JOURNAL OF PERSONALITY ASSESSMENT, *76*(3), 496–516

Effectiveness of NEO–PI–R Research Validity Scales for Discriminating Analog Malingering and Genuine Psychopathology

David T. R. Berry
Department of Psychology
University of Kentucky

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

Jessica Smerz, Jason C. Rinaldo, Alison Caldwell-Andrews, and Ruth A. Baer
Department of Psychology
University of Kentucky

We investigated the research validity scales for the NEO Personality Inventory–Revised (NEO–PI–R) proposed by Schinka, Kinder, and Kremer (1997): Positive Presentation Management (PPM) and Negative Presentation Management (NPM). Additionally, an experimental analog to the Minnesota Multiphasic Personality Inventory–2's (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) *F – K* index was calculated by subtracting the raw score on PPM from the raw score on NPM (NPM – PPM). In 2 studies, all indexes showed significant between-group differences when samples of analog malingerers ($n = 97$) were contrasted with psychiatric outpatients ($n = 272$). The sensitivity and specificity of these validity indexes indicated that although none performed well in extremely low base rate environments, the NPM and NPM – PPM indexes showed promise when the base rate of faking bad rose to higher levels.

The Five-factor model of personality (FFM) has enjoyed growing acceptance among personality researchers in recent years (Costa & McCrae, 1995; Goldberg & Saucier, 1995; Widiger & Trull, 1997). This model proposes that five major factors

capture the majority of important variance in personality characteristics. These factors include Extraversion (or Positive Emotionality; E), Neuroticism (or Negative Emotionality; N), Agreeableness (A), Conscientiousness (C), and Openness to Experience (or Intellect; O). The most widely evaluated and accepted measure of the FFM is the revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992b), which assesses five domains corresponding to the factors of the FFM as well as 30 facets (six more specific correlated facets within each domain). The FFM and its operationalization in the NEO–PI–R have a number of strengths, such as striking a balance between comprehensive personality description and oversimplification, serving as a common foundation and a useful reference point for communication among researchers and clinicians, and offering a framework for the somewhat amorphous task of personality assessment that is easily translated into assessment feedback.

The utility of the NEO–PI–R for describing normal personality dimensions has led to the suggestion that information provided by the inventory might be useful for addressing a variety of clinical questions and issues (Piedmont, 1998). For example, Costa and McCrae (1992a) proposed that data from the inventory might assist therapists in understanding clients, formulating diagnoses, establishing rapport, anticipating the course of treatment, and selecting the optimal form of treatment. However, the recommendation that the NEO–PI–R be widely used in clinical settings was questioned on several grounds by Ben-Porath and Waller (1992). Among the concerns raised by these authors was that: "A first and essential step in any clinical assessment is the determination of the quality of the data that will serve as the basis for the evaluation" (p. 14). Because the NEO–PI–R includes no validity checks beyond a single self-report question baldly asking if the test taker answered the questions honestly and accurately, Ben-Porath and Waller argued that the test should not be used as a stand-alone measure in clinical assessment, although leaving open the possibility that it could be used to supplement other inventories that included validity scales.

To some extent, judgments regarding the importance of the omission of validity scales on the NEO–PI–R may turn on the base rates of response distortion in the setting in which the test is being used. As has long been recognized, even a highly accurate indicator of the presence of a particular condition may actually decrease classification accuracy if used when the condition in question has a low base rate (Meehl & Rosen, 1955). Thus in settings where the base rate of response distortion is likely to be very low, such as anonymous research projects using volunteers or certain types of clinical work in which the client and clinician have closely aligned interests (e.g., helping the client to improve), even a highly accurate indicator of the presence of a response set may be surpassed in overall accuracy by simply playing the base rates and predicting that no test taker has a response bias.

Although to date the NEO–PI–R has apparently been used primarily in settings where the base rates of invalid responding would seem likely to be very

low, some clinical settings involve work with individuals who potentially have very strong reasons to manipulate their responses to achieve a desired presentation. Strong, Greene, and Schinka (2000), using taxometric analyses, estimated base rates of faking bad on the Minnesota Multiphasic Personality Inventory–2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) in clinical settings as approximately 27% for psychiatric inpatients and 19% for general inpatients at Veterans Affairs medical center settings. Frederick, Crosby, and Wynkoop (2000) reported data from a well-validated test of response sets administered to 737 male pretrial defendants referred to the mental health evaluation service of the U.S. Medical Center for Federal Prisoners. They found that 43.7% of these forensic evaluees produced invalid performances on the Validity Index Profile (VIP; Frederick, 1997). These reports suggest that in at least some clinical referral streams where important consequences hinge on the outcome of a psychological evaluation, the base rate of noncompliant approaches to testing may be significant.

Another issue that arises in considering the issue of validity scales for the NEO–PI–R is the extent to which the test is vulnerable to feigning. Rogers (1997, p. 379) concluded that nearly all psychological measures are susceptible to response sets. More directly, several studies have addressed the question of whether the NEO–PI–R is vulnerable to deliberate attempts to manipulate one's presentation on the test. Paulhus, Bruce, and Trapnell (1995) found that individuals taking the NEO–FFI (Five-Factor Index; a short version of the NEO–PI–R that produces only domain scores) altered their domain scores systematically as their instructions to minimize problems changed. Scandell and Wlazelek (1996) reported orderly differences in domain scores from the NEO–FFI depending on whether the instructions participants were given involved faking good, faking bad, or answering honestly. Furnham (1997) demonstrated that domain scores on the NEO–FFI were highly susceptible to feigning. Ross, Bailley, and Millis (1997) used the NEO–PI–R in an analog faking study and concluded that it was clearly vulnerable to faking. Yang, Bagby, and Ryder (2000) found that Chinese psychiatric inpatients with evidence for faking good on the NEO–PI–R had significantly greater differences between self-ratings and spouse ratings for the N and E domains than those who did not have such evidence. Caldwell-Andrews, Baer, and Berry (2000) found significant differences in domain scores on the NEO–PI–R between groups given standard instructions versus instructions to fake bad or fake good. Across these studies, a common pattern for faking bad involved elevating N and suppressing all or most of the remaining domain scores. Additionally, Caldwell-Andrews et al. reported that the criterion-related validity of NEO–PI–R results, as indexed by independent measures, was significantly reduced in groups instructed to feign relative to groups given standard instructions. Thus, the available evidence suggests that the NEO–PI–R and NEO–FFI may be significantly affected by a test taker's conscious attempts to alter his or

her presentation on the instruments, and furthermore, that such distortions may adversely affect the criterion-related validity of the test.

Responding to the perceived need for response set indicators on the NEO–PI–R, Schinka, Kinder, and Kremer (1997) developed research validity scales from the item pool of the NEO–PI–R. Three scales, developed through a combination of rational, empirical, and psychometric methods, were shown to distinguish between protocols produced under varying instructions. One scale is designed to detect inconsistent responding (INC), another to detect positive presentation management (PPM), and most important for this article, a third scale is intended to detect negative presentation management (NPM). These scales are not used to adjust scores on the domains, as is done with the MMPI–2 *K* scale, but rather to signal the presence of invalid approaches to the test.

In addition to examining changes in domain scores and criterion-related validity under response sets, Caldwell-Andrews et al. (2000) also evaluated the utility of the PPM and NPM scales for detecting response sets in their analog study. They identified optimal cutting scores for their groups and found that PPM and NPM were moderately effective at discriminating honest from fake-good and fake-bad responding, with PPM's sensitivity at .77 and specificity at .81, and NPM's sensitivity at .81 and specificity at .90 (definitions of these parameters are given later).

Although Caldwell-Andrews et al.'s (2000) results are supportive of the research validity scales, one important weakness of the study was the absence of genuine psychiatric patients. Meta-analyses of the fake-bad literature on the MMPI and MMPI–2 have reported that validity scale differences between students instructed to feign psychopathology and students answering honestly are much larger than those contrasting feigning students and psychiatric patients answering honestly (Berry, Baer, & Harris, 1991; Rogers, Sewell, & Salekin, 1994). One implication of this finding is that the ability of fake-bad scales to discriminate individuals with genuine psychopathology from those feigning psychopathology will likely be overestimated by studies that fail to include genuine psychiatric patients. Thus, if the NEO–PI–R research validity scales are to be applied in clinical populations, it is important to evaluate the impact of psychopathology on them. In this article we address this issue through two studies contrasting analog feigners with psychiatric patients taking the NEO–PI–R as part of a clinical evaluation.

STUDY 1

Method

Participants. Participants in the first study included psychiatric outpatients and college undergraduates. The initial clinical sample consisted of 298 outpatients

seen at a psychiatric facility in Canada (domain and facet scale data from some of these patients have previously been reported by Bagby et al., 1997; Bagby et al., 1999). These participants, none of whom was involved in compensation seeking, had completed the NEO–PI–R as part of a routine psychological evaluation prior to treatment and their data were retrieved from files and compiled anonymously. Diagnoses, made independent of NEO–PI–R results and described in the following, were based on structured clinical interviews and *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [*DSM–IV*]; American Psychiatric Association, 1994) criteria. None of the patients had answered the NEO–PI–R validity question in a way that raised concerns about the veracity of results. Results from the patients were screened by applying the algorithms described by Costa and McCrae (1992b, p. 42) to identify random, acquiescent, and counteracquiescent response sets. This resulted in elimination of 19 patients for evidence of acquiescence in their protocols and 5 for evidence of random responding. Two other patients were excluded from analyses due to technical problems with their data. Thus, of the 298 original patient protocols, 272 remained for analysis.

The 272 postscreening NEO–PI–R protocols from psychiatric outpatients were randomly divided into two groups of 136 each, with one group reserved for Study 2 and described later. For Study 1, the patient standard instruction group (PS) included the 136 patients from the first group, whose diagnoses included unipolar depression ($n = 98$), bipolar disorder, most recent episode depressed ($n = 19$), and schizophrenia ($n = 19$). The average age of this sample was 39.4 years ($SD = 11.28$), and 44% were men. Data on educational achievement and race were not available.

Undergraduates initially included 164 University of Kentucky Introduction to Psychology students participating for course credit. The mean age of this sample was 19.5 years ($SD = 1.69$), the mean education level was 13.5 years ($SD = .89$), 40.8% were men, and 93.3% were White (5.5% African American, 1.2% other race). Due to financial constraints, a smaller number of participants were assigned to the fake-bad group, which was offered monetary incentives for successful feigning, as described later. Of the undergraduate sample, 103 participants had been randomly assigned to the standard instruction control group (student standard; SS), and 64 had been randomly assigned to the fake-bad group (student faking; SF).

Materials. Standard NEO–PI–R question booklets and answer sheets were used for all participants. In addition to the NEO–PI–R, student participants also completed an informed consent agreement, a demographic questionnaire, and a posttesting manipulation questionnaire assessing understanding of and compliance with instructions. The posttesting questionnaire asked them to (a) summarize their instructions, (b) list any strategies used to answer NEO–PI–R questions, (c) rate on a 10-point Likert scale ranging from 1 (*low*) to 10 (*high*) how well they understood their instructions, (d) rate on the same scale how well they followed their instructions, and (e) report how many random responses they gave on the test. Following

the study, all students were given an informational sheet explaining the study and a document conferring research participation credit.

Procedure. Data were gathered from the undergraduate sample in groups of 3 to 10 people. Each student initially completed the informed consent agreement and the demographic questionnaire. Next, they were provided the instruction sheet appropriate for their group (discussed later) and given time to review it. They then went on to answer the NEO–PI–R questions. Following the completion of the NEO–PI–R, students were given the manipulation-check questionnaire. Finally, all student participants were given the information sheet explaining the study and the research credit document.

Instructions for the two student groups were as follows:

> SS: Please answer the questions so as to provide an accurate description of yourself.
>
> SF: Imagine that you were one of the federal employees involved in the Oklahoma City bombing. Your experience has left you so psychologically affected that you are unable to work and you are now seeking disability status through the court. You have been asked to undergo psychological testing as part of the court procedure. You wish to appear as psychologically disturbed as possible so that you will win disability status and compensation. Answer each item on the test in such a way that you will look psychologically disturbed, but be sure to do this in a believable manner.

Those participants assigned to the SF condition were informed that those among the half of their group who best imitated a disturbed employee would each receive $20 in addition to course credit. Following completion of data collection, each student in the SF group received a faking success score computed by subtracting the *T* score for their NPM value from the *T* score for their N domain score. Those who fell in the upper half of the distribution were mailed checks.

Data from student participants who failed various criteria were excluded from analyses to increase internal validity. The following exclusion rules were applied: (a) a score of 4 or less on the 10-point Likert ratings of comprehension and adherence to instructions; (b) 25 or more self-reported random responses (> 10% of questions); (c) exclusion rules from the NEO–PI–R manual for acquiescence, nay-saying, and random responding; and (d) failing to answer the NEO–PI–R validity question or answering it negatively. Thirteen student participants were eliminated for giving a rating of 4 or less on the posttesting feedback questions regarding comprehension of or adherence to instructions; 4 for failing the Costa and McCrae (1992b) exclusion rules regarding acquiescence, nay-saying, or random responding; 3 for failing to answer the NEO–PI–R validity question regarding honesty and completeness of answers, and 1 for answering "disagree" to the validity question regarding honesty and completeness. Group sizes remaining for analyses were 94 for the SS group and 52 for the SF group.

Data from the PS group (psychiatric outpatients) were archival and thus posttesting feedback questions were not available for these participants. However, as noted earlier, their NEO–PI–R results were screened using rules described in the manual for detecting acquiescence, nay-saying, and random responding.

Results and Discussion

No attempt was made to match the student groups with the patient group on demographic variables. One-way analyses of variance (ANOVAs) on age indicated significant differences, $F(2, 279) = 188.1$, $p < .0001$, with the patient group significantly older than the student groups, which did not differ from each other (PS $M = 37.7$, $SD = 11.5$; SS $M = 19.0$, $SD = 1.5$; SF $M = 19.2$, $SD = 1.2$). A chi-square procedure on sex distribution across the three groups failed to reach significance, $\chi^2(2, N = 282) = .138$, $p > .05$.

Table 1 presents mean scores by group for NEO–PI–R domains. There were statistically significant effects for one-way ANOVAs for all domains (all ps $< .01$). Pairwise follow-up contrasts using Tukey's honestly significant difference (HSD) procedure and $p < .05$ were calculated to assess the origin of the overall differences. Relative to the SS group, the PS group had significantly higher N scores and significantly lower scores on all remaining domains, indicating differences between student and psychiatric patients on these personality domains. Consistent with past research contrasting college students faking bad with college students answering honestly, the SF group had significantly higher scores than the SS group for N and significantly lower scores for E, O, A, and C. Considering the contrasts of feigning students with psychiatric patients (SF and PS), although a similar pattern was present for N, E, and O, a different pattern was present for remaining domains, with A significantly higher for feigning students than for psychiatric patients, and no significant difference between the feigning students and patients for

TABLE 1
Study 1: NEO–PI–R Domain *T* Scores of Student and Patient Groups

	SS		*SF*		*PS*			
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*
Neuroticism	48.4_a	11.0	64.3_b	10.8	59.4_c	5.4	69.0	.0001
Extraversion	51.0_a	10.6	24.3_b	14.0	43.7_c	5.2	140.5	.0001
Openness	52.9_a	10.8	39.6_b	10.8	43.8_c	5.2	49.1	.0001
Agreeableness	52.6_a	9.7	42.1_b	13.5	34.2_c	7.8	99.8	.0001
Conscientiousness	48.8_a	10.4	43.5_b	10.7	41.2_b	6.0	21.5	.0001

Note. Means with the same subscript do not differ significantly (Tukey's honestly significant difference $p < .05$). NEO–PI–R = NEO Personality Inventory–Revised; SS = student standard instructions; SF = student fake-bad instructions; PS = patient standard instructions.

TABLE 2
Study 1: NEO–PI–R Validity Indicator Raw Scores of Student and Patient Groups

	SS		*SF*		*PS*		*d*			
	M	*SD*	*M*	*SD*	*M*	*SD*	*SF Versus SS*	*SF Versus PS*	*F*	*p*
PPM	18.4_a	4.1	15.0_b	4.5	22.0_c	3.0	–0.8	–2.01	73.1	.0001
NPM	9.8_a	3.8	17.6_b	5.1	13.8_c	4.5	1.81	0.81	55.1	.0001
NPM – PPM	-8.6_a	6.5	2.6_b	8.5	-8.2_a	5.5	1.54	1.67	60.8	.0001

Note. Means with the same subscript do not differ significantly (Tukey's honestly significant difference $p < .05$). NEO–PI–R = NEO Personality Inventory–Revised; SS = student standard instructions; SF = student fake bad instructions; PS = patient standard instructions; *d* = Cohen's *d* a popular metric for standardizing group differences (*d* = M1 – M2/Sp [pooled standard deviation]); PPM = Positive Presentation Management; NPM = Negative Presentation Management.

C. Thus, although these results are comparable to past research contrasting feigning and honest students, a somewhat different pattern for domains A and C is found when feigning students are contrasted with psychiatric patients. Taken together with the data from the contrast of SS and PS groups, these results suggest that student and psychiatric control groups may not be entirely comparable on mean NEO–PI–R domain scores.

Table 2 presents results from the PPM and NPM scales for the three groups as well as an experimental index derived by subtracting the raw PPM score from the raw NPM score (NPM – PPM). This score is intended to compare the tendency to present negatively with the tendency to present positively as indexed by the NPM and PPM scales, respectively. Thus, a score of 0 on this calculation indicates that the two tendencies are equivalent, a negative sign for the difference score indicates more of a tendency to present positively (NPM < PPM), and a positive sign for the difference score indicates more of a tendency to present negatively (NPM > PPM). The logic of this index is based on the MMPI and MMPI–2 *F* – *K* index, which contrasts the tendency to fake bad as assessed by the *F* scale with the tendency to fake good as reflected by the *K* scale. The *F* – *K* index has been reported to perform well as an indicator of faking bad in meta-analyses of the detection of malingering on the MMPI (Berry et al., 1991) and the MMPI–2 (Rogers et al., 1994).

One-way ANOVAs calculated on NPM, PPM, and NPM – PPM indicators showed significant effects for all three (all *p*s < .01). Follow-up pairwise comparisons using Tukey's HSD procedure and a $p < .05$ indicated significant differences on PPM for all comparisons. The highest scores were obtained by the PS group, intermediate scores by the SS group, and the lowest scores by the SF group, suggesting that the student fakers suppressed their positive presentation management tendencies. For NPM, there were also significant pairwise differ-

ences for all contrasts, with the highest scores obtained by the SF group, intermediate scores coming from the PS group, and the lowest scores from the SS group. A different pattern was obtained for the NPM – PPM index. Here, there were no significant differences between SS and PS, whereas both of these groups were significantly lower than the SF group. Thus, the NPM – PPM index equated students answering honestly and patients presumed to be answering honestly, whereas both the NPM and PPM indexes showed significant differences between the SS and PS groups.

An alternative way of evaluating these data is to refer to effect sizes (Rosenthal, 1991). Cohen's *d*, a popular metric for standardizing group differences, is calculated as $d = M1 - M2/Sp$, where M1 is the mean of the first group on the scale under consideration and M2 is the mean for the second group. The difference between M1 and M2 is referenced to the pooled standard deviation for the two groups. Thus, a *d* score may be conceptualized as a standard score (e.g., a *z* score), with 0 indicating no effect, 1 indicating that M1 is 1 pooled *SD* unit higher than M2, and –1 indicating that M1 is 1 pooled *SD* unit lower than M2. Cohen's *d*, like other effect size indicators, has the advantage of allowing comparison of differences between groups on different scales, and is commonly used in meta-analytic reviews.

The *d* values in Table 2 use the mean of the SF group as M1 and quantify the differences between the groups on the validity scales using the pooled standard deviation as a metric. Thus, for PPM, the SF group is .8 pooled *SD* units lower than the SS group ($d = –.8$), whereas the SF group has scores approximately 2 pooled *SD* units lower than the PS group ($d = –2.01$). For the NPM scale, the SF group had scores 1.81 pooled *SD* units higher than the SS group, but only .81 pooled *SD* units higher than the PS group. This result, already noted in a different format earlier, in which analog student malingerers have larger effect sizes when compared to student controls than when contrasted with psychiatric controls, is commonly reported in the MMPI and MMPI–2 feigning literature (Berry et al., 1991; Rogers et al., 1994). Finally, the NPM – PPM index generated roughly comparable *d* scores for the SF versus SS and SF versus PS comparisons (1.54 and 1.67, respectively), suggesting that this index performed relatively well in the crucial contrast of analog feigners and psychiatric patients. In fact, the effect size for NPM – PPM is approximately twice as large for the SF versus PS comparison as that found for the NPM scale.

Results from Study 1 indicate significant differences for all three validity indicators when comparing the college students faking bad with both college students and psychiatric outpatients answering under standard instructions. This finding provides support for the research validity scales introduced by Schinka et al. (1997). Additionally, the experimental NPM – PPM validity index, modeled on the MMPI and MMPI–2 *F – K* index, showed promise, particularly in the critical contrast of feigners and psychiatric outpatients. However, it is possible that these

indicators performed well as a result of capitalization on factors specific to the samples in this study, and thus cross-validation in new samples is required.

STUDY 2

Method

Participants. Participants in Study 2 included psychiatric outpatients and college undergraduates. Because of the greater need for data on the crucial comparison of feigners with psychiatric patients, only SF and PS groups were included in Study 2. The psychiatric patients included the 136 outpatients randomly withheld for Study 2 from the total sample of 272 and screened as described earlier. *DSM–IV* diagnoses for this group were also based on structured interviews and included 98 patients with unipolar depression, 19 with bipolar depression, most recent episode depressed, and 19 with schizophrenia. The average age of this sample was 41.0 years (SD = 10.9) and 40.4% were men. No data on educational achievement or race were available.

Undergraduate participants for the Study 2 SF group were drawn from the report of Caldwell-Andrews et al. (2000) and included only the group that completed the NEO–PI–R under fake-bad instructions. This group was initially comprised of 50 Introduction to Psychology students who were participating to fulfill research requirements. Five of these participants were excluded from analyses for various reasons described later. The 45 remaining participants had a mean age of 18.5 (SD = .9) and 31% were men.

Materials. Standard NEO–PI–R question booklets and answer sheets were used for all participants. In addition to the NEO–PI–R, student participants in the fake-bad group completed other personality instruments not germane to this study as detailed in Caldwell-Andrews et al. (2000). Students in the fake-bad group were additionally given a posttest questionnaire on which they were asked to repeat their instructions and respond to two items on a 10-point scale: I understood the instructions given, and I completed the task according to the instructions given.

Procedure. Psychiatric outpatients (PS) had completed the NEO–PI–R under standard instructions as part of their psychological evaluations. They were screened using the procedures described in Study 1. The 50 students in the fake-bad condition (SF) completed the NEO–PI–R as well as other personality instruments under standard instructions on one occasion as described in Caldwell-Andrews et al. (2000). On a second occasion, approximately 1 week following the standard administration, fake-bad participants completed the NEO–PI–R under the same fake-bad instruction set described in Study 1, followed by the posttesting questionnaire.

These participants were told that the four most convincing feigners in their group would receive prizes of $20, $15, $10, and $5. Prizes were mailed to participants following completion of data collection.

Results from the original 50 students in the fake-bad group were screened for ability to reproduce instructions; answers to the posttesting questionnaire; responses to the NEO–PI–R validity question; and the presence of acquiescence, nay-saying, or random responding. This resulted in exclusion of 5 participants, for a final sample of 45 in the student fake-bad group.

Results and Discussion

No attempt was made to match the PS and SF groups on demographic characteristics. An ANOVA on age indicated significant differences, $F(1, 179) = 193.1$, $p < .0001$, with the PS group significantly older ($M = 41.1$, $SD = 10.9$) than the SF group ($M = 18.5$, $SD = .9$). A chi-square procedure on sex distribution failed to reach significance, $\chi^2(1, N = 181) = 1.25$, $p > .05$, with the PS group 40% men and the SF group 31% men.

Table 3 presents mean scores on NEO–PI–R domains for the two groups. There were statistically significant differences between the two groups for all domains (all ps $< .01$), with the SF group scoring higher on N and lower on remaining domains. These results are somewhat different from the comparable comparisons described in Study 1 for domains A and C. Whereas Study 1 found the SF group to be significantly higher on A and equivalent on C relative to the PS group, Study 2 results found both A and C to be significantly lower for the SF group relative to the PS group.

Table 4 presents results for the two groups on the NEO–PI–R validity indicators. ANOVAs for all three were statistically significant (all ps $< .01$), with the SF group lower on PPM and higher on NPM and NPM – PPM. These differences are consistent with those reported for Study 1's contrasts of the SF and

TABLE 3
Study 2: NEO–PI–R Domain *T* Scores of Student Feigners and Patients Under Standard Instructions

	SF		*PS*			
	M	*SD*	*M*	*SD*	*F*	*p*
Neuroticism	66.7	13.4	60.0	5.4	30.0	.0001
Extraversion	22.0	16.8	43.3	5.6	179.7	.0001
Openness	36.9	12.1	43.4	5.6	28.3	.0001
Agreeableness	29.9	14.9	34.8	7.2	8.9	.0033
Conscientiousness	35.8	17.8	42.5	6.2	20.6	.0001

Note. NEO–PI–R = NEO Personality Inventory–Revised; SF = student fake-bad instructions; PS = patient standard instructions.

TABLE 4
Study 2: NEO–PI–R Validity Indicator Raw Scores of Student Feigners and Patients Under Standard Instructions

	SF		*PS*				
	M	*SD*	*M*	*SD*	*d*	*F*	*p*
PPM	13.9	6.3	21.4	3.2	–1.79	107.1	.0001
NPM	21.9	7.6	12.9	4.1	1.74	101.8	.0001
NPM – PPM	8.0	12.0	–8.5	4.8	2.27	172.7	.0001

Note. NEO–PI–R = NEO Personality Inventory–Revised; SF = student fake-bad instructions; PS = patient standard instructions; *d* = Cohen's *d* a popular metric for standardizing group differences (*d* = M1 – M2/Sp); PPM = Positive Presentation Management; NPM = Negative Presentation Management.

PS groups. The *d* score for NPM is considerably higher than that found in Study 1, as it is for NPM – PPM as well. These results again support the effectiveness of the two indicators for detecting a faking bad approach to the test at the group level, even when psychiatric patients experiencing significant psychopathology are used as a control group.

Although the data reported in Study 1 and Study 2 are useful for assessing group tendencies, they do not directly address the classification of individual test takers, which is the task faced by clinicians. The accuracy of diagnostic tests is generally described using several key parameters: sensitivity, specificity, negative predictive power (NPP), and positive predictive power (PPP; Glaros & Kline, 1988). *Sensitivity* is defined as the percentage of a group known to have a target condition who are correctly classified by a given cutting score on a predictor (e.g., test or scale) of the condition. *Specificity* is defined as the percentage of a group known not to have a target condition who are correctly classified by a given cutting score on a predictor of the condition. As clinicians do not usually administer scales for detecting a condition they already know is present (or absent), sensitivity and specificity are not directly applicable to clinical decision making.

In contrast to sensitivity and specificity, PPP and NPP are clinically applicable. *PPP* is the likelihood that an individual with a test result falling at or above a cutting score (e.g., predicted by the test to have the condition) actually has the condition. *NPP* is the likelihood that an individual with a test score falling below the cutting score (e.g., predicted by the test not to have the condition) actually does not have the condition. Sensitivity and specificity are assumed to be relatively stable across base rates and similar samples, whereas PPP and NPP are affected by the base rates of the condition. As noted earlier, when the actual base rates of a condition are very low, even a highly accurate test will rarely surpass a strategy that involves always predicting the absence of the condition. Thus, if a condition occurs in only 5% of the clients in a particular setting, simply predicting that no

client has the condition will result in 95% correct classifications, a level very difficult to exceed even using a scale possessing excellent sensitivity and specificity characteristics, a point with significant clinical implications. Using a test or scale to predict the presence or absence of a condition having a very low base rate will usually increase error rates relative to predictions relying on the prevailing base rate. Thus, it is extremely challenging, from a psychometric perspective, to identify a very low base rate condition with any incremental accuracy beyond that afforded by knowledge of the base rates.

Sensitivity and specificity parameters associated with use of a given cutting score tend to be more stable if larger numbers of individuals both with and without the condition in question are available. Thus, to calculate sensitivity rates at various cutting scores, results from the student feigners in Studies 1 and 2 were collapsed for one faking group numbering 97. Similarly, to calculate specificity rates, psychiatric outpatients from the two studies were collapsed for one patient group numbering 272 (students answering under standard instructions are not included in this group because this comparison is not likely to be clinically relevant). For calculating sensitivity and specificity, the condition in question is the presence of a fake-bad approach to the test. Thus, sensitivity is calculated on the student faking group, whereas specificity is determined in the patient group.

Table 5 presents data on sensitivity and specificity characteristics for the NPM scale when a range of cutting scores is used to predict whether a particular NEO–PI–R protocol has been answered under standard or fake-bad instructions. Caldwell-Andrews et al. (2000) used a cutting score of ≥ 16 for discriminating students faking bad from students answering honestly, with a resulting specificity of 90% and sensitivity of 81%. In this study, use of an NPM cutting score of ≥ 16 for predicting a faking bad approach to the test results in a specificity rate of 71.7 and a sensitivity rate of 73.2, a decline from the accuracy reported by Caldwell-Andrews et al., who used groups composed exclusively of students. This failure to cross-validate cutting scores derived entirely from students when including psychiatric patient groups is a common finding in the MMPI and MMPI–2 faking literature (Berry et al., 1991; Rogers et al., 1994). To approximate a specificity rate of .90 on the NPM scale for the sample reported here (the specificity rate associated with a cutting score of ≥ 16 in Caldwell-Andrews et al.'s samples), a cutting score of ≥ 20 is necessary, which has a specificity rate of .919 and a sensitivity rate of .485. A more conservative specificity rate of .95 is approximated at a cutting score of ≥ 21, which has a specificity rate of .945 and a sensitivity rate of .412.

Table 6 presents similar data for the NPM – PPM index. Because this index has not previously been studied, no a priori cutting score is available. One approach to determining a cutting score is to identify the value resulting in the highest overall classification rate (optimal cutting score). However, this strategy weights false positive and false negative errors equivalently. In light of the serious consequences for a client who is suspected of faking bad based on test results (erroneously

TABLE 5
Sensitivity and Specificity of NPM Scale for Detecting Fake-Bad Response Set in Combined Samples of Student Feigners and Psychiatric Outpatients

	NPM	
Cutting Score ≥	*Specificity*[a]	*Sensitivity*[b]
0	0.0	100.0
1	0.0	100.0
2	0.4	100.0
3	0.7	100.0
4	0.7	100.0
5	1.8	99.0
6	3.7	99.0
7	5.9	96.9
8	9.6	96.9
9	11.0	96.9
10	17.6	93.8
11	23.9	90.7
12	31.3	87.6
13	43.0	86.6
14	53.7	81.4
15	62.1	77.3
16	71.7	73.2
17	79.4	68.0
18	82.4	62.9
19	88.2	56.7
20	91.9	48.5
21	94.5	41.2
22	97.1	37.1
23	97.8	32.0
24	98.5	32.0
25	98.9	20.6
26	99.3	19.6
27	100.0	13.4
28	100.0	11.3
29	100.0	9.3
30	100.0	7.2
31	100.0	6.2
32	100.0	5.2
33	100.0	2.1
34	100.0	2.1
35	100.0	2.1
36	100.0	2.1
37	100.0	1.0
38	100.0	1.0
39	100.0	0.0

Note. Cutting scores are raw. NPM = Negative Presentation Management; Specificity = % of those in patient standard instruction group correctly classified using this cutting score; Sensitivity = % of those in student faking instruction group correctly classified using this cutting score.

[a]$n = 272$. [b]$n = 97$.

TABLE 6
Sensitivity and Specificity of NPM – PPM Index for Detecting Fake-Bad Response Set in Combined Samples of Student Feigners and Psychiatric Outpatients

	NPM – PPM	
Cutting Score ≥	*Specificity*[a]	*Sensitivity*[b]
–26	0.0	100.0
–25	0.0	100.0
–24	0.0	99.0
–23	0.0	99.0
–22	0.0	99.0
–21	0.0	99.0
–20	1.5	99.0
–19	3.7	99.0
–18	4.4	99.0
–17	5.1	99.0
–16	6.6	97.9
–15	11.1	96.9
–14	13.2	95.9
–13	15.4	95.9
–12	16.9	95.9
–11	22.8	95.9
–10	30.1	93.8
–9	36.0	92.8
–8	44.1	89.7
–7	55.1	88.7
–6	61.0	88.7
–5	72.1	86.6
–4	78.7	85.6
–3	84.6	83.5
–2	91.2	79.4
–1	97.8	76.3
0	99.3	71.1
1	99.3	63.9
2	99.3	59.8
3	99.3	57.7
4	100.0	54.6
5	100.0	53.6
6	100.0	49.5
7	100.0	44.3
8	100.0	39.2
9	100.0	30.9
10	100.0	28.9
11	100.0	25.8
12	100.0	25.8
13	100.0	24.7
14	100.0	20.6

(continued)

TABLE 6 (Continued)

	NPM – PPM	
Cutting Score ≥	*Specificity*[a]	*Sensitivity*[b]
15	100.0	18.6
16	100.0	16.5
17	100.0	14.4
18	100.0	12.4
19	100.0	11.3
20	100.0	7.2
21	100.0	5.2
22	100.0	5.2
23	100.0	3.1
24	100.0	3.1
25	100.0	3.1
26	100.0	3.1
27	100.0	3.1
28	100.0	3.1
29	100.0	3.1
30	100.0	3.1
31	100.0	2.1
32	100.0	2.1
33	100.0	1.0
34	100.0	1.0
35	100.0	1.0
36	100.0	1.0
37	100.0	0.0

Note. Cutting scores are raw. NPM = Negative Presentation Management; PPM = Positive Presentation Management; Specificity = % of those in patient standard instruction group correctly classified using this cutting score; Sensitivity = % of those in student faking instruction group correctly classified using this cutting score.

[a]$n = 272$. [b]$n = 97$.

predicting that an honest test taker is faking bad), it seems appropriate in this case to attempt to minimize false positive errors even at the expense of an increase in false negative errors. (However, it is worth noting that this reasoning is not a consensus in the field at this time, and individual clinicians and researchers must make their own decisions regarding this issue.) Explicitly adopting a strategy of reducing false positive errors for this data set, the sensitivity rates associated with cutting scores having a specificity of .90 and .95 will be considered. The specificity rate for NPM – PPM in Table 6 closest to .90 occurs at a cutting score of –2, which has a specificity rate of .912 and a sensitivity rate of .794. The specificity rate closest to .95 occurs at a cutting score of ≥ –1, which has a specificity rate of .978 and a sensitivity rate of .763. Of course, these values must be cross-validated in new samples before being accepted as stable estimates of these parameters, and

should not be used in clinical practice before such an evaluation has been conducted. PPP and NPP parameters for these scales are discussed later.

GENERAL DISCUSSION

The major finding of these studies was that NEO–PI–R validity scales were significantly different, in expected directions, for groups of analog feigners and psychiatric outpatients. Student participants who feigned psychological disturbance scored significantly lower on the PPM scale, which is intended to identify a faking good approach to the test, and significantly higher on both NPM and the experimental NPM – PPM index, which are intended to detect a faking bad approach to the test. This pattern of results contributes to the construct validity of the Schinka et al. (1997) research validity scales as well as the new NPM – PPM index, suggesting that they are in fact sensitive to a faking bad response set. The effect sizes for NPM and NPM – PPM for the critical comparisons of analog feigners and psychiatric outpatients were in the large to very large category, at .81 and 1.74 for NPM and 1.67 and 2.27 for NPM – PPM. Although these values are not quite as high as those typically reported for similar MMPI and MMPI–2 indicators (Berry et al., 1991; Rogers et al., 1994), they are quite respectable. Overall, these results suggest that NPM and NPM – PPM are sensitive to the presence of feigning at a group level and support further research work with these scales.

At the level of individual classifications, a previously proposed cutting score for use with NPM that was derived exclusively from student participants was not supported by these results. However, raising the cutting score on NPM to ≥ 20 or 21 resulted in acceptable specificity rates and moderate sensitivity rates. Relative to NPM, the NPM – PPM index produced higher sensitivity rates at comparable specificity levels. For example, adopting a .919 specificity rate on NPM (cutting score of ≥ 20) resulted in a sensitivity rate of .485, whereas a similar specificity rate on NPM – PPM (cutting score of ≥ –2, specificity = .912) resulted in a sensitivity rate of .794. At a .945 specificity rate (cutting score of > 21), NPM's sensitivity is .412, whereas NPM – PPM's associated sensitivity rate is .763 (cutting score of ≥ –1, specificity rate = .978). These results raise the possibility that NPM – PPM may have a modest superiority over NPM alone for the detection of a faking bad approach to the NEO–PI–R, although this finding requires cross-validation in new samples.

Although the importance of positive and negative predictive powers has been noted here, an explicit illustration of the effect of base rates on these parameters may emphasize the point more clearly. Table 7 presents classification parameters for the NPM and NPM – PPM indicators at several different base rates drawn from experts' subjective estimates reported by Rogers (1997) of 7.6% and 16.5% feigning in nonforensic and forensic settings respectively, and Frederick et al.

TABLE 7
Classification Statistics for NPM and NPM – PPM Indicators for the Identification of Feigning at Various Base Rates

Indicator	*Cutting Score*	*Base Rate*	*PPP*	*NPP*	*Specificity*	*Sensitivity*
NPM	≥ 20	44.0	.82	.69	.92	.49
		26.0	.68	.84	.92	.49
		16.5	.54	.90	.92	.49
		7.6	.33	.96	.92	.49
	≥ 21	44.0	.85	.67	.95	.41
		26.0	.72	.82	.95	.41
		16.5	.60	.89	.95	.41
		7.6	.38	.95	.95	.41
NPM – PPM	≥ –2	44.0	.88	.85	.91	.79
		26.0	.76	.93	.91	.79
		16.5	.64	.95	.91	.79
		7.6	.42	.98	.91	.79
	≥ –1	44.0	.97	.84	.98	.76
		26.0	.93	.92	.98	.76
		16.5	.88	.95	.98	.76
		7.6	.74	.98	.98	.76

Note. NPM = Negative Presentation Management; PPM = Positive Presentation Management; cutting score = score at or above which condition is predicted to be present; base rate = prevalence of condition (faking bad) in combined samples; PPP = positive predictive power; NPP = negative predictive power; Specificity = % of those in patient standard instruction group correctly classified using this cutting score; Sensitivity = % of those in student faking instruction group correctly classified using this cutting score.

(2000), who found a prevalence of about 44% invalid responding based on an objective indicator administered to several hundred federal prisoners referred for mental status and competency questions. The base rate of 26% is for the samples used here (97 faking bad, 272 standard instructions). Data in this table emphasize that, although specificity and sensitivity remain stable across different base rates, as the base rate of faking bad falls, PPP declines, in some cases precipitously, whereas NPP increases. Practically speaking, for NPM, attempting to identify faking bad in a setting where the condition has a very low to low base rate (7.6% or 16.5%) would result in poor PPP. For example, using a cutting score of ≥ 21 for NPM in a setting where faking bad has a base rate of 16.5%, there are only 6 chances in 10 that an individual with a positive test sign (at or above the cutting score on NPM) is actually feigning, a probability only marginally better than flipping a coin. The situation for NPM – PPM is slightly better, although PPP is still relatively modest in the very low base rate (7.6%) environment. However, as the base rate of feigning increases to 16.5%, a cutting score of ≥ –1 on NPM – PPM has respectable PPP and NPP of .88 and .95. Data from this table emphasize the

importance of considering base rates as well as PPP and NPP in determining where to set a cutting score as well as when using a test sign may increase or decrease predictive accuracy. These factors must be carefully considered before validity scales are applied in specific settings.

The authors of the NEO–PI–R have indicated that they do not feel that validity scales are needed for the test (Costa & McCrae, 1992a), and data have been presented that suggest that such scales are of negligible value in research studies of volunteer participants given the test (Piedmont, McCrae, Riemann, & Angleitner, 2000). The data on PPP and NPP presented in this article highlight the fact that, in samples that have very low base rates of response sets, validity scales would likely serve to increase error in prediction. Thus, validity scales may in fact be contraindicated in research using volunteer participants, or settings with similarly very low base rates of response sets, as suggested by Piedmont et al. (2000). However, in clinical settings where the base rate of negative impression management is higher (Frederick et al., 2000; Strong et al., 2000), or employment settings where the base rate of positive impression management is moderate to high (Rosse, Stecher, Miller, & Levin, 1998), validity scales have the potential to increase the accuracy of predictions, although further research is necessary to demonstrate that this promise is in fact fulfilled with the research validity scales for the NEO–PI–R.

There are a number of weaknesses in this study that suggest that specific validity scale results should be regarded cautiously until replicated. The use of college students in the analog feigning groups, although generally accepted in the initial phase of research with a feigning indicator, limits the generalizability of the findings, as potentially does the lack of matching on demographic characteristics for the student and inpatient groups. The level of monetary incentives offered to the feigning groups is obviously much less than available in the real world, and may not have motivated these analog feigners to apply the same level of effort as those with important contingencies hinging on the outcomes of their evaluations. It is also important to note that the psychiatric patients in these samples were outpatients, and the findings may not generalize to more seriously disturbed populations. Additionally, the vast majority of these outpatients were diagnosed with unipolar depression, with bipolar depression and schizophrenia diagnosed much less commonly. Thus, these results may not apply to populations with other disorders. The sensitivity and specificity characteristics of NPM and NPM – PPM at various cutting scores require cross-validation before they may be applied to clinical decision making, and these results are not recommended for clinical use, although research exploring the replicability of the findings in large samples would be helpful. The new feigning index described here, NPM – PPM, also requires evaluation in new samples to determine if its moderate superiority relative to NPM is maintained in other groups. Further research is needed to address these and other concerns regarding the potential for, and clinical utility of, detecting response sets on the NEO–PI–R.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bagby, R. M., Bindseill, K. D., Schuller, D. R., Rector, N. A., Young, L. T., Cooke, R. G., Seeman, M. V., McCay, E. A., & Joffe, R. J. (1997). The relationship between the five factor model of personality and unipolar, bipolar, and schizophrenia. *Psychiatry Research, 70,* 83–94.

Bagby, R. M., Costa, P. T., McCrae, R. R., Livesley, W. J., Kennedy, S. H., Levitan, R. D., Joffe, R. T., Levitt, A. J., Young, L. T., & Seeman, M. V. (1999). Replicating the five factor model of personality in a psychiatric sample. *Personality and Individual Differences, 27,* 1135–1139.

Ben-Porath, Y. S., & Waller, N. (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., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–598.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

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

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

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

Costa, P. T., & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin, 117,* 216–220.

Frederick, R. I. (1997). *Validity Index Profile manual.* Minneapolis, MN: National Comupter Systems.

Frederick, R. I., Crosby, R. D., & Wynkoop, T. F. (2000). Performance curve classification of invalid responding on the Validity Indicator Profile. *Archives of Clinical Neuropsychology, 15,* 281–300.

Furnham, A. F. (1997). Knowing and faking one's five-factor personality score. *Journal of Personality Assessment, 69,* 229–243.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity and predictive value model. *Journal of Clinical Psychology, 44,* 1013–1023.

Goldberg, L. R., & Saucier, G. (1995). So what do you propose that we use instead? A reply to Block. *Psychological Bulletin, 117,* 221–229.

Meehl, P. J., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Paulhus, D. L., Bruce, M. M., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin, 21,* 100–108.

Piedmont, R. L. (1998). *The Revised NEO Personality Inventory: Clinical and research applications.* New York: Plenum.

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.

Rogers, R. (1997). Current status of clinical methods. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 373–397). New York: Guilford.

Rogers, R., Sewell, K., & Salekin, R. (1994). A meta-analysis of feigning on the MMPI–2. *Assessment, 2,* 81–89.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

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., 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.

Scandell, D. J., & Wlazelek, B. G. (1996). Self-presentation strategies on the NEO–FFI: Implications for detecting faking. *Psychological Reports, 79,* 1115–1121.

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

Strong, D. R., Greene, R. L., & Schinka, J. A. (2000). A taxometric analysis of MMPI–2 infrequency scales (*F* & *Fp*) in clinical settings. *Psychological Assessment, 12,* 166–173.

Widiger, T. A., & Trull, T. J. (1997). Assessment of the five-factor model of personality. *Journal of Personality Assessment, 68,* 228–250.

Yang, J., Bagby, R. M., & Ryder, A. G. (2000). Response styles and the revised NEO Personality Inventory: Validity scales and spousal ratings in a Chinese psychiatric sample. *Assessment, 1,* 389–402.

David T. R. Berry
Department of Psychology
Kastle Hall
University of Kentucky
Lexington, KY 40506–0044
E-mail: dtrb@pop.uky.edu

Received July 27, 2000
Revised September 26, 2000