Journal of Personality and Social Psychology
2000, Vol. 78, No. 3, 582–593

0022-3514/00/$5.00 DOI: 10.1037//0022-3514.78.3.582

On the Invalidity of Validity Scales: Evidence From Self-Reports and Observer Ratings in Volunteer Samples

Ralph L. Piedmont
Loyola College

Robert R. McCrae
National Institute on Aging

Rainer Riemann and Alois Angleitner
University of Bielefeld

Because of the potential for bias and error in questionnaire responding, many personality inventories include validity scales intended to correct biased scores or identify invalid protocols. The authors evaluated the utility of several types of validity scales in a volunteer sample of 72 men and 106 women who completed the Revised NEO Personality Inventory (NEO-PI-R; P. T. Costa & R. R. McCrae, 1992) and the Multidimensional Personality Questionnaire (MPQ; A. Tellegen, 1978/1982) and were rated by 2 acquaintances on the observer form of the NEO-PI-R. Analyses indicated that the validity indexes lacked utility in this sample. A partial replication (N = 1,728) also failed to find consistent support for the use of validity scales. The authors illustrate the use of informant ratings in assessing protocol validity and argue that psychological assessors should limit their use of validity scales and seek instead to improve the quality of personality assessments.

Most personality assessment relies on the use of questionnaires. Respondents are typically presented with a series of items and asked to indicate whether or to what extent each item accurately describes their personality. Questionnaires are also used to gather informant ratings of personality traits. In both cases, the method assumes that respondents understand the procedure and are both willing and able to provide reasonably accurate responses. Usually respondents are: The accumulated construct validity of all personality questionnaires is compelling evidence of the soundness of this method of measurement.

But it has been clear for decades that questionnaire responses are also susceptible to a variety of distortions. Respondents may misunderstand the instructions or misread the questions. They may be uncooperative and respond at random, or without careful thought. They may not have insight into their own personality traits, or they may answer defensively. They may make deliberate attempts to present a favorable or unfavorable picture of themselves. They may agree or disagree with items indiscriminately, or they may overuse or avoid extreme responses. If such distortions are severe and common, the scale will fail the usual tests of convergent and discriminant validity. But even if a scale works well in general, the scores of some respondents may be severely distorted, yielding inaccurate and misleading personality profiles.

The validity of individual test protocols is a great and legitimate concern of all personality assessors, particularly clinicians, and psychometricians have devised a variety of *validity scales* intended to assess the accuracy of self-reports (e.g., Arbisi & Ben-Porath, 1995). Almost all these scales examine a pattern of responses and attempt to infer the credibility of the test as a whole. Scores on validity scales are typically used either to identify suspicious protocols that may be discarded or to adjust scores on other, substantive scales. Infrequency (e.g., Jackson, 1984) and social desirability (e.g., Edwards, 1957) scales are common examples.

Validity scales are included in most personality assessment inventories, and researchers and clinicians rely heavily on them. In their 1996 *Annual Review* chapter, Butcher and Rouse stated flatly that "personality assessment instruments, in order to be effective, must have validity scales that can appraise the subject's level of cooperation, willingness to share personal information, and degree of response exaggeration" (Butcher & Rouse, 1996, p. 94). It would doubtless be desirable to have such information, but there is a growing suspicion that current validity scales do not provide it (Borkenau & Ostendorf, 1992; Nicholson & Hogan, 1990; Smith, 1997).

There are dozens of studies on the validity of validity scales (e.g., Schinka, Kinder, & Kremer, 1997; Stein, Graham, & Williams, 1995), but most of them do not speak directly to the utility of such scales in practice. The typical study uses a faking paradigm, in which participants (or computers) are asked to simulate some form of distortion (say, fake good). Such studies often show that the validity scale does distinguish between faking and control

Ralph L. Piedmont, Department of Pastoral Counseling, Loyola College; Robert R. McCrae, Personality, Stress, and Coping Section, Gerontology Research Center, National Institute on Aging, National Institutes of Health; Rainer Riemann and Alois Angleitner, Department of Psychology, University of Bielefeld, Bielefeld, Germany.

We thank Auke Tellegen, Joseph Ciarrocchi, Paul T. Costa, Jr., Lewis R. Goldberg, Jeffrey Sanders, William Sneck, and Bill Chaplin for their helpful comments on a draft of this article.

Correspondence concerning this article should be addressed to Ralph L. Piedmont, Department of Pastoral Counseling, Loyola College, 7135 Minstrel Way, Columbia, Maryland 21045. Electronic mail may be sent to piedmont@vax.loyola.edu.

conditions, from which the authors conclude that it is able to detect the bias in question. What these studies normally do not consider is the extent to which high scores on a validity scale may be obtained when the protocol is actually valid. This is a very serious problem: Lim and Butcher (1996), for example, showed that a cutoff score that discriminated faking bad from honest student respondents with 100% accuracy identified fully 30% of a sample of presumably honest psychiatric patients as faking bad. How serious the problem of false positives is depends crucially on the base rate of invalid responding in the sample. If "deliberate efforts to distort responses are rare in clinical applications of psychological tests" (Jackson, 1989, p. 22), then false positives may far outnumber true positives, and using the validity scale may prove counterproductive. This is even more likely to be true for research on normal volunteers.

The faking study is popular because in that design it is possible to create a group of known true positives through experimental manipulation. By contrast, in naturalistic conditions researchers do not ordinarily know whether a protocol is valid or not (that, of course, is why they want validity scales). The accuracy must be gauged by some external criterion. In some cases this may be a separate instrument; often it will need to be a personality assessment obtained from an independent source (e.g., a peer or clinician rating). Given this criteria information, the utility of validity scales can be assessed: Do they accurately and selectively identify protocols that are invalid as judged by an independent criterion?

A series of studies using this paradigm have almost uniformly failed to find support for the utility of validity scales (Alperin, Archer, & Coates, 1996; Borkenau & Ostendorf, 1992; Costa & McCrae, 1997; Dicken, 1963; Goldberg, Rorer, & Greene, 1970; McCrae, 1986; McCrae & Costa, 1983; McCrae et al., 1989). Hough, Eaton, Dunnette, Kamp, and McCloy (1990) used performance ratings as external criteria in a large-scale study predicting military performance; they found some utility for an infrequency scale, but none for a social desirability scale. Diener, Sandvik, Pavot, and Gallagher (1991) evaluated several validity scales in relation to subjective well-being and concluded that social desirability scales assess a substantive aspect of personality rather than a source of error variance. Smith (1997) used objective criteria such as weight, grade point average, and IQ to evaluate the utility of social desirability scales, and found "little support for the hypotheses that SD acts as a moderator, suppressor, or third variable in most psychological research" (p. iii).

The purpose of this report is to review the concept of response distortions, to evaluate empirically whether any of several validity scales is useful when used in samples of normal volunteers, and to discuss alternatives to the use of validity scales. Two studies are presented that evaluate these relationships in two completely different samples: one using an American sample of college students; the other, a large German sample of twins. For some readers, the article may seem to "beat a dead horse." Many researchers have already come to recognize the limitations of validity scales, especially when used in volunteer samples (cf. Paulhus, 1991). Auke Tellegen (personal communication, August 4, 1996) now recommends using only two validity scales out of six initially developed for the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1978/1982), and the authors of the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) acknowledge that the *K* scale may in some cases not be a valid measure of defensiveness. But many other researchers continue to rely on validity scales. MMPI-2 research routinely reports *K*-corrected scores (e.g., Putnam, Kurtz, & Houts, 1996), and measures of social desirability continue to appear as control variables in major personality journals (e.g., Gross & John, 1997).[1]

We examine two personality instruments: the MPQ and the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). Because they are still in use by some researchers (e.g., Church & Burke, 1994), and because they are excellent exemplars of several different types of validity scales, we consider all six MPQ validity scales, as well as two additional scales based on them.

In contrast to the MPQ, the NEO-PI-R expressly omits the usual validity scales, because the test authors believed that there is little empirical justification for their use. Normative data on acquiescence or nay-saying and repetitive responding is provided, but it is considered a guide to caution in interpretation rather than a basis for discarding a protocol. Other researchers, however, have proposed more conventional validity scales for the NEO-PI-R, and their utility is examined here.

The NEO-PI-R has both self-report and observer rating forms. Although most validity scales were designed for use with self-reports, the same or similar distortions (e.g., acquiescence, halo effects, carelessness) may be found with observer ratings. In this article we evaluate scales that claim to identify response bias and error in both self-reports and observer ratings.

The MPQ and NEO-PI-R substantive scales can be used as reciprocal criteria in evaluating the validity scales from each instrument. To select appropriate criteria for the MPQ scales, we consulted correlations reported by Tellegen and Waller (in press). These suggested 13 significant correlations (e.g., MPQ Stress Reaction with NEO-PI-R Neuroticism). To evaluate the utility of NEO-PI-R validity scales, we reversed the roles of predictor and criterion'for example, we used MPQ Well-Being as a criterion for NEO-PI-R Extraversion. Corresponding NEO-PI-R domain scores from different observers were also used as criteria (e.g., self-reported Openness was a criterion for peer-rated Openness).

In the first sets of analyses we treated all validity scales as continuous variables. Effects such as social desirability, acquiescence, and random responding may vary in degree, and the more they are present, the more they should reduce the validity of the substantive scale score. This is the reasoning used in applying the MMPI-2 *K* correction and in such standard procedures as partialing out social desirability scores. In subsequent analyses we used extreme scores on validity scales to identify potentially aberrant protocols, a method commonly used to identify invalid protocols in clinical practice.

[1] We reviewed all articles published in the *Journal of Personality and Social Psychology: Personality Processes and Individual Differences* (JPSP), *Personality and Individual Differences* (PAID), and *Journal of Personality Assessment* (JPA) for 1998 to document use of validity scales. For JPSP, 7% (n = 9) of the 125 articles published involved some type of response bias analysis or correction. In PAID, 16% (n = 32) of the 196 articles included validity scales; and of the 60 regular articles in JPA, 43% (n = 26) did so. It appears that the use of validity scales is widespread.

Systematic Bias in Personality Questionnaires and Detecting It

Self-reports or observer ratings can be inaccurate because the respondent systematically distorts his or her responses. This may happen unconsciously, as in the case of defensiveness, or it may result from a deliberate attempt to create a falsely favorable (or unfavorable) impression. Since Edwards“ (1957) influential volume on social desirability, validity scales intended to measure that response set have proliferated. In the MPQ, two approaches are used. The first scale, Unlikely Virtues (VIRTUES), consists of items that assert a highly desirable but improbable quality. High scorers on this scale are thought to be describing themselves in a falsely favorable way. This scale resembles the Marlowe"Crowne Social Desirability scale (Crowne & Marlowe, 1960) in rationale.

A second scale'the Desirable Response Inconsistency, or DRIN' uses pairs of items that are opposite in content but both highly desirable. Answering *true* to both would suggest that the respondent is attending to desirability rather than item content, and the total number of *true* responses is taken as an indication of socially desirable responding. Very low scores would suggest faking bad. Two ways of scoring DRIN are examined here. Because faking good is presumably more common than faking bad, the simple score on DRIN can be considered an index of invalidity. If, however, a significant number of respondents faked bad, DRIN would show a curvilinear relation to validity. A second index, DRIN2, is therefore calculated as the deviation in either direction from the neutral point of the scale.

Schinka et al. (1997) proposed two scales composed of NEO-PI-R items that are intended to reflect Positive Presentation Management (PPM) and Negative Presentation Management (NPM). Like Tellegen"s VIRTUES scale, PPM is intended to identify respondents who falsely claim uncommon virtues; NPM is supposed to identify those who claim uncommon faults and thus appear to be faking bad.

To the extent that personality scale scores are contaminated by systematic bias, they will be invalid. If, however, the bias can be measured, it can be statistically controlled. In order to assess systematic bias, one must use a criterion that is independent of the rating. One self-report could not be used as the criterion for evaluating another, because a shared bias like social desirability would tend to inflate the correlation between them even when they were invalid. However, a self-report is unlikely to share a systematic bias with an observer rating, so the latter is an appropriate criterion. Bias scales must be evaluated with hetero-method criteria. In this study we use personality ratings from a different respondent as the independent criterion (cf. McCrae & Costa, 1983).[2]

The customary statistical approach to the examination of systematic bias is the suppressor analysis (Cohen & Cohen, 1975). Although suppressor effects were first discussed in the context of multiple regression, a more intuitive approach uses semipartial correlations. To test suppressor effects, the simple correlation of the predictor with the criterion is compared with the semipartial correlation controlling for the suppressor variance in the predictor. If the validity scale successfully measures systematic error variance, removing it should increase the correlation of the predictor with the criterion. Failure to find such an improvement in prediction would argue against the utility of the validity scale as an index of bias. The suppressor analyses are outlined in Table 1.

Table 1
Outline of Planned Suppressor Analyses of the Bias Scales of the NEO-PI-R and MPQ

Predictor	Suppressor	Criterion	Analysis
MPQ self	VIRTUES, DRIN	NEO mean peer	13 × 2 = 26
NEO self	PPM, NPM	NEO mean peer	5 × 2 = 10
NEO first peer	PPM, NPM	MPQ self	13 × 2 = 26
	PPM, NPM	NEO self	5 × 2 = 10
	PPM, NPM	NEO second peer	5 × 2 = 10
Total			82

Note. NEO-PI-R = Revised NEO Personality Inventory; MPQ = Multidimensional Personality Questionnaire; VIRTUES = Unlikely Virtues; DRIN = Desirable Response Inconsistency; PPM = Positive Presentation Management; NPM = Negative Presentation Management.

Random Error and Detecting It

Random error is any nonsystematic source of variance that reduces the validity of a scale score. Random responding' deliberately marking answers without regard to the item content'is the most obvious source of random error, but carelessness or misunderstanding items can also contribute. If a scale has balanced keying, acquiescent responding contributes a kind of random error, because it is not confounded with substantive scores. Several approaches to detecting random error have been proposed.

Tellegen (1988) outlined how inconsistency scales are developed and applied. At the heart of this approach is the assumption that individuals should respond consistently to items having similar content. The True Response Inconsistency (TRIN) scale is made up of item pairs that are *opposite* in content, and inconsistency is scored when an individual answers *true* to both. High scores on TRIN indicate a tendency to answer *true* indiscriminately (an acquiescence effect), whereas low scores suggest nay-saying. As with DRIN, we used two scorings of TRIN. Because acquiescence is probably more common than nay-saying, the simple count of *true* responses can be used as a measure of invalidity. If, however, there are a significant number of individuals with a nay-saying response set, then both high and low TRIN scores will tend to be invalid. We therefore calculated a variable labeled TRIN2 as the absolute value of the difference between TRIN and the neutral point of the scale.

A second MPQ validity scale is the Variable Response Inconsistency (VRIN) scale. VRIN consists of *pairs* of items that are very similar in content, and it is scored configurationally. Responding *true* to one item in a pair and *false* to the other is conceptually inconsistent, and high scores on the VRIN scale

[2] It might be argued that peer ratings are not fully independent criteria, because targets might present themselves falsely in daily life and deceive their peers about their true personality. If that self-presentation style were accurately assessed by a validity scale, controlling for it would reduce self–other agreement but would increase the validity of the assessment. This article presumes that such systematic and long-term deception is rare and that peer ratings—especially from multiple raters—are essentially independent of self-reports.

suggest an individual who is responding to items in an indiscriminate manner.

Researchers have proposed similar validity scales for the NEO-PI-R. Schinka et al. (1997) developed an Inconsistency (INC) scale resembling Tellegen“s VRIN. Ten pairs of items were selected that were strongly positively correlated. The sum of the absolute values of the differences of responses to each pair is interpreted as evidence of inconsistency. L. R. Goldberg (personal communication, January 25, 1995; cf. Goldberg & Kilkowski, 1985) created two other measures of inconsistency. SYNCORR is defined as the correlation for each individual across a set of item pairs with similar (synonymous) content; ANTCORR is the correlation across item pairs with opposite (antonymous) content. Semantically consistent responding should lead to high values of SYNCORR and low values of ANTCORR.

A different approach was used to create the MPQ Associative Slips (SLIPS) scale. It consists of a separate set of items that require respondents to identify which of two response words is least (or most) similar to a stimulus word. The correct answers are obvious, and an error is therefore interpreted as evidence of carelessness or noncooperation. SLIPS functions like an infrequency scale, and many errors would suggest an invalid protocol. Finally, the MPQ also contains a global Invalidity scale (INVAL) that combines scores from other scales, and so it might be particularly sensitive to random error.

Moderated Regression Analyses

Suppressor analyses are not appropriate for an examination of random error, because indexes of the extent of random error are not systematically related to content scale scores. Measures of random responding thus cannot be used to correct content scale scores; they are useful only for identifying invalid protocols. A different statistical approach, moderated regression analysis, can be used to examine the utility of validity scales designed to detect random responding.

Whether the distortion is produced by systematic bias or random error, invalid protocols should show lower correlations with external criteria than valid protocols. The most straightforward way to examine the utility of any validity scale is thus to divide protocols into more and less valid groups on the basis of the validity scale and compare correlations with criteria. Although conceptually appealing, this approach is statistically inefficient, because it fails to make use of all the information in continuously distributed variables. Moderated regression is usually recommended as a superior technique (Paunonen & Jackson, 1985; Tellegen, 1988; Tellegen, Kamp, & Watson, 1982).

In moderated regression analyses, the predictor and moderator variables are entered on the first steps of a hierarchical regression, and the interaction of these two variables is entered on the next step. A significant regression coefficient for the interaction term would indicate that the relationship between the predictor and criterion varies over levels of the moderator variable. If the validity scale operates as hypothesized, a significant interaction will be found such that agreement between predictor and criterion is lower for individuals who score in the invalid range of the scale.

What criteria should be used in moderated regression analyses? As in suppressor analyses, hetero-method criteria must be used to evaluate validity related to systematic biases. But because random error reduces the validity of a scale with respect to all criteria, random responding on both of two self-reports scales would not inflate their correlation; rather, it would tend to reduce it. As a consequence, mono-method as well as hetero-method criteria can legitimately be used to evaluate validity scales that purport to assess sources of random error. Moderated regression analyses are summarized in Table 2.

Detecting Aberrancy

The analyses proposed so far have considered all validity indicators as continuous variables, on the assumption that test scores may be distorted in varying degrees by response sets and styles. An alternative assumption is that most scores are valid but that even in a volunteer sample a small percentage may be seriously distorted. Such test aberrancy (Tellegen, 1988) may be due to deliberate faking, to grossly uncooperative responding, or to simple clerical error“say, answering across rows instead of down columns on an optically scanned answer sheet.

Validity scales are thus commonly used to screen out a small number of suspicious protocols. For example, Church and Burke (1994) discarded about 4% of the MPQ records on the basis of the INVAL scale, and 3% of their NEO-PI records on the basis of an inconsistency scale they created. It is, however, only a presumption that these cases were invalid. Perhaps only a few responses were made in error, with little effect on the substantive scale scores; perhaps apparently inconsistent responses were both accurate in describing these unusual individuals.

A straightforward way to test these alternatives would be to compare validity coefficients for putatively valid and invalid subsamples. That approach was used by Hough et al. (1990), who showed small but consistent differences for groups classified on the basis of a random responding scale, but no consistent differences for groups classified on the basis of social desirability, poor impression, or self-knowledge scales. These null findings are noteworthy because Hough et al. began with a sample of

Table 2
Outline of Planned Moderated Regression Analyses of the Bias and Error Scales of the NEO-PI-R and MPQ

Predictor	Moderator	Criterion	Regression
MPQ self	MPQ Error Scales	NEO self	13 × 5 = 65
	MPQ Error, Bias Scales	NEO mean peer	13 × 8 = 104
NEO self	NEO Error Scales	MPQ self	13 × 3 = 39
	NEO Error, Bias Scales	NEO mean peer	5 × 5 = 25
NEO first peer	NEO Error, Bias Scales	MPQ self	13 × 5 = 65
	NEO Error, Bias Scales	NEO self	5 × 5 = 25
	NEO Error, Bias Scales	NEO second peer	5 × 5 = 25
Total			348

Note. NEO-PI-R = Revised NEO Personality Inventory; MPQ = Multidimensional Personality Questionnaire. MPQ Random Error Scales = VRIN (Variable Response Inconsistency), TRIN (True Response Inconsistency), TRIN2, SLIPS (Associative slips), INVAL (Invalidity). MPQ Systematic Bias Scales = VIRTUES (Unlikely Virtues), DRIN (Desirable Response Inconsistency), DRIN2. NEO Random Error Scales = INC (Inconsistency), SYNCORR (correlation, synonymous), ANTCORR (correlation, antonymous). NEO Systematic Bias Scales = PPM (Positive Presentation Management), NPM (Negative Presentation Management).

over 9,000 respondents, so even the invalid subsamples were quite large.

In the present study, the much smaller sample size prohibits these analyses. Even comparisons of the full sample with a "purified" sample eliminating aberrant respondents scoring in the top 5% of the distribution for each validity scale[3] are not persuasive; eliminating 9 cases from a sample of 178 is unlikely to have much effect, even if all 9 are clearly invalid. Tests of individual validity scales are thus not feasible in this study, but it is possible to test the entire set of validity scales by designating as aberrant any respondent who scores in the top 5% on any one of the scales. That analysis will contrast respondents who are invalid for any of several reasons (inconsistency, acquiescence, socially desirable responding) with respondents who, according to every available index, have provided honest and accurate data.

The two studies that follow evaluated the efficacy of MPQ and NEO-PI-R validity scales. The first study evaluated all relevant validity measures in a sample of American undergraduate students. The second study offered a partial replication in a large German sample of twins.

Study 1

Method

Participants. Participants consisted of 178 individuals (106 women, 72 men) between the ages of 17 and 78 (M = 29.1, SD = 12.03). All participants volunteered. One hundred one were college students; the rest were graduate students in a pastoral counseling program. Participants were requested to have two individuals who had known them for at least 6 months rate them on a personality questionnaire. One hundred sixty-eight participants returned two peer evaluations, while the remaining 10 returned one. Of the 346 raters, 207 were women and 138 were men, with 1 rater not indicating gender. Overall, raters knew their targets quite well, with an average length of acquaintance of over 10 years (M = 10.5, range from 3 months to 45 years). Raters were also asked to indicate how well they knew the participants on a 1 (*not really that well, casual acquaintance*) to 7 (*know each other very well, close friends*) Likert scale. A mean rating of 6.1 (SD = 1.2) was obtained, indicating that the peer raters believed themselves to be very well acquainted with the participant. Self–peer and peer–peer agreement for the college students in this sample has been reported previously (Piedmont, 1994).

Measures. The NEO-PI-R (Costa & McCrae, 1992) is a 240-item questionnaire developed through rational and factor analytic methods to measure the five major factors or domains of personality: Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A), and Conscientiousness (C). Items are answered on a 5-point scale ranging from *strongly disagree* to *strongly agree*, and scales are balanced to control for the effects of acquiescence. There are two forms of the NEO-PI-R, for self-reports (Form S) and observer ratings (Form R); we used both in this study. Form R contains the same questions as Form S phrased in the third person. Normative internal consistency estimates for the domain scales for adults range from .86 to .92 for Form S, and from .89 to .95 for Form R (Costa & McCrae, 1992). Six-year stability coefficients range from .68 to .83 in both self-reports and peer ratings for Neuroticism, Extraversion, and Openness. Three-year retest coefficients of .63 and .79 were seen for brief versions of the A and C domain scales (Costa & McCrae, 1988). NEO-PI-R scales have shown evidence of convergent and discriminant validity across instruments, methods, and observers (Costa, McCrae, & Dye, 1991; McCrae & Costa, 1992; Piedmont & Weinstein, 1993) and have been related to a number of life outcomes including frequency of somatic complaints, ability to cope with stress, and burnout (Costa & McCrae, 1989; Piedmont, 1993). Proposed NEO-PI-R validity scales are described in the introduction of the present article.

The MPQ (Tellegen, 1978/1982) is a 300-item, true–false questionnaire designed to clarify important dimensions of personality (see Tellegen & Waller, in press, for a description of how the MPQ was developed). There are 11 primary scales assessed by the MPQ: Well-Being, Social Potency, Achievement, Social Closeness, Stress Reaction, Alienation, Aggression, Control, Harm Avoidance, Traditionalism, and Absorption. The median alpha reliability for the scales is .85 (range = .76–.90), and 30-day retest coefficients obtained on a college sample had a median value of .89 (range = .82–.91). The MPQ has demonstrated substantial construct and criterion validity (Church & Burke, 1994; Tellegen & Waller, in press). MPQ validity scales are described in the introduction of the present article.

Procedure. Participants were given the packet of test materials to complete on their own time. Participants completed the NEO-PI-R and MPQ in counterbalanced order to control for any order effects.

Participants contacted two individuals who had known them for at least 6 months and gave them an envelope that contained Form R of the NEO-PI-R. Raters completed this form at a time and place separate from the person they were rating. After completing their ratings, they placed all the materials in the envelope provided, sealed the envelope, and returned it to the participant. When all materials were completed, the participant returned them directly to the experimenter.

Unless otherwise noted, all NEO-PI-R ratings use scores aggregated across the two raters. The raw scores, along with scores from all the other scales, were converted to t scores based on normative data provided in the respective test manuals.

Analyses. For the suppressor analyses we compared zero-order with semipartial correlations. Under the hypothesis that validity scales measure irrelevant variance that can and should be removed, each of these semipartial correlations should be higher than the corresponding zero-order correlation. For the moderated regression analyses, we first examined the statistical significance of the interaction term, and then, where significant, we determined the direction of the effect. Power calculations showed that, with 178 cases, we had an 80% chance of detecting a significant (alpha = .05) interaction that accounted for as little as 3.6% of the total variance in the criterion (equivalent to a simple correlation of .19); we had a 50% chance of detecting an interaction that accounted for only 1.8% of the variance in the criterion. Note, however, that because all analyses were conducted on the same sample, statistical tests were not independent. For the aberrancy analyses we tested the significance of the difference between convergent validity coefficients in aberrant and normal subsamples.

Results and Discussion

Preliminary analyses examined descriptive statistics for each of the validity scales used in this study. There was a wide range of scores on each of the validity scales for both the MPQ and NEO-PI-R. Our sample showed near normative values on all MPQ validity scales save VIRTUES and VRIN, where these participants scored somewhat higher. Results for the NEO-PI-R validity scales were consistent with information available from other researchers (L. R. Goldberg, personal communication, January 25, 1995; Schinka et al., 1997). These findings gave confidence that there was sufficient variability (both statistically and substantively) on

[3] Lewis R. Goldberg suggested these analyses and this cutpoint, and we explored them by examining the cross-observer validity of NEO-PI-R domain scores in the full sample and in subsamples screened on the basis of each of the validity scales in turn. None of the screenings had much effect on the magnitude of the correlations, and the small changes observed were as likely to decrease as to increase the validity coefficients.

the moderator scales to warrant their use in the regression analyses and that the present sample was comparable to other volunteer research samples.

Suppressor analyses. For each of the analyses outlined in Table 1, we calculated semipartial correlations between the predictors (controlling for the validity scale) and criteria, and we compared these values to the corresponding zero-order correlations. Of the 82 comparisons, the semipartial correlations were larger than the simple correlation only 18 times (22%). The largest increase in predictive validity was only from .28 to .34. In 14 instances there was no difference to the second decimal place between the semipartial and simple correlation, and in 50 instances (61%), the semipartial correlation was actually smaller than the corresponding simple correlation. It appears that "correcting" the scales actually reduced their predictive validity in most cases.

Such a result can be explained by noting that the hypothesized suppressor variable may actually have substantive content that is related to the criterion (cf. Diener et al., 1991). Individuals who score high on PPM, for example, are not merely pretending to have desirable qualities; they are rated by their peers as being in fact lower in Neuroticism, $r(176) = -.21$, $p < .01$, and higher in Conscientiousness, $r(176) = .23$, $p < .01$. Such data imply that content scales are not contaminated by biases; instead, validity scales are contaminated by substantive content.

Moderated regression analyses. To evaluate the utility of the validity scales as moderator variables, we performed a series of regression analyses (see Table 2 for an overview). Each analysis tested the hypothesis that the validity of a content scale (e.g., MPQ Well-Being) in predicting a criterion (e.g., NEO-PI-R Extraversion) would be moderated by standing on an validity scale (e.g., MPQ VRIN). Note that the moderator variable is taken from the same instrument and respondent as the predictor variable (because it is the validity of that test administration that the moderator scale is supposed to assess), whereas the criterion is from a different instrument or respondent (because it is supposed to be an independent data source). Note also that both random error and systematic bias scales can be evaluated using a criterion from a different observer, whereas only random error scales can be appropriately evaluated when the same individual provides both predictor and criterion data. A total of 348 regressions were thus examined.

Results from the first step of these regressions speak to the unmoderated validity of the substantive scales. All of the associations were in the expected direction, and 309 of them (89%) were significant. Clearly, the content scales of the MPQ and NEO-PI-R show substantial cross-instrument and cross-observer validity. In stark contrast, despite the considerable statistical power, the Predictor × Moderator interaction terms were significant in only 20 regressions (6%), almost exactly the number to be expected by chance. No pattern was apparent among these effects with respect to the substantive scales or validity scales involved.

Another way to illustrate the direction of the effect in these 20 cases is by examining the correlation between predictor and criterion for individuals above and below the median on the validity scale. These correlations are reported in Table 3. One would of course hypothesize that correlations would be higher among individuals with more valid data; yet Table 3 shows that this is true for only 8 of the 20 analyses. In fact, the median correlation is

Table 3
Correlations Between Predictor and Criterion Scales Separately for More and Less Validity Score Groups

			Validity group	
Predictor	Criterion	Moderator	More	Less
Achievement	R-Conscientiousness	VRIN	.21	.45
Social Closeness	R-Extraversion	SLIPS	.30	.51
Social Potency	R-Openness	SLIPS	.16	.43
Stress Reaction	R-Neuroticism	DRIN2	.58	.35
Social Closeness	R-Extraversion	DRIN	.23	.61
Well-Being	R-Extraversion	TRIN	.31	.47
Harm Avoidance	(low) R-Openness	DRIN2	.26	−.08
Aggression	(low) R-Agreeableness	INVAL	.47	.30
Well-Being	S-Extraversion	TRIN	.57	.56
Stress Reaction	S-Neuroticism	SLIPS	.77	.72
Social Potency	S-Openness	SLIPS	.02	.33
Harm Avoidance	(low) S-Openness	SLIPS	−.01	.24
S-Neuroticism	(low) Well-Being	S-ANTCORR	.62	.27
S-Openness	R-Openness	S-SYNCORR	.44	.62
R-Openness	Well-Being	R-SYNCORR	.34	.22
R-Extraversion	Social Potential	R-INC	.49	.52
R-Extraversion	S-Extraversion	R-INC	.62	.58
R-Openness	S-Openness	R-INC	.51	.54
R-Extraversion	S-Extraversion	R-SYNCORR	.58	.62
R1-Neuroticism	R2-Neuroticism	R-SYNCORR	.27	.54
Median			.39	.49

Note. R = peer-rated scale; S = self-reported scale. Criteria marked "(low)" are reflected such that all validity coefficients should be positive. VRIN = Variable Response Inconsistency; SLIPS = Associative slips; DRIN = Desirable Response Inconsistency; TRIN = True Response Inconsistency; INVAL = Invalidity; ANTCORR = correlation, antonymous; SYNCORR = correlation, synonymous; INC = Inconsistency.

substantially higher in the putatively less valid than in the more valid group. These analyses do not support the utility of any of the proposed moderator scales in this sample.

Screening out invalid protocols. To examine the effect of screening aberrant protocols from the sample, we divided the sample into aberrant ($n = 73$) and normal ($n = 103$) subsamples. The former consisted of individuals with extreme scores (top 5% in this sample) on at least 1 (and as many as 4) of the 13 self-report validity indicators; the latter consisted of individuals who had valid scores by all 13 criteria.

We computed convergent validity coefficients for the MPQ and NEO-PI-R scores for both subsamples, using the appropriate mean peer rated NEO-PI-R domain score as the criterion. As Table 4 shows, none of the 18 comparisons showed a statistically significant difference between the two correlations, and the median correlations were quite similar: .37 for the normal sample, .38 for the aberrant. Although all of the "aberrant" cases appeared suspicious for one reason or another, as a group they appear to be no less valid than other respondents are.

Overall, the pattern of findings in this study clearly demonstrates that none of the validity scales proposed for the MPQ or the NEO-PI-R functions as a suppressor or moderator of their respective content scales' validity in this sample. Perhaps that might be attributed to the sample—the pastoral counseling students, in particular, may be more conscientious and honest than most volunteer respondents. Perhaps it is attributable to the instruments used: Baer, Ballenger, Berry, and Leben (1997) reported that adolescents admit occasional random responses when completing the adolescent version of the MMPI. But together with previous research (e.g., Borkenau & Ostendorf, 1992; Smith, 1997), the present results support the conclusion that validity scales generally do not provide useful information for detecting response distortions or aberrant cases in normal volunteer samples.

Study 2

Given the importance placed on validity scales in applied contexts, claims disputing their value must be based on a serious effort to evaluate their efficacy. Study 1 reported data from a single sample of merely moderate size. Moderator effects may be subtle and may be detectable only in very large samples. Study 2 addresses these issues by evaluating validity scales of the MPQ and NEO-PI-R in a large sample of German volunteers, thus providing an opportunity to evaluate the usefulness of validity scales cross-culturally. A failure to find consistent effects for these scales would provide evidence of their lack of utility in two different cultures.

The analyses presented here provide a partial replication and extension of those in Study 1. All analyses use a hetero-method criterion (peer ratings), but within self-reports we extend analyses to include cross-instrument moderators—that is, we examine NEO-PI-R validity scales as moderators of MPQ validity, and vice versa. These analyses are appropriate under the assumption that response sets or styles used in the completion of one questionnaire may also operate in the completion of others. These analyses parallel the common situation in which a separate validity measure (e.g., the Marlowe–Crowne Social Desirability scale) is administered in conjunction with a personality questionnaire. Finally, given the very large sample size, it was possible to split the sample, allowing cross-validation of any observed effects.

We followed two analytic strategies; in each, the criteria were derived from peer ratings on the short version of the NEO-PI-R, the NEO-Five-Factor Inventory (NEO-FFI). First, we conducted a series of suppressor analyses partialing out indexes of systematic bias (VIRTUES, PPM, and NPM) from the self-report scales. Next, we performed moderated regression analyses. We used MPQ and NEO-PI-R self-reported scores as predictors, and we examined

Table 4
Convergent Validity Coefficients Between NEO-PI-R Observer Ratings and NEO-PI-R and MPQ Self-Reports for the Normal and Aberrant Samples

Comparison	Aberrant sample ($n = 75$)	Normal sample ($n = 103$)	z difference
S-Neuroticism with R-Neuroticism	.39**	.52**	−1.06
S-Extraversion with R-Extraversion	.67**	.52**	1.50
S-Openness with R-Openness	.50**	.56**	−0.54
S-Agreeableness with R-Agreeableness	.57**	.44**	1.12
S-Conscientiousness with R-Conscientiousness	.52**	.38**	1.13
Well-Being with (low) R-Neuroticism	.30*	.29*	0.07
Well-Being with R-Extraversion	.42**	.36**	0.45
Well-Being with R-Openness	.31**	.27*	0.28
Social Potential with R-Extraversion	.54**	.46**	0.69
Social Potential with R-Openness	.30**	.28*	0.14
Achievement with R-Openness	.06	.16	−.65
Achievement with R-Conscientiousness	.32*	.32**	0.00
Social Closeness with R-Extraversion	.34*	.43**	−.68
Stress Reaction with R-Neuroticism	.49**	.45**	0.33
Aggression with (low) R-Agreeableness	.36**	.38**	−0.15
Control with R-Conscientiousness	.44**	.35**	0.69
Harm Avoidance with (low) R-Openness	.13	.06	0.45
Absorption with R-Openness	.33*	.26**	0.49

Note. R = peer-rated scale; S = self-reported scale. Criteria marked "(low)" are reflected such that all validity coefficients should be positive.
* $p < .01$. ** $p < .001$, two-tailed.

validity scales from each instrument (VIRUES, VRIN, PPM, NPM, and INC) as moderators. This analytic strategy yielded 54 suppressor analyses and 90 moderated regression analyses in each half of the sample.

Method

Participants. Participants were twins who were recruited for a behavior genetic study of personality and temperament traits (see Riemann, Angleitner, & Strelau, 1997, for more details). For the present analyses data from all participants, including twins whose cotwin did not complete the questionnaire, were used. All twin pairs were approached through announcements in newspapers, magazines, radio and TV stations or through twin clubs and twin meetings. Twins were recruited from all parts of Germany and were heterogeneous with respect to educational and occupational status. One twin from each pair was assigned to Subsample A, the other to Subsample B. There were 887 (700 female and 187 male) Subsample A participants (mean age = 32.17, *SD* = 12.81) and 841 (639 female and 202 male) Subsample B participants (mean age = 31.93, *SD* = 12.97).

Materials. A number of temperament and personality questionnaires for self- and peer reports were mailed to participants. Among these measures were the German versions of the NEO-PI-R (Costa & McCrae, 1992; Ostendorf & Angleitner, 1994), the MPQ (Tellegen, 1978/1982; Angleitner & Ostendorf, 1999), and a peer-report version of the NEO-FFI (Borkenau & Ostendorf, 1993; Costa & McCrae, 1989). Proposed validity scales for the NEO-PI-R were scored in the German version as in the English version. The German version of the MPQ has been shortened to 276 items by Tellegen, who authorized the German version, and only two validity scales are scored: Unlikely Virtues and VRIN. Two items of the original VRIN were dropped in the shortened German version of the MPQ.

Procedure. Questionnaire data were gathered by mail. We mailed peer rating questionnaires to twins, who were instructed to give them to two peers who knew the respective twin but (preferably) not the cotwin very well. Peers were mostly friends (62%), relatives (16%), spouses (10%), and colleagues (9%) who knew the participants for 11.06 (*SD* = 10.46) years on average; most of them (82%) judged their acquaintance with the target person as "very good" or "good." Very few (1%) indicated that they had little or very little knowledge about the target. The majority of peers were female (62%). Peers sealed their ratings in envelopes that were mailed by the twins back to investigators.

Twins completed self-reports on the NEO-PI-R and MPQ about 10 months later. They were instructed to complete the questionnaires independently and return them by mail to investigators.

Results and Discussion

Suppressor analyses. In Subsamples A and B separately, PPM, NPM, and VIRTUES were partialed in turn from the self-report NEO-PI-R and MPQ scales. The 108 residual scores were then correlated with the appropriate NEO-FFI mean peer rating, and we compared each correlation with the corresponding zero-order correlation. If PPM, NPM, and VIRTUES measure presentation styles rather than personality traits, removing variance attributable to them should increase correlations with external criteria.

In fact, however, semipartial correlations were very similar to the zero-order correlations, and where they differed, they were usually smaller rather than larger.[4] The median uncorrected validity coefficient across the two subsamples was .38. The median semipartial validity coefficients, correcting for PPM, NPM, and VIRTUES, respectively, were .32, .38, and .37. The net effect of "correcting" scores was to decrease validity, especially when PPM was used. It appears that those individuals who present themselves most positively are the ones who are deemed by their peers actually to have the most positive traits.

Moderated regression analyses. Each of these 180 analyses tested the hypothesis that standing on one of the validity scales would moderate the validity of the self-reported content scales in predicting the peer-rated criterion. We entered the predictor and validity scales on the first step; in 160 cases (89%) the predictor was significantly related to the criterion, attesting to substantial cross-method validity in the German data. The second step of the regression analyses added the interaction between the content and validity scales. In Subsample A, only 5 of 90 interaction terms (6%) were significant, about what one would expect to find by chance. In Subsample B, 17 (19%) of the interactions were found significant, of which 14 were in the hypothesized direction. However, no effects were significant in the same direction in both subsamples; that is, no cross-validated moderator effects were found.

We split subsamples into hypothetically more and less valid groups on the basis of validity scale scores (as in Table 3, Study 1). Correlations of predictor with criterion within these groups were very similar in more valid (median absolute correlations = .36 and .40 for Subsamples A and B, respectively) and less valid groups (medians = .34 and .40). When we examined the five validity scales separately, the VRIN scale showed the most promising results: The median absolute correlation in groups with lower VRIN scores was .41; in those with higher VRIN scores it was .35. Lower VRIN scores were associated with higher validity coefficients in 27 of the 36 comparisons.

In general, results from these very large samples provide little evidence for the utility of validity scales. When used as suppressor variables, correlations with external criteria tend to decrease rather than increase—a strong contraindication for their use. When analyzed as continuous moderator variables, no consistent results emerged. Only when more and less valid subsamples were defined on the basis of VRIN scale scores was there a trend for putatively less valid self-reports to show lower correlations with mean peer ratings. These findings are consistent with results reported by Archer, Fontaine, and McCrae (1998), who found somewhat diminished validity among individuals identified by high scores on the MMPI-2 VRIN scale.

General Discussion

Validity Scales in Volunteer Samples

In two volunteer samples, from the United States and Germany, a wide range of validity scales failed to enhance the validity of personality assessments. Although other validity scales (e.g., Paulhus, 1998) might have shown different results, the scales examined here were cleverly designed and carefully constructed; why, then, did they fail? One possibility is that the scales function in unanticipated ways. People who wish to make a favorable impression may endorse the items of the PPM, but so may people who are truly well-adjusted and conscientious. Social desirability

[4] Detailed results are available from Ralph L. Piedmont.

scales in general appear to be contaminated with substantive variance (Kozma & Stones, 1987; McCrae & Costa, 1983). In the case of inconsistency scales, researchers may make the unrealistic assumption that careful and honest respondents will invariably make consistent responses. However, as Tellegen himself (1988) pointed out, intraindividual variability may arise for any number of reasons, including the probabilistic nature of trait indicators, the intrusion of state fluctuations on trait scores, and the influence of more than one trait on an item response.

A second reason for the failure of validity scales may be the relatively low base rate of biased or careless responding. Tellegen (1988) noted that it is possible to reduce false positives by setting stringent cutoff points for validity scales, but such an approach may result in screening out so few cases that it is simply not worth the effort. Most researchers would discard an answer sheet in which half the items were left blank or all were answered *strongly agree,* but short of such gross indicators of invalidity, there does not seem to be a good scientific rationale for the routine use of validity scales in research settings. Discarding cases on the basis of indexes with no proven utility at best wastes respondents' time and researchers' money and at worst may bias the sample in ways as yet unknown.

Researchers are sometimes caught between their own interpretation of the literature and the views of reviewers and editors; validity scales are doubtless often used simply to anticipate possible reviewers' objections. Unfortunately, this practice reinforces the received wisdom that validity scales are useful. At a minimum, it would seem, researchers should report results with and without screening or correcting scores and should point out to readers that the use of validity scales is a matter of scientific controversy.

Data in the present study do not speak directly to the utility of validity scales in applied and clinical settings—for that, the analyses reported here would need to be repeated in clinical, forensic, and I/O samples. But the present results do form a context for reviewing and interpreting some of the available literature (e.g., Barrick & Mount, 1996; Ones, Viswesvaran, & Reiss, 1996).

The detection of invalid protocols has always been a central concern in clinical practice (Ben-Porath & Waller, 1992). Because clinicians are called on to treat individual patients, they must take a more idiographic view of test data than researchers typically do. Including one instance of purely random responding in a sample of 1,000 would have no noticeable effect on means or correlations, but a clinician trying to understand a patient on the basis of that protocol might make serious errors in diagnosis and treatment.

For these reasons it is perfectly understandable that clinicians would want to use validity indexes. It is far from clear, however, that existing indexes are truly useful in detecting invalid protocols. For example, Archer et al. (1998) showed that, although not entirely useless, the MMPI-2 VRIN screen resulted in the identification of many false positives or of only partially invalid protocols.

A Multimethod Alternative

Test users truly concerned about the validity of individual protocols cannot rely on the results of faking studies—or even of nomothetic studies (like the present one) using external criteria—to evaluate the worth of a validity indicator. As Ben-Porath and Waller (1992) pointed out, "the validity of a psychological test, when used in clinical assessment, must be evaluated and established for each individual to whom the test is administered" (p. 16). If that is true for substantive scales, it must surely also be true for validity scales—yet that seems to entail the need for an endless series of validity scales to confirm the interpretability of other validity scales. There is, however, another approach entirely: The best evidence on protocol validity, and the best alternative to the use of validity scales, comes from the comparison of self-report scores with independent assessments, on a case-by-case basis.

Figure 1, for example, shows the NEO-PI-R profile of a female college student, one of the 11 cases in the Study 1 sample who scored above the 95th percentile on the INC scale, and who might thus be suspected of random responding. Information from mean peer ratings (dotted line) provides a basis for judging the validity of her self-reports (solid line). Scores on the five factors are given at the left of the profile; toward the right are scores on the 30 facet scales, grouped by domain.

The two profiles are not identical, but they do show considerable agreement, most notably with respect to very high standing on Extraversion. The extent of agreement can be quantified by profile agreement statistics (McCrae, 1993), available with the computer interpretation, which show an overall coefficient of profile agreement of .80—higher than the average self–other agreement in volunteer samples. It is extremely unlikely that such close agreement would be seen if this individual were responding randomly, and it would certainly be a mistake to discard these self-report data.

But observer rating data do more than merely validate questionable self-reports. They are valuable in their own right as information on the way the individual is perceived by others, and they can be averaged with self-reports to provide a more precise estimate of true scores on personality variables. Perhaps most important, they also call attention to areas of substantial disagreement that once again call for additional information gathering (cf. Piedmont, 1998). For example, profile agreement statistics in this case suggest that there are meaningful differences on four facet scales: O1, Fantasy; A2, Straightforwardness; A6, Tender-Mindedness; and C3, Dutifulness. There are many possible reasons for these discrepancies (McCrae, Stone, Fagan, & Costa, 1998): Perhaps the target is thinking of her behavior in situations the observers never see; perhaps the raters have different standards of comparison; perhaps the individual lacks insight into some aspects of her personality. In a clinical setting these are all issues that could be fruitfully pursued.

Although personality psychologists occasionally interpret and report case studies (e.g., Nasby & Read, 1997), most research is conducted at the level of the group, where individual protocols need not be interpreted. Bias and error of measurement are legitimate concerns there, too, however, and multimethod assessment is again a valuable alternative to the use of validity scales. Aggregation across multiple sources tends to reduce random error of measurement, and replication across methods can increase confidence that results are not due to systematic biases.

Conclusion

Obtaining supplementary observer ratings is more laborious than screening or statistically adjusting self-reports, but, to date, there is little evidence that screening and adjusting are effective.

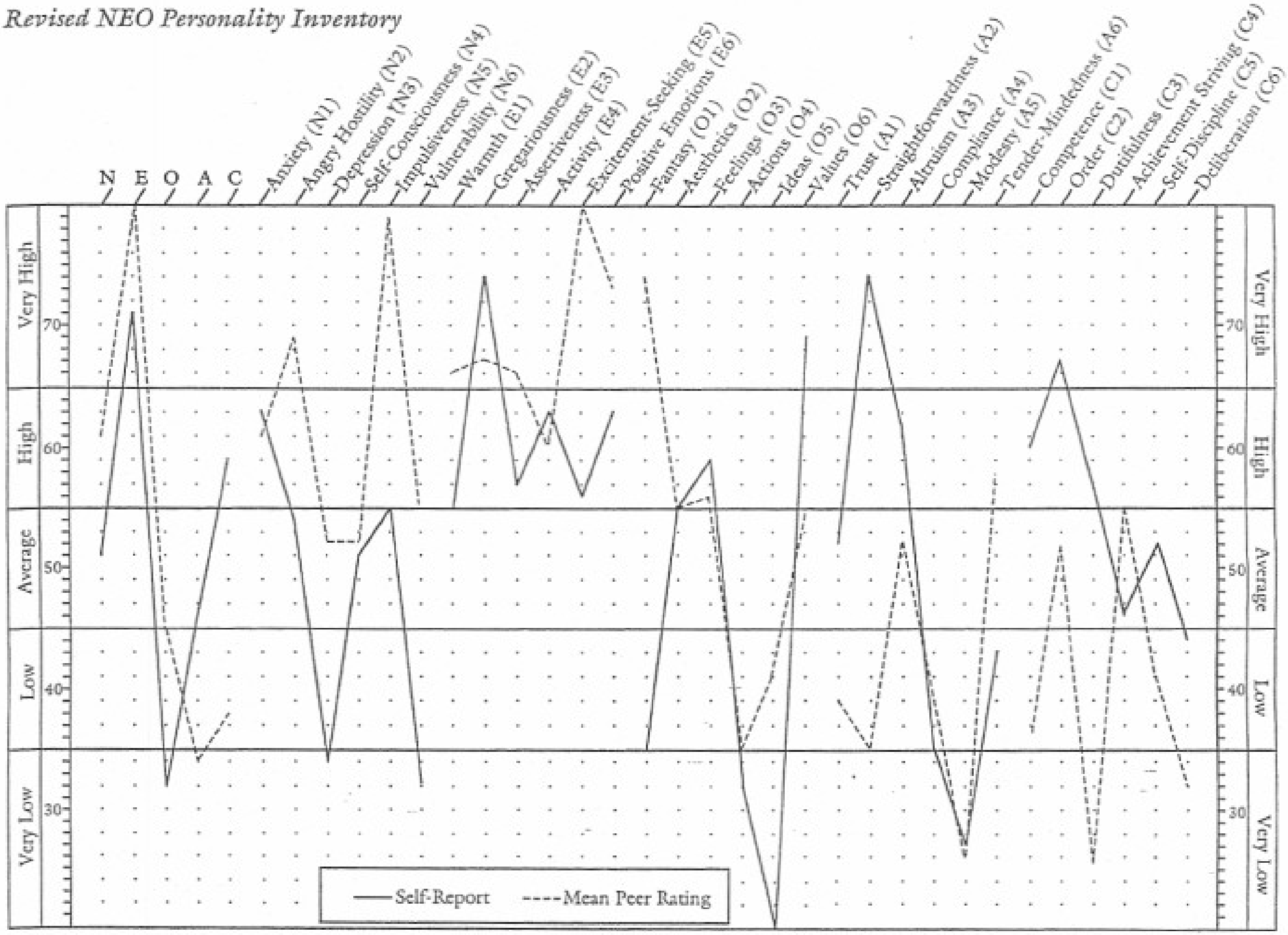

Figure 1. Self-reported (solid line) and mean peer-rated (dotted line) personality profiles of a female college student scoring above the 95th percentile on a measure of response inconsistency. N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. Profile form reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the NEO Personality Inventory—Revised, by Paul Costa and Robert McCrae, Copyright 1978, 1985, 1989, 1992 by PAR, Inc. Further reproduction is prohibited without permission of PAR, Inc.

Even the most sophisticated validity scales offer very limited guidance about the meaning of test responses. Perhaps attention should shift from detecting invalidity toward improving the quality of assessment. Every effort should be made to motivate the respondent and ensure that instructions are understood. Although immensely useful, questionnaires are not an infallible method of assessing personality, and validity scales will not make them so. Instead of blind faith in validity scales, clinicians and researchers must rely on the use of well-validated instruments, the development of rapport with the respondent, and the judicious comparison of multiple sources of data in interpreting results.

References

Alperin, J. J., Archer, R. P., & Coates, G. D. (1996). Development and effects of an MMPI-A *K*-correction procedure. *Journal of Personality Assessment, 67,* 155–168.

Angleitner, A., & Ostendorf, F. (1999). *Manual zur deutschsprachigen Fassung des Multidimensional Personality Questionnaire (MPQ) von Tellegen* [Manual for the German version of Tellegen's Multidimensional Personality Questionnaire (MPQ)]. Unpublished manuscript, University of Bielefeld, Bielefeld, Germany.

Arbisi, P. A., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The infrequency-psychopathology scale, *F(p). Psychological Assessment, 7,* 424–431.

Archer, R. P., Fontaine, J., & McCrae, R. R. (1998). The effects of two MMPI-2 validity scales on basic scale relations to external criteria. *Journal of Personality Assessment, 70,* 87–102.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68,* 139–151.

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

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.

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6,* 199–214.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI)* [NEO Five-Factor Inventory]. Göttingen, Germany: Hogrefe.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen A., & Kaemmer, B. (1989). *MMPI2: Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

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

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66,* 93–114.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Costa, P. T., Jr., & McCrae, R. R. (1988). Personality in adulthood: A 6-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology, 54,* 853–863.

Costa, P. T., Jr., & McCrae, R. R. (1989). Personality, stress, and coping: Some lessons from a decade of research. In K. S. Markides & C. L. Cooper (Eds.), *Aging, stress and health* (pp. 267–283). New York: Wiley.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

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

Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12,* 887–898.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

Dicken, C. (1963). Good impression, social desirability, and acquiescence as suppressor variables. *Educational and Psychological Measurement, 23,* 699–720.

Diener, E., Sandvik, E., Pavot, W., & Gallagher, D. (1991). Response artifacts in the measurement of subjective well-being. *Social Indicators Research, 24,* 35–56.

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

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48,* 82–98.

Goldberg, L. R., Rorer, L. G., & Greene, M. M. (1970). The usefulness of "stylistic" scales as potential suppressor or moderator variables in predictions from the CPI. *ORI Research Bulletin, 10,* 1–31.

Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology, 72,* 435–448.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Port Huron, MI: Research Psychologists Press.

Jackson, D. N. (1989). *Basic Personality Inventory manual.* Port Huron, MI: Sigma Assessment Systems.

Kozma, A., & Stones, M. J. (1987). Social desirability in measures of subjective well-being: A systematic evaluation. *Journal of Gerontology, 42,* 56–59.

Lim, J., & Butcher, J. N. (1996). Detection of faking on the MMPI-2: Differentiation among faking-bad, denial, and claiming extreme virtue. *Journal of Personality Assessment, 67,* 1–25.

McCrae, R. R. (1986). Well-being scales do not measure social desirability. *Journal of Gerontology, 41,* 390–393.

McCrae, R. R. (1993). Agreement of personality profiles across observers. *Multivariate Behavioral Research, 28,* 25–40.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888.

McCrae, R. R., & Costa, P. T., Jr. (1992). Discriminant validity of the NEO-PI-R facet scales. *Educational and Psychological Measurement, 52,* 229–237.

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

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66,* 285–313.

Nasby, W., & Read, N. W. (1997). The life voyage of a solo circumnavigator: Integrating theoretical and methodological perspectives. *Journal of Personality, 65,* 785–1068.

Nicholson, R. A., & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist, 45,* 290–292.

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

Ostendorf, F., & Angleitner, A. (1994, July). *Psychometric properties of the German translation of the NEO Personality Inventory (NEO-PI-R).* Poster presented at the 7th Conference of the European Association of Personality Psychology, Madrid, Spain.

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

Paulhus, D. L. (1998). *Manual of the Balanced Inventory of Desirable Responding (BIDR-7).* Toronto, Ontario, Canada: Multi-Health Systems.

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92,* 486–511.

Piedmont, R. L. (1993). A longitudinal analysis of burnout in the health care setting: The role of personal dispositions. *Journal of Personality Assessment, 61,* 457–473.

Piedmont, R. L. (1994). Validation of the rater form of the NEO-PI-R for college students: Toward a paradigm for measuring personality development. *Assessment, 1,* 259–265.

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

Piedmont, R. L., & Weinstein, H. P. (1993). A psychometric evaluation of the new NEO-PI-R facets scales for Agreeableness and Conscientiousness. *Journal of Personality Assessment, 60,* 302–318.

Putnam, S. H., Kurtz, J. E., & Houts, D. C. (1996). Four-month test–retest reliability of the MMPI-2 with normal male clergy. *Journal of Personality Assessment, 67,* 341–353.

Riemann, R., Angleitner, A., & Strelau, J. (1997). Genetic and environmental influences on personality: A study of twins reared together using the self- and peer report NEO-FFI scales. *Journal of Personality, 65,* 449–475.

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

Smith, H. L. (1997). *The structure and utility of social desirability scales in psychological research.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Stein, L. A. R., Graham, J. R., & Williams, C. L. (1995). Detecting fake-bad MMPI-A profiles. *Journal of Personality Assessment, 65,* 415–427.

Tellegen, A. (1978/1982). *Brief manual of the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 621–663.

Tellegen, A., Kamp, J., & Watson, D. (1982). Recognizing individual differences in predictive structure. *Psychological Review, 89,* 95–105.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.), *Handbook of adult personality inventories.* New York: Plenum.

Received February 2, 1999
Revision received July 26, 1999
Accepted August 2, 1999 ■