Psychological Assessment
2003, Vol. 15, No. 3, 333–339

1040-3590/03/$12.00 DOI: 10.1037/1040-3590.15.3.333

Positive Impression Management and Its Influence on the Revised NEO Personality Inventory: A Comparison of Analog and Differential Prevalence Group Designs

R. Michael Bagby
University of Toronto and Centre for Addiction and Mental Health

Margarita B. Marshall
University of Toronto

Participants (n = 22) completed the Revised NEO Personality Inventory (NEO PI–R) as part of an authentic job application. Protocols produced by this group were compared with "analog" participants (n = 23) who completed the NEO PI–R under standard instructions and again under instructions designed to mimic the test-taking scenario of the job applicants (the "fake-good" condition). Participants completing the NEO PI–R under fake-good instructions and the job applicants scored lower on the Neuroticism and higher on the Extraversion scales than did the participants responding under standard instructions. Analog participants in the fake-good condition scored higher on the Extraversion and lower on the Agreeableness scales than did the job applicants. These results suggest that outcomes from analog designs are generalizable to real-world samples where response dissimulation is probable.

A long-held belief among many personality researchers is that self-report instruments designed to measure personality and psychopathology are vulnerable to response bias (Edwards, 1953; Hogan & Nicholson, 1988). It is assumed that in many situations, test takers may be motivated to respond to items in a manner that maximizes a desired outcome. An individual applying for insurance compensation for psychiatric disability may intentionally exaggerate or even fabricate symptoms associated with mental illness in order to procure financial award. This type of responding is commonly referred to as "faking bad," or negative impression management. Others may be in situations in which it is in their best interests to underreport symptoms of mental illness that they experience or claim to possess desirable personality traits that they know to be untrue. One such scenario might be a job applicant seeking employment and therefore motivated to underreport the presence of psychopathology or endorse what he or she believes the potential employer would view as highly desirable traits. This type of responding has been labeled "faking good," or positive impression management.

Most measures of personality and psychopathology include special scales designed to detect the presence and influence of fake-bad and fake-good responding. Widely used instruments such as the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher et al., 2001) and the Personality Assessment Inventory (Morey, 1991), for example, have validity scales, which are used to assess fake-bad and fake-good responding. One notable exception is the Revised NEO Personality Inventory (NEO PI–R; Costa & McCrae, 1992c), which is the most frequently used instrument to assess today's most prominent model of personality—the Five-Factor Model of Personality (FFM). The NEO PI–R was designed to measure the FFM, and it provides five personality domain scores that correspond to five broad dimensions of personality: Neuroticism, Extraversion, Openness-to-Experience, Agreeableness, and Conscientiousness. Although Costa and McCrae (1992c, 1997) have been criticized for not including validity scales on the NEO PI–R (Ben-Porath & Waller, 1992; Butcher & Rouse, 1996), they justify the continued exclusion of such scales on a number of different grounds, including the existence of a third-person (informant) version of the instrument that can be used, among other purposes, when there is suspicion of nonveridical responding (Costa & McCrae, 1992a, 1992b, 1992c) and the finding that correction adjustments based on validity scales rarely increase the predictive validity of individual content and clinical scales (McCrae et al., 1989). Perhaps the most compelling argument set forth by these personality researchers is that the "analog design," typically employed to develop and evaluate the effectiveness of validity scales, has not yet been shown to be generalizable to samples where faking is clearly known or is suspected to have occurred. These latter samples are called "known groups" and "differential prevalence groups," respectively (Rogers, 1997).

Despite Costa and McCrae's clear stance on the use of validity scales, Schinka, Kinder, and Kremer (1997) constructed research validity scales for the NEO PI–R to assess response dissimulation: the Positive Presentation Management (PPM) scale, designed to detect response styles that reflect "claimed uncommon virtues and/or denied common faults" (p. 129) and the Negative Presen-

R. Michael Bagby, Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada, and Centre for Addiction and Mental Health, Toronto, Ontario, Canada; Margarita B. Marshall, Department of Psychology, University of Toronto.

This research was completed as part of an independent study for course credit on the part of Margarita B. Marshall and funded, in part, by a Senior Research Fellowship from the Ontario Mental Health Foundation awarded to R. Michael Bagby.

We thank Andrew G. Ryder and Paul T. Costa Jr. for their comments and suggestions on earlier versions this article.

Correspondence concerning this article should be addressed to R. Michael Bagby, Centre for Addiction and Mental Health, Clarke Site, 250 College Street, Toronto, Ontario M5T 1R8, Canada. E-mail: michael_bagby@camh.net

tation Management (NPM) scale, designed to detect response styles that reflect "claimed uncommon faults and/or denied common virtues" (p. 129). Overall, results from studies evaluating these validity scales suggest that the NPM and PPM are effective in detecting, respectively, fake-bad and fake-good responding (Berry et al., 2001; Caldwell-Andrews, Baer, & Berry, 2000; Reid-Seiser & Fritzsche, 2001; Schinka et al., 1997; Young & Schinka, 2001). All of these studies that provided support for these validity scales, however, used analog research designs.

Although no investigation has examined whether analog samples are similar to differential prevalence group or known-group samples, researchers have assumed, at least implicitly, that one can generalize from analog to differential prevalence group and known-group designs. Such empirically unsubstantiated assumptions have prompted some to suggest that an entire line of research examining fake-bad and fake-good response styles may have committed a Type II error (Piedmont, McCrae, Riemann, & Angleitner, 2000). That is, investigators using analog designs may be assuming falsely that research participants instructed to fake in the experimental context do not differ from persons in known-group and differential prevalence group samples, when in actuality they do. If, however, persons in differential prevalence groups and participants in analog designs who are instructed to fake good perform similarly or differ in similar ways from those in analog designs who respond honestly (i.e., standard instructions), the generalization from analog studies can be more confidently extended to the applied setting.

The goal of this study was to compare directly test results from the NEO PI–R completed by respondents under a condition in which positive impression management was highly likely (differential prevalence group sample) with test results produced by research participants who were provided with information and instruction in an experimental context designed to mimic the same test-taking scenario (analog sample). In particular, we sought to examine two issues. First, do individuals from a differential prevalence group sample produce NEO PI–R personality domain scale elevations similar to those produced by participants in analog research who are instructed to fake good, and do the scale elevations of these two groups differ from those of research participants who take the NEO PI–R under standard (honest) instructions? In addition, do the recently developed research scales designed by Schinka et al. (1997) to assess fake-good and fake-bad responding distinguish among the three groups? To this end, an analog sample composed of aspiring actors solicited from local acting schools first completed the NEO PI–R under standard (i.e., honest responding) instructions and then were readministered this test under instructions to respond in such a way as to maximize their chances of successfully acquiring a role on a local and highly publicized "reality TV" show (i.e., the fake-good condition). The results from these two test-taking conditions were then compared with one another and with the test results from a sample of bona fide reality TV applicants—the differential prevalence group—who had previously completed the NEO PI–R as part of the actual selection procedure.

We hypothesized that the analog research participants responding in the standard instruction condition would score higher on the Neuroticism and lower on the Extraversion scales than when responding under fake-good instruction; the same pattern of results should emerge in the comparison of those analog research participants responding in the standard instruction condition and those participants in the differential prevalence group sample. No differences in scores on Openness-to-Experience, Agreeableness, and Conscientiousness were expected, as results from previous investigations have not revealed reliable differences across standard and fake-good instructions for these personality traits (Ballenger, Caldwell-Andrews, & Baer, 2001; Caldwell-Andrews et al., 2000).

We also hypothesized that research participants responding in the condition in which they were instructed to fake-good would score significantly higher on the PPM and lower on the NPM validity scales than when responding under standard instructions; this hypothesis was in line with results from some previous studies that used analog research designs (e.g., Ballenger et al., 2001; see also Baer & Miller, 2002, for a review). Again, under the assumption that analog designs are comparable to differential prevalence group samples, we also predicted that the participants in the differential prevalence group sample would score higher on the PPM and lower on the NPM than those in the analog sample, who completed the NEO PI–R under standard instructions.

Method

Participants

The differential prevalence group sample consisted of 25 finalists from a competition to become hosts for a popular reality TV show based in a major North American city. From this sample of 25, 8 were to be selected. As part of the selection process, all applicants were required to complete a psychological evaluation. At the time of their assessment, informed consent was obtained from all participants, indicating their awareness of the purpose of testing. Because the results of these measures would have an impact on their success or failure to obtain one of the eight positions as a TV "host," there was a strong likelihood that the respondents were highly motivated to present themselves in a manner that they believed would enhance their selection.[1] Thirteen of the applicants were men, 11 were women, and 1 was in the process of changing from a woman to a man. Twenty-three of the 25 applicants were single and had never been married, and applicants were between 19 and 29 years of age ($M = 23.0$ years, $SD = 3.16$). One participant was a Canadian "First Nations" person (this term is the accepted designation for indigenous peoples of Canada), 2 participants were Canadians of African descent, and the remaining 23 (92%) were Canadians of European descent. Twenty-one of the 25 applicants (84%) were currently employed at the time of testing, and 14 of these 21 indicated that they were working as actors or performers. All but one (who had a high school degree) had at least 4-year university degrees.

The analog sample was recruited through postings on Internet discussion boards for local actors on the basis of the advertisement used to recruit the applicants in the differential prevalence group. Inclusion and exclusion criteria for this group were based on the demographics of the differential prevalence group sample. The inclusion criteria were that participants be

[1] As noted, the applicants in the differential prevalence sample were informed that their responses on the personality measure were to be used as part of the selection process. There are two reasons to believe that applicants in the differential prevalence sample were highly motivated to alter their responses. First, participants on other reality TV programs (e.g., "Survivor") have received positive career benefits as a result of media exposure. Second, applicants in the present study were aware that as part of their job, they would be required to conduct interviews with well-known stars in the entertainment industry as well as to create and host their own programs. Such experiences would considerably enhance the resumes of successful applicants and have a great impact on their careers.

between 19 and 29 years of age and currently employed or seeking employment as an actor or other performer at the time of testing. Applicants were excluded if their current or usual occupation (full or part time) was not acting or performing.[2] Twelve of the analog participants were men, and 16 were women. Of the 28 participants, 25 were single and had never been married, and 3 participants were married. Two participants in this sample were Canadians of African descent, and the remaining 26 (93%) applicants were Canadians of European descent. The mean age of the research participants was 23.8 years ($SD = 3.10$). All of these participants had university degrees.

Between-groups comparisons were conducted by means of *t* tests and the chi-square statistic to determine if the analog sample differed from the differential prevalence sample on the basis of demographic variables. None of the demographic variables differed significantly between the analog and differential prevalence samples for sex, marital status, age, race/ethnicity, or employment status.

Measurement

Personality domain scales. The NEO PI–R was used to assess the five personality domains of the FFM. The NEO PI–R comprises 240 self-report items answered on a 5-point Likert format scale, with separate scales for each of the five domains. Each scale consists of six correlated facets or subscales with 8 items, for a total of 48 items for each scale. In this study, only the five personality domain scales were used.

Validity scales. Two NEO PI–R research validity scales developed by Schinka et al. (1997), the PPM and NPM, were used to assess response style. The PPM and NPM scales comprise items selected from the pool of 240 NEO PI–R items using an empirical/rational scale strategy (see Schinka et al., 1997, Study 1).[3] Both the PPM and NPM scales comprise 10 items; 6 of the 10 of the items on the PPM and all 10 items on the NPM are negatively keyed. The PPM scale includes 2 Neuroticism items, 3 Extraversion items, 3 Openness-to-Experience items, 1 Agreeableness item, and 1 Conscientiousness item. Two items from each of the five personality domains reside on the NPM scale.

Procedure

Persons who responded to these advertisements had the study explained to them by telephone, and if they agreed to participate, a testing session was booked. Written informed consent was obtained from participants upon their arrival at the session. The number of participants in each session ranged from 2 to 8 people. Sessions lasted approximately 4 hr and were completed in two parts.[4] In the first part respondents were administered the tests under the standard instructions and told that after completing the first testing session, they would be taking the tests again in a second session under different (but unspecified) instructions. After completing the first part of the session, they had a 30-min break (to minimize the effect of fatigue and boredom) and were given $5 for a beverage and a snack. When the participants returned from the break, they were given the tests a second time, with the instruction to respond to the questions as if they were trying to maximize their chances of gaining a role on a popular and local reality TV show. The instructions were as follows:

> Imagine that you are auditioning for a job as a television host for ["Name of Show"] and that as part of the selection and audition process you must undergo psychiatric examination and psychological testing.[5] We would now ask that you answer the questions on the psychological test in a manner that you believe would enhance your chances of being selected to be a television host. One thing to keep in mind is that you want to respond in a manner that is believable, but at the same time enhance the potential of your being selected. Remember that the research participant in this study who produces the most believable and desirable profile will receive an extra $100.

We then explained that test results in this second session would be compared with those of a sample of individuals who had taken these tests as part of their application for the actual reality TV show; the respondent who produced a personality profile that most closely matched the "averaged" personality profile of the bona fide applicants would be awarded $100, in addition to that individual's remuneration for participating. Following the completion of the second testing session, all research participants were paid $75.

Results

Protocol and Data Screening

Protocols were first screened for acquiescence, nay-saying, random responding, and incomplete response sets according to the guidelines in the NEO PI–R manual (Costa & McCrae, 1992c). In the analog sample, under standard instructions, 1 respondent was found to have engaged in acquiescence, 1 engaged in random responding, and 1 did not complete the protocol. Protocols from these research participants (i.e., protocols completed by them under both instruction conditions) were removed from the sample. Similarly, two additional protocols were dropped from the analog sample, as participants in the fake-good instruction condition had extreme personality domain scale scores that deviated markedly from the rest of the analog sample in the fake-good instruction condition.

For the differential prevalence group sample, 3 protocols were removed: 1 for lack of clarity regarding which norms to use to score the NEO PI–R protocol (i.e., the individual involved in a sex-change operation), and 2 because of random responding. In sum, a total of 68 test protocols remained—46 from the analog sample (23 under standard and 23 under fake-good instructions) and 22 in the differential prevalence group sample.

Experimental Manipulation Check

Rogers (1997) strongly recommended that research participants in analog dissimulation studies be questioned on their understanding of the instructions, as some studies have shown that many participants may not have comprehended them adequately. In order to determine if the analog research participants understood the instructions to fake-good provided in that instructional condition, we administered a postexperimental questionnaire, which included two open-ended discussion topics: (a) "In your own

---

[2] None of the analog participants was a finalist for the first season of the program, from which the differential prevalence group sample was drawn. However, 2 participants in the analog sample had applied for positions on the second season, and of these, 1 was a finalist.

[3] A third research scale, the Inconsistency scale, was also created by Schinka et al. (1997) for the purpose of detecting random responding. However, because the main purpose of the present study was to examine impression management, this scale was not used in the analyses.

[4] Participants also completed the MMPI–2 (Butcher et al., 2001), the Fundamental Interpersonal Relations Orientation—Behavior (Schnell & Hammer, 1993), and the BarOn Emotional Quotient Inventory (Bar-On, 1997); however, these measures were not included for analysis in the present article.

[5] The name of the TV show in question is withheld in order to preserve anonymity of respondents.

words, please explain the instructions you received for the second half of the study" and (b) "Please briefly explain what strategies you employed to respond as if you were applying for the [Name of Show]." Two students (1 graduate student and 1 advanced undergraduate student) then reviewed and rated the responses from each of the participants and used the following scheme to code the answers: "Did not appear to have understood the instructions," "most likely understood the instructions," and "definitely understood the instructions." Both raters agreed that 21 of the 23 research participants either most likely or definitely understood the instructions on the basis of their responses to the first topic; one rater thought that 2 participants did not appear to understand the instructions, whereas the other rater thought these same 2 participants most likely understood the instructions. A review of the responses to the second topic, however, revealed that all subjects articulated a strategy reflecting that they most likely understood or definitely understood the instructions. Thus, no analog participant was eliminated on the basis of failure to understand instructions.

Mean Differences

Personality domain scales. Three sets of planned comparisons for the domain scales were performed with *t* tests. The first set of planned comparisons consisted of a within-group (repeated) analysis (i.e., analog standard instructions vs. analog fake-good instructions). The next two sets of planned comparisons were between groups (i.e., analog sample/standard instructions vs. differential prevalence group sample; and analog sample/fake-good instructions vs. differential prevalence group sample). Because most comparisons were based on a priori assumptions, Bonferroni correction was not applied, and the *p* value was set at .05. Effect size calculations (Cohen's *d*; Cohen, 1988) were used to supplement the mean difference analyses.

The means and standard deviations for the personality domain scores for each of the three groups are displayed in the upper portion of Table 1. In the repeated analysis, the analog sample research participants responding under fake-good instructions as opposed to those responding under standard instructions scored significantly lower on the Neuroticism scale, $t(22) = 3.21$, $p < .01$, $d = -1.37$, and significantly higher on the Extraversion scale, $t(22) = 6.02$, $p < .01$, $d = 2.57$, and the Conscientiousness scale, $t(22) = 2.43$, $p < .05$, $d = 1.04$, respectively. For the between-groups analysis in which the analog research participants responding under standard instructions were compared with the differential prevalence sample, the former group scored significantly higher on the Neuroticism scale, $t(43) = 4.16$, $p < .01$, $d = -1.18$, and significantly lower on Extraversion scale, $t(43) = 2.98$, $p < .01$, $d = 0.91$, than the latter group. For the between-groups analysis in which the analog research participants responding under instructions to fake-good were compared with the differential prevalence group sample, the former group scored significantly higher on the Extraversion scale, $t(43) = 4.16$, $p < .01$, $d = -1.27$, and significantly lower on the Agreeableness scale, $t(43) = 2.34$, $p < .05$, $d = 0.72$, than did the latter group.

Validity scales. The means and standard deviations for the research validity scales (the PPM and NPM) are displayed in the lower portion of Table 1. For the PPM, there was a significant difference between the analog research participants' responding under standard instructions compared with their responding under fake-good instructions, $t(22)$ 2.67, $p < .01$, $d = -1.14$; however, contrary to the hypothesis, those responding under standard instructions scored *higher* than those responding under fake-good

Table 1
Means, Standard Deviations, and Effect Sizes for Domain and Validity Scales of the Revised NEO Personality Inventory

	Analog design				Differential prevalence group		Cohen's *d*		
	Standard instructions		Fake-good instructions						
NEO PI–R	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*1	*d*2	*d*3
Domain scales									
Neuroticism	58.93_a	12.55	46.55_b	15.88	46.66_b	8.48	−1.37	−1.18	0.01
Extraversion	58.67_a	10.80	77.04_b	8.73	66.96_c	7.44	2.57	0.91	−1.27
Openness	67.06	10.40	67.42	10.18	65.26	11.08	0.06	−0.17	0.21
Agreeableness	42.08	13.19	36.18_a	18.08	46.68_b	11.07	−0.48	0.39	0.72
Conscientiousness	44.56	11.17	51.90	15.49	48.32	9.68	1.04	0.37	−0.28
Validity scales									
PPM	23.34_a	2.74	21.87_b	2.47	21.36_b	2.72	−1.14	−1.19	−0.12
NPM	7.57	3.10	7.39	5.04	7.18	3.02	0.40	−0.13	−0.21

Note. Values for the personality domains scores are standardized (T) scores ($M = 50$, $SD = 10$); values for the PPM and NPM are raw scores. Row means with different subscripts are significantly different at minimum $\alpha = .05$. NEO PI–R = Revised NEO Personality Inventory; *d*1 = analog/standard instructions sample vs. analog/fake-good instructions sample; *d*2 = analog/standard instructions sample vs. differential prevalence group; *d*3 = analog/fake-good instructions sample vs. differential prevalence group; PPM = Positive Presentation Management scale; NPM = Negative Presentation Management scale.

instructions. The PPM score was significantly higher for the analog research sample responding under standard instructions than for the differential prevalence group, $t(43) = 2.44$, $p < .05$, $d = -1.19$, a finding also contrary to what was hypothesized. The PPM scale scores did not differ significantly between the differential prevalence group sample and analog fake-good group. There were no significant differences between any of the groups for the NPM.

Discussion

The primary goals of this study were (a) to attempt to replicate results from previous (analog) studies that have used the traditional repeated measures analog design to investigate the influence of fake-good responding on the personality domain scores on the NEO PI–R and (b) to compare NEO PI–R personality scale scores of the research participants in the analog sample completed under standard and fake-good instructions with those completed by individuals in the differential prevalence group sample. Comparing the NEO PI–R domain scale scores between these pairs of groups begins, we believe, to address the issue of whether results from analog design samples are generalizable to differential prevalence group samples. This issue of generalizability is critical because most research directed at developing, refining, and validating "validity" scales relies almost exclusively on analog designs, and results from analog designs have significant implications for the effective and meaningful use of validity scales in applied contexts.

Consistent with results from previous analog design studies (Reid-Seiser & Fritzsche, 2001; Ross, Bailley, & Millis, 1997; Rosse, Stetcher, Miller, & Levin, 1998), and supporting one of our hypotheses, the results from the present study demonstrate that research participants decreased their Neuroticism scores and increased their Extraversion scores when instructed to fake-good when completing the NEO PI–R compared with when they completed this test under standard instructions (i.e., responding honestly). Also as hypothesized, the differential prevalence group sample scored significantly lower on Neuroticism and higher on Extraversion than did the analog sample responding under standard instructions. The analog research participants responding under fake-good instructions also scored higher on Conscientiousness than when responding under standard instructions, although this had not been hypothesized. There was no statistically significant difference between the analog research participants responding under standard instructions and the differential prevalence group in terms of their scores on the Conscientiousness scale, but the difference was in the predicted direction. Because studies examining fake-good responding that employ within-group designs typically produce larger effect size differences than between-groups designs (Baer & Miller, 2002), as was the finding in the current study, it is conceivable that a larger sample would have detected a between-groups effect.

Because the analog fake-good instruction group and the differential prevalence group differed similarly from the analog standard instruction group for three of the five personality domains (most with medium-to-large effect size differences), we believe there is some evidence to suggest that results from fake-good analog designs can be generalized to differential prevalence group designs. The pattern of these results points not only to the overall effectiveness of the experimental manipulation typically employed in analog studies but also to the overall similarity between the analog fake-good condition and the differential prevalence group sample.

At the same time, a number of unexpected differences emerged among the groups. Although both the analog research participants responding under the instructions to fake-good and the participants in the differential prevalence group scored significantly higher on the Extraversion scale than did the analog research participants responding under standard instruction, the Extraversion scores of the analog fake-good group also exceeded those of the differential prevalence group. The analog fake-good research participants also scored lower on Agreeableness compared with the differential prevalence sample, although there was no significant difference in Agreeableness scores between the analog research participants responding under standard and fake-good instructions.

These results might be attributed to the fact that the television show in question was already "on the air" at the time the analog group was tested. This temporal difference across the analog and differential prevalence group samples may have allowed the analog respondents to become more familiar with the kind of personalities most likely to be hired, compared with the bona fide applicants, who had no access to such information because the show had not been previously aired. In general, extraversion characterized most of the actors, and antagonism and competitiveness (i.e., low Agreeableness) among the participants were two of the most salient features of the show. Thus, the analog sample under instructions to respond in order to maximize their chances of getting on the show may have perceived excessive extraversion and low agreeableness to be highly desirable in this context.

No mean differences were detected with respect to the Openness-to-Experience dimension across any of the three groups. Previous studies have demonstrated inconsistent differences across standard and analog fake-good instructions with respect to Openness-to-Experience. For example, Ballenger et al. (2001) found that Openness-to-Experience did not distinguish analog fake-good research participants from honest respondents in a clinical sample. Conversely, Caldwell-Andrews et al. (2000) found that Openness-to-Experience scores did differ significantly across analog fake-good and honest conditions. One way to make meaning of these discrepant results, including those of the current investigation, is to consider that the scenarios provided in the experimental instructions or the perceived demands of the real-life assessment situations may elicit context-specific desirable traits. For example, in one situation, high scores on Openness-to-Experience might be seen as a particularly important characteristic, whereas in another context, low or high scores on Agreeableness might be seen as the most desirable. This interpretation echoes Rogers's (1997) caution that careful attention be paid to the specificity of instructions when designing dissimulation studies and generalizing results from them. Perhaps what can be said at this point is that in most contexts in which some form of fake-good responding can be expected, Neuroticism scores are likely to be decreased; Extraversion and, to a lesser extent, Conscientiousness scores are likely to be increased. Openness-to-Experience and Agreeableness domain scores are apt to be more variable and situation specific.

In this study the performance of the research validity scales developed by Schinka and colleagues (1997) to detect fake-good responding was much less than optimal. The failure to replicate the findings of prior work regarding the validity of these scales sug-

gests the possibility that the experimental manipulation employed in the analog portion of the current study was unsuccessful. This interpretation, however, seems unlikely, as consistent group differences between the fake-good and standard instructions conditions and between the standard instructions and differential prevalence group samples emerged for the Neuroticism, Extraversion, and Conscientiousness domain scales. In a similarly designed, repeated measures analog study, Ballenger et al. (2001) also reported that the PPM was unable to distinguish NEO PI–R protocols completed under fake-good versus standard instructions, although the balance of the evidence suggests that the PPM can make such distinctions (see e.g., Caldwell-Andrews et al., 2000; Schinka et al., 1997; Young & Schinka, 2001).

As indicated earlier, one explanation for discrepant results across different studies may be located in the instructions provided to participants in fake-good conditions, which may differentially influence scale elevations on both the domain and validity scales of the NEO PI–R. There is evidence to suggest, for example, that the PPM is more highly correlated with measures of self-deceptive enhancement than with impression management (e.g., Reid-Seiser & Fritzsche, 2001). Instructions (or assessment situations) that elicit overt attempts to engage in positive impression management, as was the case in this study, may not produce strong effects for a scale more sensitive to self-deception. More studies are needed to explore these issues, and should future evidence emerge supporting the need for NEO PI–R validity scales, we believe that separate scales for these two types of fake-good responding should be developed. We also think that careful consideration should be given to the development of any validity scale that is composed *exclusively* of items that reside on personality domain scales, because items designed to assess personality traits that have proven construct validity are unlikely to provide unequivocal meaning with respect to response dissimulation, especially with scales composed of relatively few items.

Several limitations of the current study must be acknowledged. First, the fact that the analog sample was tested almost 10 months later than the differential prevalence group sample may have contributed to some of the differences across these two groups. Another limitation is that the differential prevalence group sample was only assessed on one occasion. Ideally, it would have been best to assess the individuals in this sample a second time, either before or after the selection process. Personality assessment results outside of the context of the selection process for these applicants would have permitted a more definitive conclusion regarding the generalizability of analog and differential prevalence group designs and samples. Every effort was made, however, to match the research participants in the analog sample demographically (i.e., age, sex, education) and vocationally (i.e., career choice and aspirations) to the individuals who composed the differential prevalence group sample. This was done with the expectation that closely matched groups would collectively produce similar personality profiles. There is the possibility, nonetheless, that these two groups did have different "baseline" personality profiles. We believe, however, that it is unlikely that this potential difference could account for the outcomes of the current study, as the specific patterns of results obtained for the differential prevalence group sample and the analog group responding to fake-good instructions, relative to the same analog research participants responding to standard (honest) instructions (i.e., lower on Neuroticism, higher on Extraversion and Conscientiousness), were too similar. Moreover, this same pattern of personality scale alteration has been observed in previous analog studies.

We also recognize that the sample size was small and that this limitation not only compromised the generalizability of the results but also prohibited an analysis of the NEO PI–R facet scales. Because facet scales have proven to be excellent predictors of behavior and offer a more fine-grained personality profile than do domains scales (see, e.g., Reynolds & Clark, 2001), future studies with large samples examining facet scores would certainly prove useful. Finally, as no study prior to the current investigation has directly compared differential prevalence group samples to analog samples instructed in a manner to mimic real-world scenarios, future studies comparing analog samples with other types of assessment scenarios are needed. Such efforts will begin to clarify the relation between results from analog samples and research designs and those from differential prevalence group samples, and perhaps even from known-group samples. Accumulation of results from such studies will, in time, address the important issue of the external validity of the widely employed experimental approach to the study of test response dissimulation.

Notwithstanding the need for replication and extension, it is worth emphasizing that the results from the current investigation do suggest that outcomes from analog design studies are likely generalizable to real-world settings. Against this background, the accumulated evidence from previous analog studies and from the current investigation, which indicate that test profiles are altered under instructions to fake-good, has implications for the use of the NEO PI–R and other tests like it. In assessment contexts in which positive impression management is likely, such as personnel selection, the assessor needs to be cognizant that such instruments are susceptible to impression management response bias. Of course, this becomes especially problematic if the instruments used do not have scales that accurately assess the presence of response bias. One solution would be to use multiple tests or other evaluation strategies in which one of the scales or methods assesses for potential response bias. For example, the NEO PI–R and the MMPI–2, the latter of which has scales that can detect positive impression response style, could be used in combination. The NEO PI–R would provide extensive information on a variety of personality trait attributes relevant to job performance, whereas the MMPI–2 would provide information about potential psychopathology and possible response bias. We believe that although the general cost (in time and money) might be perceived as excessive, the potential benefits from such a comprehensive assessment relative to the costs tip the cost–benefit ratio in favor of such extensive testing (see, e.g., Meyer et al., 2001).

References

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI–2: A meta-analytic review. *Psychological Assessment, 14,* 16–26.

Ballenger, J. F., Caldwell-Andrews, A., & Baer, R. A. (2001). Effects of positive impression management on the NEO Personality Inventory—Revised in a clinical population. *Psychological Assessment, 13,* 254–260.

Bar-On, R. (1997). *BarOn Emotional Quotient Inventory: A measure of emotional intelligence.* Toronto, Ontario, Canada: Multi-Health Systems.

Ben-Porath, Y. S., & Waller, N. G. (1992). Normal personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4,* 14–19.

Berry, D. T. R., Bagby, R. M., Smerz, J., Rinaldo, J. C., Caldwell-Andrews, A., & Baer, R. A. (2001). Effectiveness of NEO PI-R research validity scales for discriminating analog malingering and genuine psychopathology. *Journal of Personality Assessment, 76,* 496–516.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory–2: Manual for the administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47,* 87–111.

Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO PI-R scores and their relations to external criteria. *Journal of Personality Assessment, 74,* 472–488.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Costa, P. T., & McCrae, R. R. (1992a). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13.

Costa, P. T., & McCrae, R. R. (1992b). "'Normal' personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory": Reply. *Psychological Assessment, 4,* 20–22.

Costa, P. T., & McCrae, R. R. (1992c). *Professional manual for the NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1997). Stability and change in personality assessment: The NEO Personality Inventory in the year 2000. *Journal of Personality Assessment, 68,* 86–94.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37,* 90–93.

Hogan, R., & Nicholson, R. A. (1988). The meaning of personality test scores. *American Psychologist, 43,* 621–626.

McCrae, R. R., Costa, P. T., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine, 51,* 58–65.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

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.

Reid-Seiser, H. L., & Fritzsche, B. A. (2001). The usefulness of the NEO PI-R Positive Presentation Management Scale for detecting response distortion in employment contexts. *Personality and Individual Differences, 31,* 639–650.

Reynolds, S. K., & Clark, L. A. (2001). Predicting dimensions of personality disorder from domains and facets of the five-factor model. *Journal of Personality, 69,* 199–222.

Rogers, R. (1997). *Clinical assessment of malingering and deception* (2nd ed.). New York: Guilford Press.

Ross, S. R., Bailley, S. E., & Millis, S. R. (1997). Positive self-presentation effects and the detection of defensiveness on the NEO PI-R. *Assessment, 4,* 395–408.

Rosse, J. G., Stetcher, 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.

Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research validity scales for the NEO PI-R: Development and initial validation. *Journal of Personality Assessment, 68,* 127–138.

Schnell, E. R., & Hammer, A. (1993). *Introduction to the FIRO–B in organizations.* Palo Alto, CA: Consulting Psychologists Press.

Topping, G. D., & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO–FFI. *Personality and Individual Differences, 23,* 117–124.

Young, M. S., & Schinka, J. A. (2001). Research validity scales for the NEO PI–R: Additional evidence for reliability and validity. *Journal of Personality Assessment, 76,* 412–420.

Received November 4, 2002
Revision received March 25, 2003
Accepted April 2, 2003 ■