Development and initial validation of validity scales for the NEO-Five Factor Inventory

Donald J. Scandell\textsuperscript{a, b,*}

\textsuperscript{a}Department of Social Work, Southern Illinois University at Edwardsville, Edwardsville, 662026-1450, USA
\textsuperscript{b}Missouri Institute of Mental Health, St Louis, MO, USA

Received 21 May 1999; received in revised form 1 November 1999; accepted 13 December 1999

Abstract

Based upon the methodology established by Schinka, Kinder, and Kremer (Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research validity scales for the NEO-PI-R: Development and initial validation. Journal of Personality Assessment, 68, 127–138), a set of validity scales were developed for the NEO-Five Factor Inventory (NEO-FFI, Costa, P. T., & McCrae, R. R. (1992). Revised NEO Personality Inventory and NEO Five Factor Inventory Professional Manual. Odessa, FL: Psychological Assessment Resources). In Study 1, 111 undergraduates completed the NEO-FFI and item means, standard deviations, and intercorrelations were used to construct measures of Positive Impression Management, Negative Impression Management, and Inconsistent Responding. In Study 2, 146 participants completed the NEO-FFI under one of five instructional sets (control, fake good, fake bad, graduate psychology and police academy admissions). A set of randomly produced NEO-FFI profiles were added to this data set. ANOVA results provided support for the utility of the validity scales, as they were differentially sensitive to random responding, positive and negative impression management in hypothesized ways. \copyright 2000 Elsevier Science Ltd. All rights reserved.

Keywords: NEO-FFI validity scales; Response bias

1. Introduction

The Five-Factor model of personality has become one of the dominant theoretical
perspectives in personality psychology (Goldberg, 1993). The most commonly used measures of
the Five-Factor Model are the two versions of the NEO: the 240-item NEO Personality
Inventory-Revised (NEO PI-R) and the 60-item NEO-Five Factor Inventory (NEO-FFI, Costa
& McCrae, 1992). These forms of the NEO have been employed in a wide range of applied
and research settings including personnel selection, health, career counseling, psychotherapy
and basic personality research (Costa & McCrae, 1992). However, questions have been raised
concerning the susceptibility of these inventories to deceptive responding (Ben-Porath &
Waller, 1992). Unlike other widely used personality measures such as the Minnesota
Multiphasic Personality Inventory, the NEO PI-R and NEO-FFI were not originally developed
with validity scales. Therefore, the self-presentation strategies of respondents could not be
assessed directly. Some have contended that the lack of validity scales made the NEO
vulnerable to undetected faking, an important concern in assessment.

Recently, Schinka, Kinder, and Kremer (1997) developed three validity scales for the NEO-
PI-R: Positive Impression Management, Negative Impression Management, and Inconsistency.
Schinka and his associates found that these scales were differentially sensitive to faking good,
faking bad, and random responding. Unfortunately, the validity scales developed for the NEO
PI-R cannot be used with the NEO-FFI because they share only a fraction of common items.
As such, it is necessary to develop new validity scales based exclusively on items from the
NEO-FFI pool.

There have been three studies investigating “faking” on the NEO-FFI. The results have
consistently demonstrated that respondents are able to manipulate their test performance in
such a way as to create the impression they desire, thereby invalidating interpretations based
upon test scores (Paulhus, Bruce, & Trapnell, 1995; Scandell & Wlazelek, 1996; Topping &
O’Gorman, 1997). Although Scandell and Wlazelek (1996) developed some general guidelines
concerning the detection of faking based upon extreme scores, to date no validity scales have
been developed for the NEO-FFI. The purpose of this research is to develop validity scales for
the NEO-FFI.

Development of validity scales for the NEO-FFI would have ramifications for both basic
and applied research. The ability to detect faking may increase the range of situations
appropriate for instrument use, and it may also aid basic researchers to understand the
processes involved in positive and negative self-presentation on personality inventories.
Moreover, validity scales may provide important information concerning the test-taking
approach of respondents including motivational concerns and self-presentational strategies.

2. Study 1

The purpose of Study 1 was to select items for inclusion on the NEO-FFI Positive
Impression Management (PIM), Negative Impression Management (NIM), and Inconsistency
(INC) scales. The data used in Study 1 were collected as part of a previously published study
examining the relationship between private self-consciousness and NEO-FFI reliability indices
(Scandell, 1998).
2.1. Participants

Participants were recruited from undergraduate classes at a northeastern state university. Volunteers (84 women and 27 men) were paid US $5.00 for their participation. Their mean age was 20.5 years, the median was 20.0 years, and the mode was 18.0 years. Eighty-nine percent of the sample were between 18 and 22 years old.

2.2. Measures

All participants were asked to complete the Self-Consciousness Scale (SCS; Fenigstein, Scheier, & Buss, 1975), the NEO-FFI and a brief demographic questionnaire. Since data from the SCS were not included in the current study, properties of this measure will not be described here.

The NEO-FFI (Costa & McCrae, 1992) is a brief 60-item version of the NEO PI-R. As with the NEO PI-R, the NEO-FFI is a measure of the Five Factor Model of personality and yields scores on the following domains: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Each item is scored on a five-point scale. Estimates of internal consistency for this inventory have ranged from 0.68 for Agreeableness to 0.86 for Neuroticism (Costa & McCrae, 1992). Scores on the NEO-FFI are highly correlated with those on the NEO PI-R, with coefficients ranging from 0.87 to 0.92 except for Agreeableness (0.77). The NEO-FFI was developed by culling those items from the NEO PI-R that demonstrated the best discriminant and convergent validity (Costa & McCrae, 1992).

2.3. Procedure

Items were selected based upon the procedures outlined by Schinka et al. (1997). For the PIM and NIM scales, items deviating from the mean in an extreme positive or negative direction were examined as potentially relevant to the validity scales. Schinka et al. (1997) defined extreme mean scores as 0.5 standard deviations above or below the item score midrange of 2.0. Although this definition was originally applied in this study, the actual standard deviations of selected items approached 1 standard deviation above or below the midrange score of 2.0. Thirty-four items (18 PIM and 16 NIM) were initially selected for examination as potential validity scale items based upon item means and standard deviations. Selection of final PIM and NIM items included consideration of extremity of the item means, item standard deviations, corrected item-total scale correlations, alpha if item deleted, squared multiple correlation, domain membership and directionality of item. An attempt was made to select two items from each of the NEO-Five Factor Inventory domains. Within each domain, the directionality of items were considered. On the NEO-Five Factor Inventory, some items are scored in a positive direction such that endorsement of an item results in a higher domain score. Other items are scored in a negative direction, such that disagreement with an item results in a higher score. In order to avoid confounding item directionality with scores on the validity scales, an effort was made to select one positively and one negatively keyed item in each of the five personality domains.

Selection of items for the INC scale was based upon an examination of inter-item
correlations. Scores on the INC scale were intended to reflect inconsistency in responding to pairs of items that were highly correlated. Pairs of items that were correlated above 0.7 were considered for scale inclusion. Employment of this criteria resulted in 66 item-pairs being selected for initial consideration for inclusion on the Inconsistency scale. Two item-pairs from each of the five domains of the NEO-FFI were selected. Additionally, no item was included twice. In general, pairs of items demonstrating the highest correlations were selected.

3. Results

Table 1 presents descriptive statistics for the three validity scales, and Cronbach’s alpha for the PIM and NIM scales. The internal consistency levels obtained for the PIM and NIM scales in the present study were consistent with those obtained by Schinka et al. (1997) for the NEO PI-R. The final PIM and NIM scales each consisted of 10 items. Items on the PIM scale included: N11, N12, E2, E9, O4, O11, A1, A8, C8 and C11. For Items N11, N12, E2, O11, A1, and C8, the “strongly disagree” response category is scored zero using the original NEO-FFI scoring. For items E9, O4, A8, and C11, using the scoring criteria provided for the original NEO-FFI, the “strongly disagree” category is scored a four. Directionality of items is not confounded with the PIM scale, since six items are scored in one direction, with the remaining four items scored in the opposite direction.

The NIM scale included the following items: N1, N3, E6, E12, O2, A5, A6, A9, C2, and C6. Directionality of item response is not confounded with the NIM scale. For five of the items, the response “strongly disagree” corresponds to a score of zero (N3, E2, O11, A1, C8). For the other five items on the NIM scale the “strongly disagree” response corresponds to a score of four (N1, E9, O4, A8, C11).

The INC scales included the following pairs of items: N11–N9 (r = 0.79); N6–N5 (r = 0.71); E11–E8 (r = 0.78); E7–E9 (r = 0.70); O5–O9 (r = 0.73); O11–O4 (r = 0.73); A1–A2 (r = 0.81); A10–A8 (r = 0.81); C12–C10 (r = 0.88); C7–C4 (r = 0.84). To construct the INC score, the raw score difference between each item pair was calculated, and the absolute value of this figure was determined. These ten absolute values were summed to provide the final INC scale score. This procedure is a replication of the methodology employed by Schinka et al. (1997).

Table 2 provides the correlations between the NEO-FFI domain scales and the three validity scales. The correlations between the three validity scales and the NEO-FFI domain scales in

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity scale alpha coefficients, scale means, and scale standard deviations in the developmental study</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale</th>
<th>Scale M</th>
<th>Scale SD</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIM</td>
<td>29.33</td>
<td>4.79</td>
<td>0.67</td>
</tr>
<tr>
<td>NIM</td>
<td>16.79</td>
<td>4.90</td>
<td>0.51</td>
</tr>
<tr>
<td>INC</td>
<td>7.14</td>
<td>3.49</td>
<td>–</td>
</tr>
</tbody>
</table>

*a Note. PIM = Positive Impression Management scale; NIM = Negative Impression Management scale; INC = Inconsistency scale; scale scores ranges are 0–40 for each scale.*
the present study are consistent with those obtained by Schinka et al. (1997). Also, the correlations obtained between the validity scales and the NEO-FFI domain scores in the present study are similar to those reported by Morey (1991) for correlations between the validity scales of the Personality Assessment Inventory and NEO-PI-R domain scores. Although many of the correlations obtained between the validity scales and the NEO-FFI domains in the current study are significant and in the moderate range, they are consistent with previous findings.

4. Study 2

In Study 2, the validity scales developed in Study 1 were applied to a data set in order to determine the sensitivity of these scales to instructional set. The data used in Study 2 were collected as part of a previously published study (Scandell & Wlazelek, 1996) examining the fakeability of the NEO-FFI.

4.1. Participants

Participants were recruited from undergraduate psychology classes at a northeastern state university. As inducement for participation, participants were offered bonus points on an upcoming examination. The sample included 89 women and 57 men. Participants ranged in age from 18 to 52 years, with both a median and mode of 21 years of age.

4.2. Procedure

Participants were randomly assigned to one of five instructional sets for completion of the NEO-FFI: fake good (present self in best possible light), fake bad (present self in worst

Table 2
Comparison of validity scale and NEO domain scale correlations in present study with Schinka et al. (1997)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Neuroticism</th>
<th>Extraversion</th>
<th>Openness</th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIM</td>
<td>-0.51</td>
<td>-0.59</td>
<td>0.40</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td>NIM</td>
<td>0.44</td>
<td>0.37</td>
<td>-0.43</td>
<td>-0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>INC</td>
<td>0.23</td>
<td>0.07</td>
<td>-0.25</td>
<td>0.02</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Note: PIM = Positive Impression Management scale; NIM = Negative Impression Management scale; INC = Inconsistency scale. Scales corrected for item overlap. The NEO-FFI was used in the present study, referred to as Scandell in Table 2. The NEO-PI-R was employed in the Schinka et al. (1997), referred to in the Table as Schinka. Level of significance was not reported in the Schinka et al. study, so a comparison could not be made. All correlations above 0.20 were significant in the present study at the 0.05 level.
possible light), police (present self to maximize chances of admission to the police academy), psychology (present self to maximize chances of admission to a graduate psychology program), and control (standard directions). To this existing data set, 29 computer generated random-response NEO-FFI profiles were added (Random group).

It was predicted that participants who were instructed to present themselves in a favorable manner (fake good, psychology and police conditions) would score significantly higher on the Positive Impression Management validity scale than individuals in the control condition. Further, it was predicted that individuals who were instructed to depict themselves in a negative manner (fake bad) would score significantly higher on the Negative Impression Management validity scale. Finally, the randomly generated profiles were predicted to have significantly higher scores on the Inconsistency scale in comparison to the control condition.

5. Results

Three one-way analyses of variance were performed with PIM, NIM, and INC scores serving as the dependent variables and instructional set (control, fake good, fake bad, Police, Psychology, and Random) as the independent variable (see Table 3). In terms of PIM scores, a significant main effect was obtained for instructional set, $F(5, 168) = 91.01, p < 0.0001$. Post hoc analysis (Tukey HSD, alpha = 0.05) revealed that participants in the fake good and psychology groups had significantly higher PIM scores than participants in all of the other conditions. The mean PIM scores of the control condition were significantly higher than the random and fake bad conditions. Also, individuals in the police condition had significantly higher PIM scores than participants in fake bad and random conditions. However, PIM scores in the police condition were not significantly different from those in the control, fake good, or psychology conditions.

In terms of NIM scores, ANOVA results indicated a significant effect for instructional set, $F(5, 168) = 21.35, p < 0.0001$. Post hoc analysis (Tukey HSD, alpha = 0.05) revealed participants in the fake bad condition had the highest NIM scores, significantly different than all of the other groups. The mean scores on the NIM scale for the fake good, psychology, and police conditions were all significantly lower than the mean scores for the other conditions, but

<table>
<thead>
<tr>
<th>Group</th>
<th>Control</th>
<th>Good</th>
<th>Psych</th>
<th>Police</th>
<th>Bad</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIM</td>
<td>48.09&lt;sub&gt;b&lt;/sub&gt;</td>
<td>56.87&lt;sub&gt;a&lt;/sub&gt;</td>
<td>58.50&lt;sub&gt;a&lt;/sub&gt;</td>
<td>54.92&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>5.54&lt;sub&gt;d&lt;/sub&gt;</td>
<td>31.49&lt;sub&gt;c&lt;/sub&gt;</td>
</tr>
<tr>
<td>NIM</td>
<td>49.11&lt;sub&gt;b&lt;/sub&gt;</td>
<td>34.75&lt;sub&gt;c&lt;/sub&gt;</td>
<td>32.45&lt;sub&gt;c&lt;/sub&gt;</td>
<td>36.44&lt;sub&gt;c&lt;/sub&gt;</td>
<td>59.45&lt;sub&gt;a&lt;/sub&gt;</td>
<td>41.72&lt;sub&gt;bc&lt;/sub&gt;</td>
</tr>
<tr>
<td>INC</td>
<td>49.40&lt;sub&gt;b&lt;/sub&gt;</td>
<td>42.48&lt;sub&gt;b&lt;/sub&gt;</td>
<td>42.53&lt;sub&gt;b&lt;/sub&gt;</td>
<td>43.67&lt;sub&gt;b&lt;/sub&gt;</td>
<td>43.67&lt;sub&gt;b&lt;/sub&gt;</td>
<td>69.76&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> Note. Across rows, groups that have a subscript in common are not significantly different from each other.

<sup>b</sup> PIM = Positive Impression Management scale; NIM = Negative Impression Management scale; INC = Inconsistency scale.
were not different from each other. The control and random groups were midway between the means for the fake bad condition and the cluster of fake good conditions (fake good, police, and psychology instructional sets).

A significant effect was found for instructional set on INC scores, \(F(5, 167)=23.48, p < 0.0001\). Post hoc analysis (Tukey HSD, alpha=0.05) indicated that the random group had significantly higher INC scores than all of the other conditions. The police, fake good, psychology, control, and fake bad conditions were not significantly different from each other.

6. Discussion

The results of this study provide preliminary support for the use of the PIM, NIM, and INC scales as a research tool. When these validity scales were applied to a sample of profiles in which participants were instructed to intentionally manipulate their responses, the NIM, PIM and INC scales demonstrated sensitivity to group differences. Specifically, individuals who were instructed to fake good or produce a profile that would maximize their odds at being selected for graduate psychology admissions, had significantly higher PIM scores than did participants in the control group. However, the PIM scale was not sensitive to the police instructional set relative to the control condition. This suggests caution in using the PIM scale for use in the analysis of individual profiles. Moreover, the PIM scale may be not be able to detect more sophisticated forms of impression management.

The NIM scale was designed to detect an intentionally negative self-presentational response style, and preliminary results obtained in this study suggests that it was able to do so. Individuals in the fake bad condition produced profiles that had significantly higher NIM scores than any of the other conditions. However, since only a general “fake bad” instructional set was included in the validation study, it is unknown if more sophisticated faking bad could avoid detection by the NIM scale. Future research may consider using a condition such as instructing participants to maximize odds of not being selected by the military by producing a negative profile, but to avoid detection of lying.

Support for the INC scale was provided in Study 2. The INC scale was sensitive to random responding, but insensitive to the effects of all other instructional sets employed in this study. However, here too caution must be advised. The random profiles used in this study were computer generated. However, in clinical and applied settings, careless or “random” responding may not produce the same profiles as those generated by computers. In clinical or applied settings, random responding may consist of choosing the same option over and over, or the same pattern repetitively (0, 4, 0, 4). Additionally, random responding may itself be variable. An individual may lapse in and out of random responding. It is unknown how the INC scale would perform under those conditions, and future research should compare random responding generated by actual participants with those produced through computer generations.

The results suggest that the validity scales may perform best when interpreted as an whole. For instance, the PIM scale did not differentiate individuals in the police condition from those in the control group, but the NIM scale did so. This suggests that participants in the police condition were trying to manage their self-presentation by denying negative attributes rather
than endorsing positive ones. In the examination of more sophisticated forms of impression management, researchers may need to consider both endorsement of positive attributes and negation of negative aspects of self.

The issue of the significant correlations between the validity scales and the NEO-FFI domain scores deserves attention. These significant correlations do not necessarily invalidate the validity scales. First, the correlations obtained in the present study are consistent with the level of correlations obtained by other authors (Morey, 1991, Schinka et al., 1997). Moreover, it is not surprising that the validity scales are correlated to the NEO-FFI domain scales, as there are significant intercorrelations between the NEO-FFI domain scales themselves (Tokar & Swanson, 1995). Although some may argue that the significant correlations between the validity and domain scales suggests that the validity scales are simply a measure of common variance across the domains of the NEO, the results of Study 2 contradict this interpretation. If the validity scales were simply a measure of common variance, they would not have differentiated between the various faking groups.

The primary issue is not whether the validity scales and domain scales are correlated, but what is the psychological meaning of this correlation. At a more fundamental level, the question is what do validity scales measure and what contributions do validity scales offer to test interpretation. Costa and McCrae, the developers of the NEO-FFI, intentionally did not include validity scales on the NEO-FFI and NEO-PI-R because they believe that validity scales are of limited value. They consider social desirability a core aspect of personality and that its influence in test scores should not be removed. McCrae and Costa (1983) obtained results which suggested that controlling for the effects of social desirability actually reduced the correlations between self-report and spouse rating of personality. However, others suggest that “faking” can invalidate test performance (Topping & O’Gorman, 1997), and that the influence of social desirability should be removed, controlled, or at least recognized in the assessment situation (Ben-Porath & Waller, 1992; Dunette, McCartney, Carlson, & Kirchner, 1962).

Results obtained by Scandell and Wlazelek (1999) suggest that the differentiation between “faking” and “honest” responding may be blurred. These authors found that self-presentation on the NEO-FFI was related to both perceptions of environmental contingencies (i.e., instructions to fake) and self-perception of personality. Their results suggested that individuals use their personality as the baseline from which to fake. As such, even when “faking” on the NEO-FFI, respondents’ scores still reflected their personality.

Paulhus’s insights concerning the nature of social desirability may be useful in reconciling the divergent viewpoints concerning validity scales. In his original formulation, Paulhus (1984) suggested that social desirability is composed of two components: impression management and self-deception. In impression management, individuals deliberately attempt to manipulate their responses to create specific impressions. In self-deception, an unconscious process, individuals report and believe that they have traits which they do not. Paulhus considered impression management to be a greater threat to test validity than self-deception, as self-deception is a reflection of personality. Future research may examine how the Positive and Negative Impression Management scales developed in the present study are related to Paulhus’s constructs of impression management and self-deception, as measured by the Balanced Inventory of Desirable Responding (Paulhus, 1998). If the validity scales are found to be significantly related to impression management, but not to self-deception, this would suggest
that the validity scales are responsive to deliberate attempts on the part of respondents to self-present in particular ways. This finding would imply that test profiles with high validity scores might not be valid. However, if the validity scales are found to be related to self-deception but not impression management, this would suggest that the validity scales are assessing an aspect of personality. In this situation, high validity scores may not “invalidate” obtained scores, but they may be interpreted as a characterological style of the respondent.

Recently, Paulhus has reformulated his conceptualization of social desirability (Paulhus & John, 1998). He suggests that there are two types of social desirability, labelled egoist and moralistic bias. In egoistic bias, related to the Need for Power, individuals endorse positive trait descriptions. In moralistic bias, related to the Need for Approval, individuals deny having negative traits. This reformulation suggests that the Positive Impression Management scale may be related to egoistic bias, while scores on the Negative Impression management scale may be related to moralistic bias. These hypotheses await testing.

Future research should examine the operation of the PIM, NIM and INC scales in clinical and applied settings. At this point, the NEO-FFI validity scales developed here should not be used to identify individual profiles that may be biased by negative or positive self-presentation, or random responding. The scales would be more appropriately used as a research tool to examine the relationship between scores on the validity scales and theoretically relevant constructs such as impression management and self-deception or egoistic and moralistic bias. Additionally, future research may examine the relationship between the validity scales developed in this study and other validity scales such as those of the MMPI. Future research should seek to examine the psychological meaning of scores on the validity scales, and the value of the current validity scales may be in providing this opportunity.

References


