is an individual-differences variable.

> *Hypothesis 2:* Applicants' score changes between T1 and T2 will form a normal distribution with a mean of zero and 5% falling outside of a 95% confidence interval (CI) around the mean.

Every applicant's score will have an associated error band based on the standard error of measurement of the scale. The standard error of measurement (SE_{msmt}) is an estimate of the error in an individual's test score; it is useful for placing an error interval around observed test scores (Thurstone, 1927) and comparing them to a known distribution. The SE_{msmt} is based on the reliability of a specific measure; the more reliable the test, the smaller the SE_{msmt}, thereby increasing the ability to detect significant score changes. Tests that meet psychometric (Nunnally, 1978) and professional (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) standards for reliability provide a sufficiently small SE_{msmt} for hypothesis testing. Guion (1998, p. 233) explained that SE_{msmt} can be used in personnel decisions to determine whether a person's score differs significantly from a hypothetical true score. Cascio, Outtz, Zedeck, and Goldstein (1991) used the SE_{msmt} with a 95% CI to establish score bands for selection purposes. Although T2 scores are not expected to be exactly the same as T1 scores, they should be within a 95% CI of the T1 score approximately 95% of the time.

> *Hypothesis 3:* Between T1 and T2, applicants will lower their scores as often as they increase them.

Although applicants should try to increase their scores, their ability to do so will be constrained by their habitual styles of impression management during interviews, by the complexity of the personality inventory items, and by their lack of knowledge about the personal requirements of the job. McFarland and Ryan (2000) noted that, although individuals can deliberately increase their test scores, it is not clear that they will try to in an actual employment testing situation. In an experimental manipulation, McFarland and Ryan found considerable variance across individuals in the extent of score change on different types of noncognitive measures.

> *Hypothesis 4:* The changes in scores between T1 and T2 will not affect the factor structure of the assessment.

We predicted (a) an a priori confirmatory factor analysis (CFA) model will fit the T1 scores as well as the T2 scores, (b) such score changes as occur will only introduce error variance into the model, and (c) a structural equation model (SEM) that treats failing the T1 battery as an intervention and includes a latent variable for change will fit the data less well than a model that ignores the intervention. These analyses are similar to those used in prior faking studies (cf. Smith & Ellingson, 2002).

Method

Sample. The study included 5,266 adults from a population of 266,582 who applied for a customer service job with a nationwide U.S. employer in the transportation industry. The population was 60% male and 40% female; race/ethnicity for Whites, Blacks, Hispanics, Asians, and American Indians was 42%, 23%, 11%, 6%, and 1%, respectively. The sample was 64% male and 36% female; race/ethnicity for Whites, Blacks, Hispanics, Asians, and American Indians was 35%, 28%, 13%, 8%, and 1%, respectively. Race/ethnicity was not reported for 17% of the population and 15% of the sample. Applicants completed a selection battery that included a personality inventory, an English comprehension test, and a cognitive ability test. The test battery was computerized and administered in a proctored test center. Applicants had to exceed cutoff scores on the designated scales on all measures. Not exceeding any cutoff score constituted failure on the battery—high scores on one measure could not compensate for low scores on another. The same scoring rules applied to both T1 and T2 administrations. The sample included only applicants who failed the test battery at T1 and reapplied for the same job after a minimum of 6 months. The same test battery was readministered at T2.

Measure. The applicants completed the HPI (R. Hogan & Hogan, 1995). The HPI is a 206-item, true–false inventory of normal personality designed to predict occupational performance. The inventory contains seven primary scales that align with the five-factor model (FFM) of personality (Digman, 1990; Goldberg, 1993; J. S. Wiggins, 1996). To link the current study with previous faking research, we used the HPI–Reduced, a shortened five-scale version of the HPI developed by Smith and colleagues (Smith, 1996; Smith et al., 2001; Smith & Ellingson, 2002). Using a quasi-confirmatory approach, Smith (1996) examined several existing measures of the FFM and identified consistencies in the measures of each construct. Using 20 HPI subscales (i.e., Homogenous Item Composites consisting of 4–6 items each), Smith and Ellingson chose item content that reflected cross-test FFM consistencies. From this analysis, they concluded, "the shortened version is a five-factor version of the HPI that capitalizes on the broad base of theoretical and empirical research supporting a five-factor conceptualization of personality" (p. 214).

Analyses. We calculated change scores (i.e., algebraic difference scores) for all retest applicants by subtracting their T1 raw scores from their T2 raw scores for each of the five HPI scales. A positive change score indicates a scale increase; a negative change indicates a scale decrease. We then calculated the reliabilities of the change scores based on the reliabilities of the T1 and T2 scores. Change scores are sometimes criticized for unreliability (J. R. Edwards, 1994; J. R. Edwards & Parry, 1993), but they are appropriate and necessary for use in within-subjects research (cf. Rogosa, Brandt, & Zimowski, 1982; Tisak & Smith, 1994). Moreover, previous research shows that differences scores on personality measures are appropriate data for factor analytic work of the type conducted here (Nesselroade & Cable, 1974). To calculate the reliability of change scores, we used the formula from McFarland

and Ryan (2000): $r_{dd} = (\sigma^2_d - \sigma^2_{ed})/\sigma^2_d$, where $\sigma^2_{ed} = \sigma^2_{T1}(1 - r_{T1}) + \sigma^2_{T2}(1 - r_{T2})$, with T1 representing the first score, T2 the second score, and σ^2_d the variance of the change score. Then, we calculated the SE_{msmt} of the change score using the change score reliability and variance for each scale.

Analyses focused on the magnitude and the direction of scale score change. The literature contains a number of methods for assessing change in test scores; we used three of the most appropriate methods. First, we calculated the correlations between the T1 and T2 scores for each scale. If applicants are unable to change their scores, then the T1–T2 correlations will be functionally the same thing as test–retest reliability coefficients, roughly comparable to the test–retest correlations presented in the HPI manual. Second, we constructed the distribution of change scores between T1 and T2 for the five personality scales. To identify the scores that changed by more than chance, we followed methods suggested by Cascio et al. (1991) and calculated a 95% CI for each scale using the SE_{msmt} and compared change scores in the distribution with this interval. Also, we calculated a distribution for the sum of the change scores across the five scales to evaluate change across the entire profile.

Third, we used a latent factor model to assess the extent to which impression management at T2 caused the construct validity of the test to deteriorate. This required three analyses: (a) fitting a confirmatory model to the T1 data; (b) fitting the identical confirmatory model to the T2 data; and (c) fitting a SEM to the combined T1, T2, and change score data. Prior to model testing, we conducted a power analysis using methods outlined by MacCallum, Browne, and Sugawara (1996) for covariance structure modeling and verified that our sample sizes were sufficient for the models to be tested.

Specifically, following Smith and Ellingson (2002), we used a CFA model. The CFA model specified an oblique five-factor structure using four subscales to define each HPI–R factor; the model fitted appears in Figure 1. We tested the fit of this model with the total sample (n = 5,266) T1 data, and then with the corresponding T2 data, using the structural equation modeling program EQS, Version 6.1 (Bentler & Wu, 2006). A further model tested whether T1 scores could be considered entirely causal for T2 scores. The schematic of the model fitted appears in Figure 2. Prior to all SEM analysis, multivariate normality of the T1 and T2 data sets was examined using Mardia's (1970) normalized multivariate kurtosis.

Results

Descriptive statistics. Table 1 contains the means, standard deviations, alpha and difference score reliabilities, SE_{msmt}, and 95% CIs around the expected score (i.e., mean) for each of the five HPI–R personality scales at T1, T2, and the change scores (i.e., T2 – T1). The effect size (Cohen's d; Cohen, 1988) for the mean change score is also reported for each scale. Because computing alpha reliabilities required complete item data for each scale, the statistics for each scale are based on complete item–case data for that scale, and that is why the number of cases differs per scale (the "Reliability n" column in Table 1). For all other subsequent analyses, prorated scale scores enabled us to use the total sample: n = 5,266 cases. These analyses are not affected by range restriction. The scale variances for the study samples at T1 and T2 are close to the scale variances for the total applicant population (N = 266,582; see Table 1). Dudek (1979) provided the equation for the standard error of measurement used in this article; it is specifically appropriate for computing the standard deviation of expected observed scores from the current observed scores, and an estimate of unreliability:

$$sem_3 = s_x\sqrt{(1 - r_{xx}^2)}$$

where

s_x = the standard deviation of observed test scores

r_{xx} = the reliability of the test.

As Nunnally and Bernstein (1994, pp. 259–260) indicated, this is the appropriate formula to be used when estimating the standard error of measurement with observed scores rather than estimated true scores as the initial score estimates.

T1–T2 correlations. Table 2 presents correlations between the scales at T1 and T2; the diagonal contains coefficient alpha reliabilities for each scale at each point in time. The within-scale correlations for T1 and T2 are an index of the stability of the test scores after failing the T1 test battery. These may be compared with the test–retest reliabilities from current research on the HPI–R (Deslauriers, Grambow, Hilliard, & Veldman, 2006), which are as follows: Emotional Stability, .92; Extraversion, .94; Openness, .92; Agreeableness, .87; and Conscientiousness, .90. That the correlations between T1 and T2 are smaller than the expected test–retest reliabilities suggests that failing the test at T1 might have affected scores at T2. The pattern of scale intercorrelations at T1 is similar to that at T2, suggesting that the personality factor structures at T1 and T2 will be similar and stable.

Observed score change distributions. Figure 3 shows the frequency distributions of T2 minus T1 score changes by personality scale for the total sample. Positive values indicate that an applicant's T2 score was larger than the T1 score. As shown by the normal curve overlaid on each graph, the five change score distributions are near normal and possess negligible skew and kurtosis except for the Agreeableness scale, which had a normalized kurtosis value of 7.46. Table 1 presents the mean for each distribution; the median and mode for each of the five distributions is 0. These results support Hypotheses 1 and 2. Average scores on two of the five scales, Emotional Stability and Extraversion, changed in a positive direction between T1 and T2 (the Extraversion T2 – T1 mean to 4 decimal places is .0038); average scores on the other three scales were lower at T2 than at T1. Mean score change on the Emotional Stability scale (20 items) was .23 raw score points higher at T2 than T1, with a standard deviation of 2.99 and a range of –13 to +15. Although statistically significant ($p < .001$), this change (i.e., d = 0.077) is not meaningful by convention (Cohen, 1988). Raw score change on the Extraversion scale (19 items) ranged from –17 to +15, with a mean of 0.00 points (SD = 3.45), which was neither a significant nor a meaningful change between T1 and T2 scores (i.e., d = 0.001). The raw score change ranged from –13 to +13 on the Openness scale (15 items), with a mean change of −0.18 points (SD = 2.62), which was statistically significant but not meaningful (i.e., d = −0.070). Mean raw score change on the Agreeableness scale (19 items) was −0.17 points (SD = 1.77), which was a statistically significant but small effect

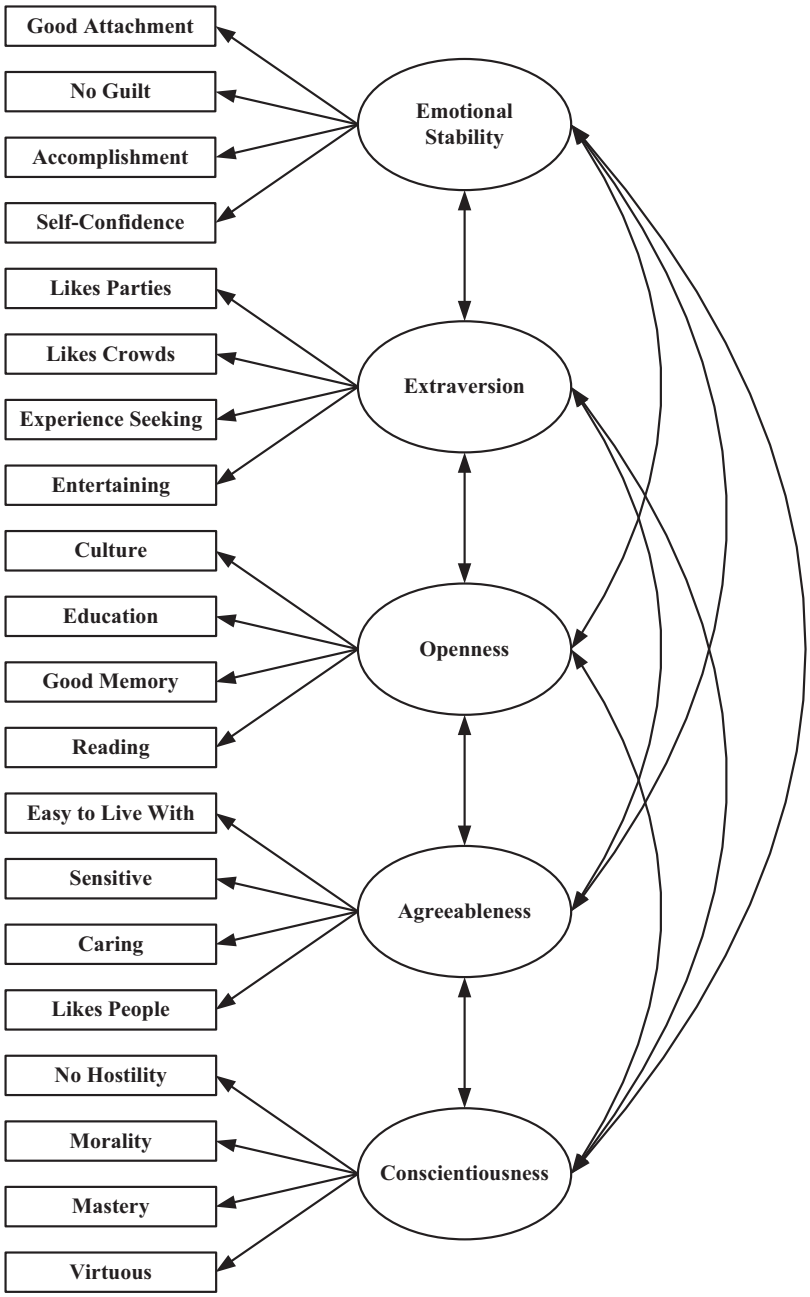

Figure 1. Confirmatory factor analysis model specified to Hogan Personality Inventory—Revised scales and fit to each of the Time 1 and Time 2 data sets.

(i.e., $d = -0.098$) and ranged from –14 to +10 points. And the raw score change ranged from –11 to +12 on the Conscientiousness scale (17 items), with a mean change of –0.04 points ($SD = 2.37$), which was neither significant nor meaningful (i.e., $d = -0.017$). These results support Hypothesis 3.

The 95% CIs around individuals' scores for each scale at T1, T2, and the T2 – T1 change score were calculated around the scale means using the respective SE_{msmt}. The SE_{msmt} estimates were calculated using the coefficient alpha reliability and standard deviation for each scale at each point in time and for the change score. The upper and lower bounds of the 95% CI at T1 and T2 for each of the five scales round to the same raw score points (with the exception of Agreeableness T1, with a lower bound CI of 15.03 vs. 14.66 for T2), indicating that for all practical purposes, any individual's T1 score was within a 95% CI at T2. The 95% CI for each of the five T2 – T1 change scores centered on a rounded raw score of zero. The upper and lower bounds of each of the five T2 – T1 95% CIs were used as comparisons to the frequency distributions of individuals' change scores on each of the five scales (see the Appendix for a full illustration of change score data).

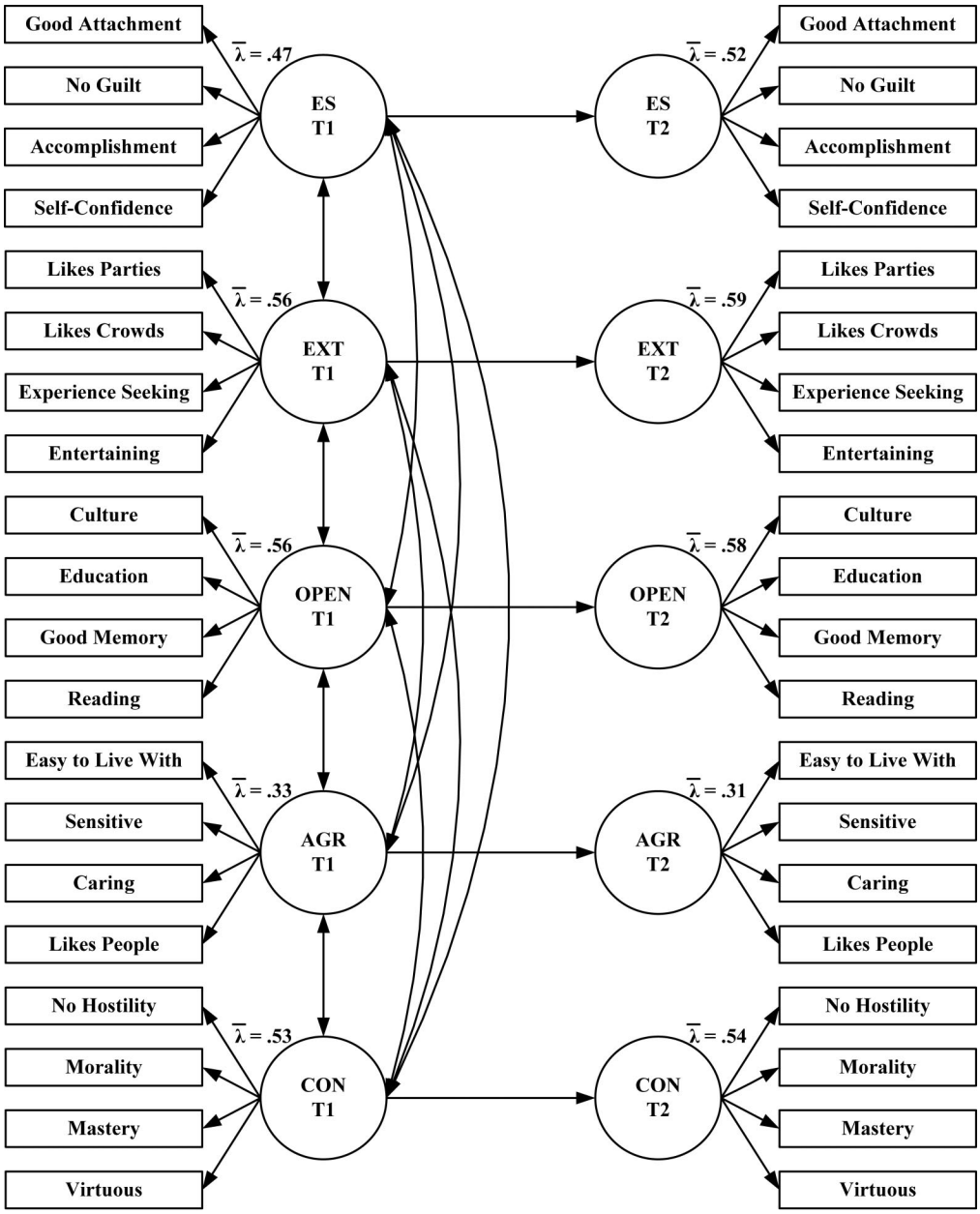

Figure 2. Structural equation model fit to combined Time 1 (T1) and Time 2 (T2) data for Hogan Personality Inventory—Revised scales. ES = Emotional Stability; EXT = Extraversion; OPEN = Openness; AGR = Agreeableness; CON = Conscientiousness. $\bar{\lambda}$ represents the average factor loadings.

Regarding the Emotional Stability scale, 3.1% of applicants changed their scores between T1 and T2 in a negative direction beyond the lower bound of the 95% CI; 4.3% changed their scores in a positive direction beyond the upper bound of the 95% CI for the same scale. For the Extraversion scale, 5.4% of applicants changed their scores beyond the lower bound of the 95% CI; 5.2% changed their scores beyond the upper bound of the 95% CI for the same scale. For the Openness scale, 3% of applicants changed their scores beyond the lower bound of the 95% CI; 3.6% changed their scores beyond the upper bound of the 95% CI for the same scale. For the Agreeableness scale, 3.3% of applicants changed their scores beyond the lower bound of the 95% CI; 1.7% changed their scores beyond the upper bound of the 95% CI for the same scale. And for the Conscientiousness scale, 3.5% of applicants changed their scores beyond the lower bound of the 95% CI; 3.2% changed their scores beyond the upper bound of the 95% CI for the same scale. Although these results support Hypothesis 3, Extraversion scores changed beyond the 5% prediction in Hypothesis 2.

On average across the five scales, 3.7% of the respondents changed their scores beyond the lower bound of the 95% CIs, and 3.6% changed their scores beyond the upper bound of the 95% CIs. Averaged across the five scales, the scores for 92.7% of the applicants were

Table 1
Descriptive Statistics by Hogan Personality Inventory—Revised (HPI–R) Scale for Time 1 (T1), Time 2 (T2), and Differences for Study 1 Applicant Sample

HPI–R scale	*M*	*SD*	Reliability *n*	Reliability	SE_{msmt}	95% CI	Cohen's *d*
Emotional Stability (range 0–20)							
Applicant population scores	15.29	3.18					
Applicant sample T1 scores	14.81	3.31	5,007	0.75	2.19	10.52, 19.10	
Applicant sample T2 scores	15.04	3.44	5,007	0.78	2.15	10.83, 19.26	
Applicant sample T2 – T1	0.23	2.99	5,007	0.40	2.74	−5.13, 5.59	0.077
Extraversion (range 0–19)							
Applicant population scores	11.39	3.71					
Applicant sample T1 scores	11.37	3.66	4,952	0.77	2.33	6.80, 15.94	
Applicant sample T2 scores	11.37	3.99	4,952	0.81	2.34	6.79, 15.96	
Applicant sample T2 – T1	0.00	3.45	4,952	0.49	3.01	−5.90, 5.91	0.001
Openness (range 0–15)							
Applicant population scores	10.34	3.18					
Applicant sample T1 scores	10.37	3.18	5,014	0.75	2.10	6.24, 14.49	
Applicant sample T2 scores	10.18	3.36	5,014	0.78	2.10	6.06, 14.30	
Applicant sample T2 – T1	−0.18	2.62	5,014	0.27	2.52	−5.13, 4.76	−0.070
Agreeableness (range 0–19)							
Applicant population scores	17.73	1.52					
Applicant sample T1 scores	17.61	1.55	4,988	0.53	1.32	15.03, 20.18	
Applicant sample T2 scores	17.43	1.78	4,988	0.61	1.41	14.66, 20.20	
Applicant sample T2 – T1	−0.17	1.77	4,988	0.24	1.71	−3.53, 3.19	−0.098
Conscientiousness (range 0–17)							
Applicant population scores	13.42	2.46					
Applicant sample T1 scores	13.49	2.53	4,970	0.68	1.86	9.85, 17.12	
Applicant sample T2 scores	13.45	2.63	4,970	0.71	1.85	9.82, 17.08	
Applicant sample T2 – T1	−0.04	2.37	4,970	0.28	2.28	−4.50, 4.42	−0.017

Note. T2 − T1 mean differences were rounded from four decimal points. Applicant population: $N = 266{,}582$. SE_{msmt} represents the standard error of measurement (fixed observed score formula). The 95% confidence interval (CI) was calculated using SE_{msmt}.

unchanged from T1 to T2. Of the 7.3% whose T1 scores changed, the scores were as likely to decrease as to increase at T2. Three (0.06%) of 5,266 applicants changed scores across all five scales beyond the 95% CIs. Twenty-two (0.4%) applicants changed their scores beyond the 95% CIs for any combination of four scales.

We also calculated a total change score by summing score changes across the five scales (range = –51 to +39). The mean of the total change score scale is –.20, the median is 0, the mode is –1, and the standard deviation is 8.40. Overall, applicants lowered their scores very slightly on the personality scales between T1 and T2.

Latent factor analysis results. We analyzed the T1 and T2 data using CFA. The CFA model included the five HPI–R scales with four subscales (cf. Smith & Ellingson, 2002); this model is shown in Figure 1. The subscales and associated error terms were not allowed to correlate across scales—only through the common latent factors. We compared parameter estimates for the model from each of the T1 and T2 data sets and then used goodness-of-fit tests to determine whether the model would provide acceptable fit to the T1 data, and whether the same model provided equivalent fit to both the T1 and T2 data.

Steiger and Lind (1980) proposed the root-mean-square error of approximation (RMSEA) as an index of overall model fit. The RMSEA is a measure of discrepancy per degree of freedom. According to Browne and Cudeck (1993), there is good model fit

Table 2
Intercorrelations Between Hogan Personality Inventory—Revised (HPI–R) Scales at Time 1 (T1) and Time 2 (T2)

HPI–R scale	1	2	3	4	5	6	7	8	9	10
1. Emotional Stability T1	*.75*									
2. Extraversion T1	.06	*.77*								
3. Openness T1	.31	.24	*.75*							
4. Agreeableness T1	.17	.33	.16	*.53*						
5. Conscientiousness T1	.48	.06	.31	.26	*.68*					
6. Emotional Stability T2	**.61**	.06	.23	.12	.32	*.78*				
7. Extraversion T2	−.01, *ns*	**.59**	.14	.19	−.01, *ns*	.09	*.81*			
8. Openness T2	.20	.21	**.68**	.10	.20	.37	.26	*.78*		
9. Agreeableness T2	.08	.24	.10	**.46**	.12	.21	.35	.22	*.61*	
10. Conscientiousness T2	.34	.07	.22	.16	**.58**	.52	.07	.36	.30	*.71*

Note. $n = 5{,}266$. Coefficient alpha reliabilities for each scale at T1 and T2 are presented in italics on the main diagonal. Values in boldface represent T1 versus T2 same-scale score correlations. Correlations $> |.04|$ are significant at $p < .01$.

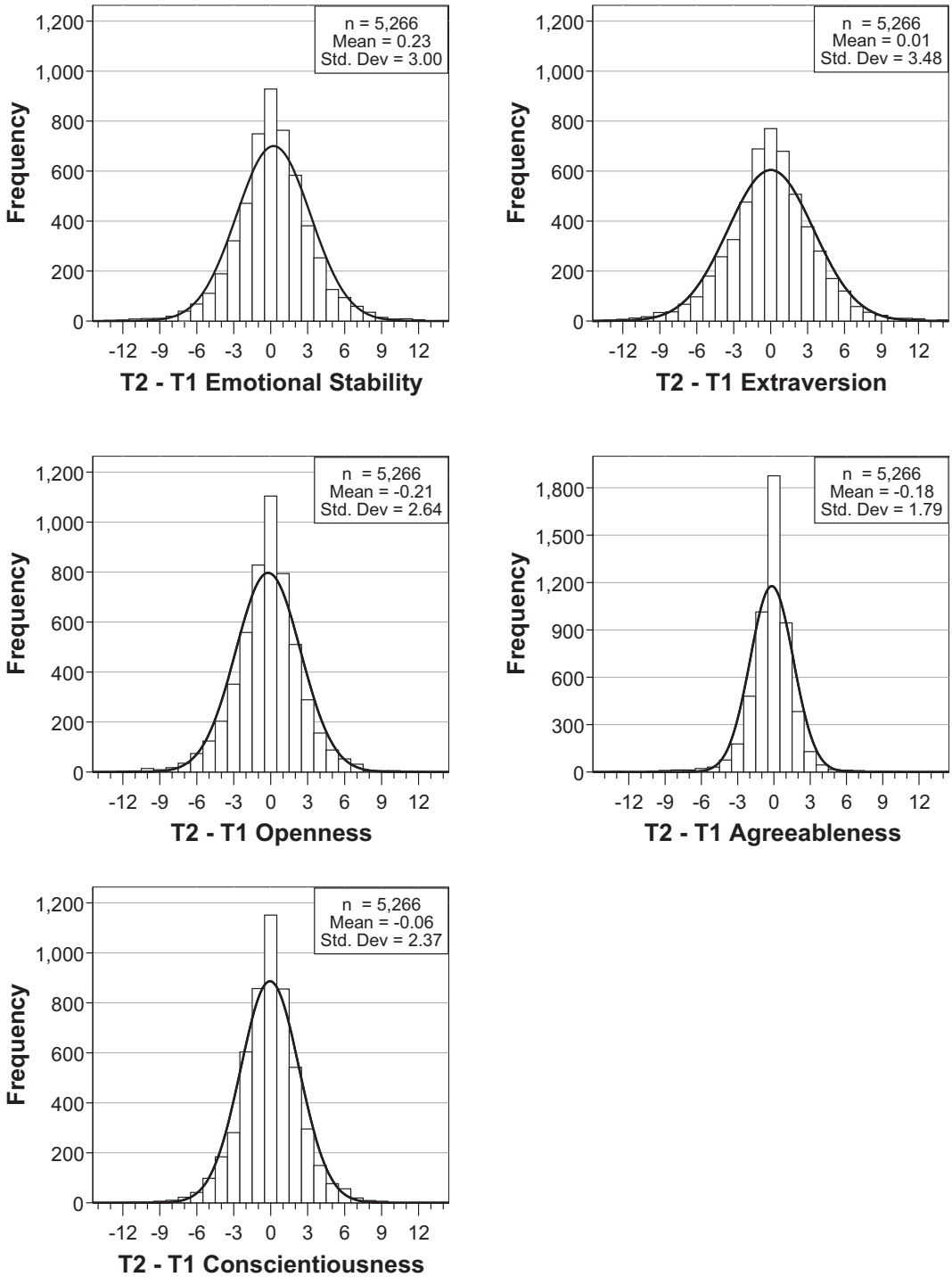

Figure 3. Applicants' Time 2 (T2) – Time 1 (T1) change scores with normal curve for each Hogan Personality Inventory—Revised scale.

if the RMSEA is less than or equal to .05 and adequate fit if the RMSEA is less than or equal to .08. Hu and Bentler (1999) suggested RMSEA = .06 as the cutoff for good model fit. The RMSEA does not require comparison with a null model and is not affected by sample size, as are chi-square tests of fit. The RMSEA also has a known distribution, related to the noncentral chi-square distribution, and thus does not require bootstrapping to establish CIs. The EQS 6.1 statistical package reports 90% CIs for RMSEA.

An initial check of Mardia's (1970) normalized multivariate kurtosis yielded values of 106.35 and 112.85 for T1 and T2 data,

respectively. Bentler (2006, p. 129) recommended this value should be between –3 and +3 for normality to be assured. Thus, given the sample size of greater than 5,000 in each model to be fitted, along with the evidence from Yuan and Bentler (1998) on the relative accuracy of the Browne (1984) chi-square fit statistic using asymptotic distribution-free (ADF) estimation with samples of this size, all SEM modeling was implemented in EQS 6.1 using the arbitrary generalized least squares (unbiased ADF) algorithm. Although generalized least squares might also have been considered as an estimation option, the results from Olsson, Troye, and Howell (1999) and Powell and Schafer (2001) contraindicate its use. We first examined the fit of the CFA model to the T1 data and then to the T2 data at the path coefficient and overall model levels. Path coefficients were all statistically significant ($p < .05$) and nearly equal across T1 and T2 data. The average factor loadings (i.e., $\bar{\lambda}$) across the four indicators for each HPI–R latent factor are presented in Figure 2. The overall model fit to the T1 data was good (n = 5,266; RMSEA = 0.050; 90% CI = 0.048, 0.052) and equivalent to overall fit to the T2 data (n = 5,266; RMSEA = 0.051; 90% CI = 0.049, 0.053), as shown by the overlap in the 90% CIs.

T1–T2 SEM. We tested a third SEM, which is essentially a "no faking beyond what took place in T1" model. The "no faking" model used T1 latent factors as the only causal paths for the T2 latent factors (see Figure 2). The T1 latent factors are indicated by their respective subscales and are modeled as oblique exogenous factors as before. Each T1 latent factor has a direct causal link to the associated T2 latent factor. The endogenous T2 latent factors are indicated by their respective subscales. This model shows the T1 latent variables anchoring the T1 and T2 observed scores. In other words, the scores at T1 are baseline and the T2 scores are caused by T1. This model tests the stability of the assessment across T1 and T2. Error terms were not correlated within or across T1 and T2 assessments. This model had reasonable fit to combined T1–T2 data (n = 5,266; RMSEA = 0.056; 90% CI = 0.055, 0.057). All path coefficients were statistically significant ($p < .05$). The model fit confirmed that latent variable scores on T2 were essentially accounted for entirely by T1 scores and that no other substantive variance remained to be explained.

Study 2

Our next question concerns individual differences in the ability to change one's scores. Figure 3 shows that T1 and T2 change scores are normally distributed, even leptokurtic, across five HPI–R scales; no more than 5.2% of the applicants at each tail of the scale distribution changed, and different people changed across different scales. Although these changes are minimal, they still exceeded the 5% prediction in Hypothesis 2. We wondered whether these changes were systematic and predictable. We first examined demographic variables and found no score changes by gender, race–ethnicity, or age.

Next, we reasoned that individual differences in *social skill*—the ability to put on an attractive performance—and *social desirability*—the tendency to present oneself in a socially desirable fashion—would predict such changes as did occur. People with good social skills can control the impressions they make on others; people with poor social skills seem unaware of the impressions they make on others. People motivated by social desirability try to make socially appropriate (as opposed to engaging) impressions on others (Paulhus, 1991); people who are not motivated by social desirability seem aloof, insensitive, odd, eccentric, deviant, or socially inappropriate (R. Hogan, 1991; N. Wiggins, 1966).

McFarland and Ryan (2000) reported that "integrity" is related to faking. They tested students using the revised NEO Personality Inventory (Costa & McCrae, 1989) in honest and faking conditions. They found that students with higher integrity and Conscientiousness, and lower Neuroticism scores, were less likely to increase their scores on the NEO than those with the opposite scores.

Hypotheses

> *Hypothesis 5:* Persons with high scores on a measure of social skill will improve their personality scores in characteristic ways if they have an incentive to do so. Persons with low scores on a measure of social skill will be unable to improve their personality scores, even when they have an incentive to do so.

The extensive literature on social skills (cf. Argyle, 1981; R. Hogan, 1991) has maintained that socially competent people have the ability to produce the desired effects in other people in social situations. Specifically, socially skilled individuals should be able to change their scores on measures of Extraversion and Agreeableness in order to appear more socially engaging than they typically might be—because they know how "to turn on the charm."

> *Hypothesis 6:* Persons with high scores on a measure of social desirability will improve their scores on personality scales in a characteristic way if they think it is necessary. Persons with low scores on a measure of social desirability will be unwilling to change their scores on personality scales, even if it seems beneficial.

Paulhus (1984, 1991) has suggested that social desirability is the tendency to make oneself seem socially appropriate. Individuals disposed to socially desirable responding should improve their scores on Emotional Stability and Conscientiousness so as to make themselves seem more virtuous and conforming than they typically might be.

> *Hypothesis 7:* Persons with high scores on a measure of integrity will be less likely to change their scores on personality scales in reapplying for a job than those with lower scores.

McFarland and Ryan (2000) found that individuals with higher scores on an integrity measure changed their NEO scale scores less than those with lower integrity scores. Emotional Stability, Extraversion, Agreeableness, and Conscientiousness scores changed most; Openness changed the least (also see Mersman & Shultz, 1998). Because integrity includes aspects of Emotional Stability, Conscientiousness, and Agreeableness (J. Hogan & Ones, 1997), these are the scales most likely to be affected by individual differences in integrity.

Method

Sample. We used three samples in this study. All were applicants for a customer service job with the same employer as in Study 1. All applicants failed the selection battery at T1 and reapplied for the same job with the same employer after 6 months. One sample completed the HPI at T1; they completed the HPI and a measure of social skills at T2. This sample included 541 adults, of whom 385 were male and 156 were female; race–ethnicity for Whites, Blacks, Hispanics, Asians, and American Indians was 31%, 25%, 14%, 12%, and 0% respectively. Race–ethnicity was not reported for 18% of the sample. The second sample completed the HPI at T1; they completed the HPI and a measure of social desirability at T2. This sample included 535 adults, of whom 343 were male and 192 were female; race–ethnicity for Whites, Blacks, Hispanics, Asians, and American Indians was 28%, 28%, 12%, 11%, and 1%, respectively. Race–ethnicity was not reported for 20% of the sample. The third sample was the same as in Study 1; they completed the HPI and a measure of integrity at T2. This sample (n = 5,266) is described in the Study 1 section.

Measures. We composed measures of Social Skill and Social Desirability using items from the International Personality Item Pool (IPIP; Goldberg, 1999; Goldberg et al., 2006). Combinations of IPIP items can be used to build scales for Social Skill and Social Desirability. The psychometric properties of these IPIP scales appear on the IPIP Web site (International Personality Item Pool, 2001).

The IPIP Social Skills scale contains 39 items with five subscales: (a) Empathy, (b) No Social Anxiety, (c) Even-Tempered, (d) Emotional Intelligence/Empathic Concern, and (e) Self-Monitoring. The coefficient alpha reliability for the Social Skills measure in this study was .76.

The IPIP Social Desirability scale contains 31 items with two subscales: (a) Unlikely Virtues and (b) Impression Management. Coefficient alpha reliability for the Social Desirability measure in this study was .86.

We used the Employee Reliability Index (J. Hogan & Hogan, 1989) as a measure of Integrity. This 18-item scale identifies persons who are honest, dependable, and good organizational citizens. The scale, which has a coefficient alpha reliability of .75, was validated using various delinquency criteria and counterproductive behavior (R. Hogan & Hogan, 1995).

Analyses. For each applicant, we calculated a change score for the personality measures between T1 and T2. We then formed a distribution of change scores for each scale, and we then examined two groups for each scale (i.e., Emotional Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). The first group contained applicants whose change score fell below the SE_{msmt} at the 95% CI; the second group contained applicants whose change score fell above the SE_{msmt} at the 95% CI. This comparison provides useful descriptive information regarding the magnitudes of scores on the Social Skills, Social Desirability, and Integrity scales, as a function of extreme T2 – T1 difference scores. Next, using the total sample data for each of Samples 1, 2, and 3, we computed a Pearson product–moment correlation between each HPI–R T1–T2 difference score and the measures of Social Skill, Social Desirability, and Integrity; these correlations were also expressed as an effect size (d).

Results

Table 3 shows the sample sizes, scale means, standard deviations, reliabilities, SE_{msmt}, and score confidence ranges for each HPI–R scale, using a combined data set of Samples 1 and 2 (n = 541 + 535 cases). The results in Table 3 can be compared directly with those in Table 1. Such a comparison reveals a high degree of consistency between data set descriptive and psychometric indices. Table 4 provides the number of cases, mean differences, and standard deviations by subset change group for Social Skills, Social Desirability, and Integrity, alongside the correlation and effect size value computed over all data for each specific sample. As can be seen, the effect sizes for Social Skills and Social Desirability were moderate to large across all five HPI dimensions; applicants with higher scores for Social Skills and/or Social Desirability tended to increase their scores on all five HPI scales at T2. Conversely, applicants with lower scores on Social Skills and Social Desirability tended to lower their scores on the five HPI scales at T2. For the Integrity analysis, applicants who increased their scores on Conscientiousness, Emotional Stability, and Openness tended to score lower on Integrity at T1.

We note the following findings in support of Hypotheses 5 and 6. First, as predicted, applicants with high scores on Social Skill raised their scores the most on the scales for Extraversion and Agreeableness, whereas applicants with low scores on Social Skills lowered their scores the most on the same scales. Second, as predicted, applicants with high scores on Social Desirability raised their scores the most on the scales for Emotional Stability and Conscientiousness, whereas applicants with low scores on Social Desirability lowered their scores the most on the same scales. Hypothesis 7 was also supported; applicants with low scores on Integrity changed significantly in the positive direction on the Emotional Stability, Conscientiousness, and Openness scales. The finding for Openness is not consistent with previous research; Openness seems to be the least changeable FFM dimension. Also, Integrity was unrelated to score changes on Extraversion and Agreeableness. This analysis benefits from the power of the sample size from Study 1.

Study 3

The repeated measures design used in Study 1 overcomes some persistent problems with prior faking research; however, the question still remains as to whether applicants fake during their first testing and then again on retesting. As one observer commented, "Maybe it is all faking all of the time." Study 3 attempts to evaluate the criticism that the applicants' scores in Study 1 did not change on retesting because they were faking on both occasions. One way to evaluate the claim that applicants were faking at T1 is to test a matched employment-related sample who, presumably, are not motivated to fake. We searched the HPI archives for studies that contained (a) job applicants, (b) for customer service jobs, (c) with nationwide employers, and (d) with employment selection test batteries that included a cognitive ability measure and the HPI. We wanted an adult sample who completed the HPI for research purposes and whose results were not used for personnel selection. We tested the following hypothesis:

> *Hypothesis 8:* Scores for job applicants who completed the HPI for research purposes will not differ substantially from

Table 3
Descriptive Statistics by Hogan Personality Inventory—Revised (HPI–R) Scale for Time 1 (T1), Time 2 (T2), and Differences for Study 2, Samples 1 and 2 Combined Applicant Sample

HPI–R scale	*M*	*SD*	Reliability *n*	Reliability	SE_{msmt}	95% CI	*d*
Emotional Stability (range = 0–20)							
Applicant population scores	15.29	3.18					
Applicant sample T1 scores	14.77	3.42	1,063	.77	2.18	10.50, 19.04	
Applicant sample T2 scores	15.04	3.58	1,063	.80	2.15	10.82, 19.25	
Applicant sample T2 – T1	0.27	3.12	1,063	.46	2.77	−5.16, 5.70	0.085
Extraversion (range = 0–19)							
Applicant population scores	11.39	3.71					
Applicant sample T1 scores	11.37	3.70	1,056	.77	2.36	6.75, 16.00	
Applicant sample T2 scores	11.32	4.07	1,056	.81	2.39	6.64, 16.00	
Applicant sample T2 – T1	−0.06	3.60	1,056	.51	3.09	−6.11, 6.00	−0.016
Openness (range = 0–15)							
Applicant population scores	10.34	3.18					
Applicant sample T1 scores	10.19	3.25	1,070	.76	2.11	6.06, 14.33	
Applicant sample T2 scores	10.13	3.43	1,070	.79	2.11	6.00, 14.25	
Applicant sample T2 – T1	−0.07	2.78	1,070	.35	2.60	−5.17, 5.04	−0.024
Agreeableness (range = 0–19)							
Applicant population scores	17.73	1.52					
Applicant sample T1 scores	17.51	1.73	1,066	.61	1.37	14.82, 20.20	
Applicant sample T2 scores	17.41	2.07	1,066	.72	1.44	14.59, 20.22	
Applicant sample T2 – T1	−0.10	2.15	1,066	.49	1.88	−3.79, 3.58	−0.048
Conscientiousness (range = 0–17)							
Applicant population scores	13.42	2.46					
Applicant sample T1 scores	13.55	2.61	1,059	.71	1.84	9.95, 17.16	
Applicant sample T2 scores	13.63	2.70	1,059	.73	1.84	10.02, 17.25	
Applicant sample T2 – T1	0.08	2.55	1,059	.39	2.34	−4.52, 4.67	0.031

Note. T2 − T1 mean differences were rounded from four decimal points. Applicant population: N = 266,582. SE_{msmt} represents the standard error of measurement (fixed observed score formula). The 95% confidence interval (CI) was calculated using SE_{msmt}.

scores for applicants who complete the HPI as part of the employment selection process.

We located one study that met these conditions. The sample (N = 141) contained applicants who were hired based on their cognitive ability scores and who also completed the HPI for research purposes. These applicants were not compensated for their time, nor were they given any results from the tests they completed.

Method

The research sample (n = 141) consisted of applicants for a customer service job with a nationwide U.S. employer in the distribution services industry. The sample was 66% male and 34% female; race–ethnicity for Whites, Blacks, Hispanics, and Asians was 28%, 35%, 6%, and 11%, respectively. Race–ethnicity was not reported for 20% of the sample. Applicants completed a selection battery consisting of an application blank and a cognitive ability test. These were followed by the HPI. Applicants were assured that the personality measure was being used for research purposes only. All measures were completed in paper-and-pencil format, and the cognitive and personality measures were administered in proctored settings. Applicants were required to answer all scored items on the application blank correctly, and they had to exceed a cutoff score on the cognitive ability test to pass the assessment phase of the screening process.

Results

Table 5 contains the personality mean scale scores for the Study 1 applicant (n = 5,266) sample at T1 and T2 and the Study 3 applicants (n = 141) who completed the HPI for research purposes. In terms of algebraic mean differences, the research sample scored higher than T1 applicants on Emotional Stability and lower on the other scales; the research sample scored lower than T2 applicants on three of the five scales. For Study 1 applicants and the research sample applicants, mean scale score differences as reflected by Cohen's d were small, with effect sizes ranging from 0.01 to 0.15. In general, applicants at T1 got the same scores as they did at T2, and both sets of scores were the same as scores for a sample of research applicants. These results support Hypothesis 8.

Discussion

Murphy and Dzieweczynski (2005) argued that the problem of faking undermines the validity of personality assessment, and this view seems to be widely shared. For example, one complaint about this study is that the data are hard to interpret because the applicants might have been faking at both T1 and T2. In contrast with this worry, our view can be summarized as "all faking all of the time"—which means that faking doesn't matter. Consider the data in Table 5. The columns contain mean scores for the applicant sample at T1, the same applicant sample at T2, and a sample of 141 research applicants for a customer service job who completed the HPI as part of a different piece of validation research during the same period of time. Applicants at T1 get the same scores as they do at T2, and both sets of scores are the same as scores for a sample of research applicants.

We believe that the concern about faking reflects a misunderstanding of the item response process. There are essentially two theories regarding what people do when they respond to items on

Table 4
Social Skills, Social Desirability, and Integrity Scale Differences for Hogan Personality Inventory—Revised (HPI–R) Score Change

	Social Skills Scale				Social Desirability Scale				Integrity Scale			
HPI–R scale	*M*−	*M*+	*r*	*d*	*M*−	*M*+	*r*	*d*	*M*−	*M*+	*r*	*d*
Emotional Stability												
M	21.79	28.29	.20	0.41	15.94	26.22	.35	0.75	13.65	10.72	−.23	−0.47
SD	10.20	3.30			6.24	3.34			2.51	2.81		
n	14	28			16	23			162	229		
Extraversion												
M	22.92	28.91	.29	0.61	21.32	23.68	.09	0.18	13.60	13.24	−.03	−0.06
SD	5.29	4.76			5.18	5.38			3.02	3.04		
n	37	35			34	31			283	272		
Openness												
M	20.67	27.28	.19	0.39	16.42	26.55	.30	0.63	14.01	12.13	−.11	−0.22
SD	4.89	4.11			6.63	3.44			2.36	3.02		
n	15	23			12	29			157	192		
Agreeableness												
M	20.86	25.56	.27	0.56	18.05	24.73	.17	0.35	13.17	11.76	−.05	−0.10
SD	4.84	4.50			5.61	4.08			2.82	3.42		
n	29	16			21	11			174	87		
Conscientiousness												
M	23.42	27.29	.21	0.43	15.89	25.76	.32	0.68	13.66	10.83	−.21	−0.43
SD	9.21	2.69			6.30	3.46			2.94	3.08		
n	24	28			18	17			182	167		

Note. *M*− represents the mean of the subgroup (on Social Skills, Social Desirability, or Integrity) whose T2 – T1 change scores were below the lower bound of the 95% confidence interval for the HPI–R scales; *M*+ represents the mean of the subgroup (on Social Skills, Social Desirability, or Integrity) whose T2 – T1 change scores were above the upper bound of the 95% confidence interval for the HPI–R scales; *r* represents the correlation between the relevant T2 – T1 change scores and the Social Skills ($n = 541$), Social Desirability ($n = 535$), and Integrity scale scores ($n = 5{,}266$); all correlations for all samples are statistically significant at $p < .05$; *d* represents Cohen's standardized effect size (calculated from *r*); ~0.2 = small, ~0.5 = medium, ~0.8 and greater = large (Cohen, 1988).

a personality measure: (a) self-report and (b) engage in impression management. These theories have very different implications for understanding faking.

Self-Report Theory of Faking

Many psychologists believe that responses to items on personality measures are self-reports; however, this is a theory of what happens, not a factual account. Self-report theory is based on two assumptions; the first concerns memory, and the second concerns communication. Self-report theory assumes that memory is like a videotape so that, when people read an item on an inventory ("I read 10 books a year"), they play back the memory videotape, compare the item with the tape, and then "report." Self-report theory also assumes that when people report, they offer factual accounts of how an item matches their memory tape. Faking involves providing inaccurate reports about the match between the content of an item and the content of memory.

Self-report theory has two problems. First, memory researchers from Bartlett (1937) to the present have argued that memories are not relatively faithful recordings of past events; they are self-serving reconstructions. Second, much communication does not concern accurate reporting of the world; it concerns trying to control the behavior of others (Dunbar, 2004). Thus, self-report theory is inconsistent both with the research regarding how memory works and with modern thinking about the nature of communication—which suggest that people construct their memories and use communication to manipulate others.

Table 5
Comparison of Applicants' Time 1 (T1) and Time 2 (T2) Hogan Personality Inventory—Revised (HPI–R) Scores With Research Incumbents' Scores

	Applicant T1		Applicant T2		Research sample		*d*	
HPI–R scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	T1	T2
Emotional Stability	14.82	3.30	15.05	3.44	15.00	3.17	0.056	−0.014
Extraversion	11.35	3.66	11.36	4.00	10.81	3.86	−0.147	−0.138
Openness	10.38	3.18	10.18	3.36	10.27	3.22	−0.036	0.028
Agreeableness	17.58	1.59	17.40	1.83	17.53	1.57	−0.029	0.073
Conscientiousness	13.50	2.53	13.44	2.63	13.36	2.45	−0.056	−0.031

Note. Applicants, $n = 5{,}266$; research sample, $n = 141$.

Impression Management Theory

Impression management theory maintains that during social interaction, most people try to maximize acceptance and status and minimize rejection and the loss of status (cf. R. Hogan, 2006). When people respond to employment interview questions, assessment center exercises, or items on a personality inventory, they behave exactly as they do during any other interaction—they try to create a particular (usually favorable) impression of themselves. In this view, faking involves changing the manner in which one typically behaves during interaction; faking involves distorting the way one normally communicates about oneself.

Note that impression management theory does not assume the existence of "real selves" inside people, and faking does not concern acting in ways that are discrepant from those real selves. In our view, assumptions about "real selves" reflect a serious misunderstanding of the nature of personality development. Consider the goals of child rearing. Small children usually act in ways that are seamlessly related to their real desires and urges. The socialization process consists almost entirely of training children to hide, or at least delay, their real desires and urges and, instead, to behave in ways that are consistent with the norms of civilized adult conduct. For self-report theory, the socialization process involves training children to fake. For impression management theory, socialization involves training children in appropriate forms of self-presentation.

The items on well-validated personality measures sample ordinary socialized adult behavior. Most adults know the rules of conduct and respond to the items in terms of social norms rather than in terms of their real desires and urges. On the other hand, criminals and other unsocialized deviants respond to personality items in ways that are closer to their real selves—and in ways that are consistent with their typical behavior. The larger point here is that it is almost impossible to distinguish faking from socialized behavior. And this means that it is very hard to assign a clear meaning to the claim that some people fake when they respond to personality measures.

More important, however, it is possible to make an empirical comparison of self-report and impression management theories of item responding (cf. Johnson & Hogan, 2006). For example, impression management theory predicts that the scores of people with good impression management skills should be more valid than scores of people whose skills are poor. Mills and Hogan (1978) asked members of a community service organization to complete the California Psychological Inventory (CPI; Gough, 1975). They then had these members rated for dominance, femininity, and social presence (all CPI scales). The discrepancy between each person's scale score and rating for that same attribute correlated –.87 with their score on R. Hogan's (1969) Empathy scale, a well-validated measure of impression management skills (Johnson, Cheek, & Smither, 1983).

Self-report theory predicts that the personality scores of people who are honest will be more consistent than scores of people who are dishonest. Conversely, impression management theory predicts that scores of people with good impression management skills will be more consistent than scores of people with poor impression management skills. Johnson (1981) tested these predictions in three separate samples; his results clearly support the impression management theory of item responding and provide no support for self-report theory—consistency in personality scale scores is related to impression management not honesty.

Returning now to the complaint that our job applicants were faking at T1 and T2, we would agree—if faking is defined as normal impression management. We are making a very simple claim. When people complete a well-validated personality measure as part of a job application process, are denied employment, reapply some time later, and then take the personality measure a second time, their scores will not change significantly on the second occasion. This claim is fully supported by our data. We also believe it is reasonable to assume that the applicants tried to improve their scores on the second occasion; on the basis of this assumption, we interpret the data as showing that, when they try, applicants are unable to improve their scores substantially. Not everyone will agree with this assumption, but the data are what they are, and this is an appropriate data set for testing the faking argument. There were no manipulations, and the data are based exclusively on processes that employers actually use and therefore apply to real selection procedures. R. Hogan, Hogan, and Roberts (1996) reviewed the faking literature and concluded that, although the data clearly show that faking does not adversely affect the validity of personality measures for employment decisions, "the issue seems somehow unlikely to go away" (p. 475).

References

Abrahams, N. M., Neumann, I., & Githens, W. H. (1971). Faking vocational interests: Simulated versus real life motivation. *Personnel Psychology, 24,* 5–12.

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

Argyle, M. (Ed.). (1981). *Social skills and work.* London: Routledge & Kegan Paul.

Arkin, R. M. (1981). Self-presentational styles. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp. 311–333). San Diego, CA: Academic Press.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin, 120,* 338–375.

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

Bartlett, F. C. (1937). *On remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.

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

Bentler, P. M. (2006). EQS Structural Equations Program (Version 6) [Computer manual]. Encino, CA: Multivariate Software.

Bentler, P. M., & Wu, E. J. C. (2006). EQS Structural Equations Program (Version 6.1) [Computer software]. Encino, CA: Multivariate Software.

Browne, M. W. (1984). Asymptotic distribution-free methods for analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Cascio, W. E., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233–264.

Christiansen, N. D., Edelstein, S., & Flemming, B. (1998, April). *Recon-*

sidering forced-choice formats for applicant personality assessment. Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Dallas, TX.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Costa, P. T., Jr., & McCrae, R. R. (1989). *The NEO PI/FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

Deslauriers, J., Grambow, D., Hilliard, T., & Veldman, L. (2006). *Test–retest reliability of the Hogan Personality Inventory and the Hogan Business Reasoning Inventory* (Tech. Rep. No. 52606). St. Louis: University of Missouri–St. Louis.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Dunbar, R. I. M. (2004). *Grooming, gossip, and the evolution of language.* London: Faber & Faber.

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.

Dwight, S. A., & Donovan, J. J. (1998, April). *Warning: Proceed with caution when warning applicants not to dissimulate.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

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

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

Edwards, J. R. (1994). Regression analysis as an alternative to difference scores. *Journal of Management, 20,* 683–689.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Review, 36,* 1577–1613.

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

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

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84–96.

Gough, H. G. (1975). *Manual for the California Psychological Inventory* (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87,* 243–254.

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

Hogan, J., & Hogan, R. (1989). How to measure employee reliability. *Journal of Applied Psychology, 74,* 273–279.

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

Hogan, J., & Ones, D. S. (1997). Conscientiousness and integrity at work. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 849–870). New York: Academic Press.

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology, 33,* 307–316.

Hogan, R. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 873–919). Palo Alto, CA: Consulting Psychologists Press.

Hogan, R. (2005). In defense of personality measurement. *Human Performance, 18,* 331–341.

Hogan, R. (2006). *Personality and the fate of organizations.* Mahwah, NJ: Erlbaum.

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual.* Tulsa, OK: Hogan Assessment Systems.

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

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. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Hough, L. M., & Furnham, A. (2003). Use of personality variables in work settings. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Comprehensive handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 131–169). New York: Wiley.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial work and organizational psychology* (Vol. 1, pp. 233–377). London: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria in fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

International Personality Item Pool. (2001). A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. Retrieved April 14, 2004, from http://ipip.ori.org/

Jackson, D. N., Wrobleski, 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, J. A. (1981). The "self-disclosure" and "self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology, 40,* 761–769.

Johnson, J. A., Cheek, J. M., & Smither, R. (1983). The structure of empathy. *Journal of Personality and Social Psychology, 45,* 1299–1312.

Johnson, J. A., & Hogan, R. (2006). A socioanalytic view of faking. In R. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking* (pp. 209–231). Greenwich, CT: Information Age.

Kelly, E. L., Miles, C. C., & Terman, L. M. (1936). Ability to influence one's score on a typical paper-and-pencil test of personality. *Journal of Personality and Social Psychology, 4,* 206–215.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435–447.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58,* 981–1007.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Marshall, M. B., DeFruyt, F., Rolland, J.-P., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO PI–R. *Psychological Assessment, 17,* 379–384.

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

Mersman, J. L., & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences, 24,* 217–227.

Mills, C., & Hogan, R. (1978). A role theoretical interpretation of personality scale item responses. *Journal of Personality, 46,* 778–785.

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

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance, 18,* 343–358.

Nesselroade, J. R., & Cable, D. G. (1974). "Sometimes, it's okay to factor difference scores"—The separation of state and trait anxiety. *Multivariate Behavioral Research, 9,* 273–284.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretic fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavioral Research, 34,* 31–58.

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Social and Personality Psychology, 46,* 598–609.

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

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics, 26,* 105–132.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90,* 726–748.

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

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology, 43,* 133–168.

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

Smith, D. B. (1996). *The Big Five in personnel selection: Reexamining frame of reference effects.* Unpublished master's thesis, University of Maryland, College Park.

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

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

Smith, D. B., & Robie, C. (2004). The implications of impression management for personality research in organizations. In B. Schneider & D. B. Smiths (Eds.), *Personality and organizations* (pp. 111–138). Hillsdale, NJ: Erlbaum.

Snell, A. F. (2006, May). *A closer look at applicant faking behavior.* Panel discussion at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors.* Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Thurstone, L. L. (1927). The scoring of individual performance. *Journal of Educational Psychology, 18,* 505–524.

Tisak, J., & Smith, C. S. (1994). Defending and extending difference score methods. *Journal of Management, 20,* 675–682.

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the 106th Annual Convention of the American Psychological Association, San Francisco, CA.

Wiggins, J. S. (1996). *The five-factor model of personality.* New York: Guilford Press.

Wiggins, N. (1966). Individual viewpoints of social desirability. *Psychological Bulletin, 66,* 68–77.

Worthington, D. L., & Schlottmann, R. S. (1986). The predictive validity of subtle and obvious empirically derived psychology test items under faking conditions. *Journal of Personality Assessment, 50,* 171–181.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51,* 289–309.

Appendix

Change Score Frequency Distributions by Hogan Personality Inventory—Revised Scale

	Frequency of change score by scale				
Change in raw score	Emotional Stability (0–20)	Extraversion (0–19)	Openness (0–15)	Agreeableness (0–19)	Conscientiousness (0–17)
−17	0	1	0	1	0
−16	0	1	0	0	0
−15	0	3	0	0	0
−14	0	3	0	1	0
−13	3	4	1	1	0
−12	4	7	3	1	1
−11	8	12	4	1	2
−10	10	17	14	4	2
−9	11	34	9	10	6
−8	19	37	17	13	10
−7	39	67	35	13	22
−6	68	97	74	22	41
−5	**111**	**180**	**124**	31	98
−4	**189**	**257**	**203**	76	**184**
−3	**321**	**326**	**351**	**178**	**281**
−2	**471**	**476**	**558**	**480**	**603**
−1	**749**	**689**	**828**	**1014**	**857**
0	**928**	**770**	**1104**	**1876**	**1151**
1	**763**	**679**	**794**	**945**	**855**
2	**583**	**507**	**510**	**383**	**542**
3	**381**	**377**	**289**	**129**	**295**
4	**253**	**280**	**156**	45	**149**
5	**126**	**170**	88	15	76
6	94	120	52	11	56
7	59	58	31	9	19
8	35	35	5	2	9
9	15	23	5	3	6
10	8	10	5	2	0
11	9	11	3	0	0
12	5	9	2	0	1
13	1	0	1	0	0
14	2	4	0	0	0
15	1	2	0	0	0
16	0	0	0	0	0
17	0	0	0	0	0
18	0	0	0	0	0

Note. Values in boldface are within the 95% confidence interval for each scale based on SE_{msmt} for change score. SE_{msmt} represents the standard error of measurement (fixed observed score formula).

Received March 29, 2006
Revision received November 16, 2006
Accepted December 20, 2006 ■