Behavioral Sciences & the Law
Behav. Sci. Law **30**: 16–27 (2012)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/bsl.1996

The Inconsistency of Inconsistency Scales: A Comparison of Two Widely Used Measures

Natalia L. Nikolova, M.A.*,†, Melissa C. Hendry, M.A.†, Kevin S. Douglas, LL.B., Ph.D.†, John F. Edens, Ph.D.‡ and Scott O. Lilienfeld, Ph.D.§

This study compared the inconsistent responding validity scales of the Personality Assessment Inventory (PAI) and the Psychopathic Personality Inventory (PPI)/PPI-Revised (PPI-R) in two correctional samples to determine the extent to which they overlap in identifying invalid profiles. Results revealed substantial differences in the way the inconsistent responding validity scales of these measures performed. In particular, the PAI identified far fewer participants as having responded inconsistently compared with the PPI/PPI-R. We discuss the implications of our findings for clinical practice, and potential concerns with the use of a single measure to identify inconsistent responding in clinical practice and research. Copyright © 2012 John Wiley & Sons, Ltd.

The utility of self-report instruments has sometimes been questioned (e.g., Baer & Miller, 2002), as their validity may be negatively impacted by such factors as inconsistent or indiscriminate responding, over- or under-reporting of symptoms, or social desirability. As a result, a number of researchers have incorporated scales that assess the validity of questionnaire protocols as a means of ascertaining the accuracy of the results (e.g., Arbisi & Ben-Porath, 1995). Piedmont, McCrae, Riemann, and Angleitner (2000), however, examined the validity of validity scales in two samples using the NEO Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992) and concluded that the scales "failed to enhance the validity of personality assessments" (p. 589). Specifically, most standard validity scales failed to act as suppressor variables as expected: controlling statistically for scores on these scales did not lead to higher correlations between personality measures and external criteria. Nevertheless, Piedmont et al. (2000) acknowledged that these results do not speak to the validity of validity scales in applied settings, such as clinical, forensic, or industrial/organizational samples, where the incentives for response distortion may be higher (see also Edens & Ruiz, 2006).

Following in the footsteps of Piedmont et al. (2000), McGrath, Mitchell, Kim, and Hough (2010) recently conducted an extensive review of published research examining the utility of validity scales and concluded that the "support for the use of bias indicators was weak" (p. 450). This assertion was based primarily on a study selection strategy that excluded all simulation designs and focused almost exclusively on studies examining either suppressor or moderator effects that were tested using multiple regression techniques.

*Correspondence to: Natalia L. Nikolova, M.A., Mental Health, Law and Policy Institute, Department of Psychology, Simon Fraser University, Burnaby, BC V5A 1S6 Canada. E-mail: nnikolov@sfu.ca
†Simon Fraser University, Burnaby, BC, Canada.
‡Texas A&M University, College Station, TX.
§Emory University, Atlanta, GA.

Using such a restrictive approach to identifying relevant research, it is perhaps not surprising that the authors identified only 41 studies for review across a wide range of contexts (e.g., general personality assessment, assessment of emotional disorders, workplace assessment, forensic assessment).

Notably, the one response style for which McGrath et al. (2010) concluded there was some supportive research evidence was in the assessment of non-systematic distortion (e.g., inattentive, inconsistent, confused, or random responding), indicating that validity coefficients were higher in a small number of studies in which such responding was statistically controlled. Note that, none of these studies were conducted with forensic or correctional samples.

In forensic and correctional settings, there are numerous motivations for feigning psychopathology (e.g., avoiding criminal sanctions by appearing insane) or minimizing socially deviant personality traits (e.g., appearing "rehabilitated" to increase the odds of early release). Non-systematic distortion also is a significant concern in such contexts, given that offenders: (a) typically have much lower reading abilities than the general population; and (b) may be required to complete personality inventories as part of inmate screening procedures for which they have little or no investment in the assessment process. For instance, Butcher et al. (2001) reported that across two general inmate samples, 15–21% of the individuals who completed the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) produced invalid profiles reflecting inattentive, inconsistent or random responding. Given the importance of accurate assessment, many multi-scale self-report personality instruments commonly used in forensic and correctional settings, such as the Personality Assessment Inventory (PAI; Morey, 1991), the MMPI-2, Psychopathic Personality Inventory (Lilienfeld & Andrews, 1996), and the Psychopathic Personality Inventory-Revised (PPI-R; Lilienfeld & Widows, 2005), incorporate an array of validity scales to assess various forms of systematic and non-systematic response distortion.

For example, both the PAI and PPI/PPI-R contain validity scales intended to measure impression management and inconsistent or variable responding. Some authors have argued that impression management (negative or positive) is often related to "valid personality variance" (e.g., McCrae & Costa, 1983; Uziel, 2010) in that it reveals psychological characteristics of individuals, especially personality dispositions (e.g., neuroticism, agreeableness). Hence, with the PAI and PPI/PPI-R, some authors argue that these scales should be used only sparingly to declare participants' responses uninterpretable (Edens & Ruiz, 2006; Lilienfeld & Fowler, 2006; Morey & Hopwood, 2007).

However, elevations on inconsistency response scales (the focus of the present research), arguably do invalidate test protocols, as they ostensibly reflect the introduction of non-systematic error variance into assessment results. Such variance is putatively irrelevant to the personality and psychopathology constructs the inventories are intended to assess. Such inconsistency scales typically are created by identifying item pairs that have similar (or opposing) content and are highly intercorrelated (e.g., $r > \pm$ 0.45; Lilienfeld & Widows, 2005), and by subsequently calculating the number of times participants responded to those item pairs inconsistently (Morey, 2007; Lilienfeld & Widows, 2005; Tellegen & Waller, 2008).

To date, few peer-reviewed studies have examined the validity of the inconsistency scales of the PAI and the PPI/PPI-R. Some research using the PAI Inconsistency (ICN) scale has shown that it can distinguish actual respondents from computer-generated profiles of random responding. For example, a cut-off *T*-score of 64 correctly identified 83.8% of

DOI: 10.1002/bsl

computer-generated random profiles, whereas only 16.4% of the clinical and 11.7% of the community/college student samples were above this score (Morey, 1991). Nevertheless, some correlational studies have revealed weak associations between the ICN scale and measures of putatively inconsistent responding embedded in other scales. For example, Morey (2007) reported minimal correlations between the ICN scale of the PAI and the inconsistency scales of the Trauma Symptom Inventory (TSI; Briere, 1995), the Structured Interview of Reported Symptoms (SIRS; Rogers, Bagby, & Dickens, 1992), and the NEO-PI-R (Costa & McCrae, 1992).

Furthermore, a few studies, described in the PPI-R professional manual (Lilienfeld & Widows, 2005), have examined the validity of the Inconsistent Responding 15 (IR15) and the Inconsistent Responding 40 (IR40) of the PPI-R. In one study, nine participants (i.e., 23%) were instructed to take the PPI-R under speeded conditions; participants in this condition produced higher IR15 and IR40 scores than participants in both the honest responding and the positive impression management conditions, but not the negative impression management condition. Another study demonstrated that, in an offender sample, the PAI ICN scale was related to both the IR15 ($r = 0.24$) and the IR40 ($r = 0.32$) scales. However, in the normative population, the ICN scale was associated with the IR40 scale ($r = 0.22$), but not with the IR15 scale (Lilienfeld & Widows, 2005). Finally, the results of one unpublished study reported in the PPI-R manual indicated that when using 200 randomly generated PPI-R protocols, most (72.5%) exceeded the cut-off with respect to protocol validity for IR15, and even more (78.5%) exceeded the IR40 cut-off, providing some evidence that these two scales can identify random, inconsistent, or careless protocols (Lilienfeld & Widows, 2005).

The focus of the present study is on the PAI and PPI/PPI-R validity scales that address inconsistent responding. We address a question of pragmatic importance to clinicians and researchers alike – do the corresponding scales across these instruments identify the same protocols as valid and invalid? This question is important in situations in which only one instrument with validity scales is used – does the assessment of non-systematic distortion on one instrument generalize to other self-report instruments?

In this study, we examined: (1) whether the validity scales of these measures identify the same inmates as producing invalid protocols on different instruments; and (2) whether one can thereby rely on only one measure to rule out cases that would have been judged invalid on the other measure. We examined this question by comparing the PAI and PPI in a large sample of offenders from the United States, and by comparing the PAI and PPI-R in a smaller Canadian sample of provincial inmates.

METHOD

Participants

Study 1

Participants were prison inmates or substance abusers enrolled in court-mandated treatment. They took part in a larger study (see Poythress, Lilienfeld, Skeem, Douglas, Edens, Epstein, & Patrick, 2010, for an overview) examining personality features and socially deviant behavior conducted in five U.S. states (Florida, Nevada, Oregon, Texas, and Utah). Eligibility criteria were as follows: (1) between the ages of 21 and 40 (although a small

DOI: 10.1002/bsl

number of younger and older participants were selected if prospective participants between 21 and 40 years old were not available on days of participant recruitment); (2) of African American (33.3%) or Caucasian (57%) descent (race data were missing for 2.9% of the participants, and 6.8% were of "other" descent); (3) able to communicate in English; and (4) not suffering from acute psychotic symptoms or mental retardation. PAI and PPI protocols were obtained for 1,741 individuals, most of whom were male (81.2%); gender data were missing for 1.7% of the participants. Of all participants, 52.3% were prison inmates and 47.7% were enrolled in a court-mandated substance abuse treatment program. The average age was 30.5 years (SD = 6.6; age data were missing for 5.1% of the participants), and majority of the participants had no high school diploma or equivalent (i.e., 30.4%).

Study 2

Participants were inmates from two correctional facilities in Western Canada who participated in a larger study validating a newly developed measure of psychopathy – the Comprehensive Assessment of Psychopathic Personality-Institutional Rating Scale (CAPP-IRS; Cooke, Hart, & Logan, 2005). Eligibility criteria were: (1) between the age of 19 and 50; (2) able to communicate in English; and (3) not suffering from acute psychotic symptoms (e.g., delusions, hallucinations, disorganized cognition or behavior). PAI and PPI-R protocols were obtained for 101 individuals (51 men and 50 women). Most participants were Caucasian (70.3%), followed by Aboriginal (16.8%), and their average age was 32.91 years (SD = 8.35). The majority of the participants (i.e., 56.4%) had no high school diploma or an equivalent.

Procedure

Study 1

As part of a larger study, individuals who met eligibility criteria were selected randomly and invited to participate. The recruitment process involved obtaining informed consent from participants followed by an enrollment interview. They were later administered IQ and reading screens. The PAI and the PPI, along with other measures, were administered over the course of two sessions, which lasted (on average) a total of 4.5 hours. The PAI was administered as a paper-and-pencil measure, whereas the PPI was entered directly into a software program. The PAI was always administered in session one, whereas the PPI was administered in session two. Each of those measures was the first to be administered in their respective sessions. For the purposes of this study, only the validity scales of the PAI and the PPI were used in data analyses. Scale scores on the PPI were not computed if 20% or more of the items on that scale were left unanswered or had multiple responses. The PAI was scored using the official PAI scoring software from the publisher, Psychological Assessment Resources, Inc.

Study 2

The PAI and the PPI-R were administered as part of a validation study of the CAPP-IRS. Following recruitment, participants were interviewed briefly to obtain information about life and family history, education, mental health, and current functioning. They were then asked to complete the PAI and the PPI-R, along with other self-report measures, as a way of

DOI: 10.1002/bsl

obtaining further information about their functioning, personality traits, and psychopathology. The order of administration of the self-report measures was counterbalanced, such that for half of the sample, the PAI and PPI-R were administered as the penultimate and final measures respectively, whereas for the other half they were administered in reverse order and as the first and second measures of a battery of tests. Both instruments were administered as paper-and-pencil measures. For the purposes of this study, only the validity scales of the PAI and the PPI-R were used in analyses. As in Study 1, if 20% or more of the items on a given PPI-R scale were left unanswered or had multiple responses, the scale score was not calculated and was therefore excluded from analyses. In addition, PPI-R profiles were considered invalid if they had more than 30 invalid items in total (i.e., items with missing or multiple responses). If PAI scales had more than 20% missing items, they were not scored.

Measures

Personality Assessment Inventory

The PAI (Morey, 1991) is a 344-item instrument designed to assess personality and psychopathology, as well as other domains of clinical concern (e.g., stress, motivation for change). It consists of 22 scales with non-overlapping item content, including four primary validity scales. Items on the PAI are scored on a four-point scale (i.e., "false, not at all true," "slightly true," "mainly true," and "very true"). An estimated grade 4 reading level is required to complete the measure. Research provides support for the reliability and validity of the PAI (e.g., Boyle & Kavan, 1995; Douglas, Hart & Kropp, 2001; Morey, 2000, 2007), as well as its convergent and discriminant validity with respect to the MMPI-2. The PAI is widely used in forensic and correctional settings for both research and clinical purposes (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Edens & Ruiz, 2005; Mullen & Edens, 2008).

The four PAI validity scales are Inconsistency (ICN), Infrequency (INF), Negative Impression (NIM), and Positive Impression (PIM). The ICN scale – the focus of the present research – comprises 10 item pairs and reflects the extent to which participants are consistent in answering questions with similar or opposing content (i.e., for half of the item pairs, items should be endorsed in the same direction, and for the other half they should be endorsed in opposite directions). The item pairs included in this scale were those most empirically interrelated at the time of development of the measure. Even though the items within each pair cover similar content (or reversely scored opposing content), the content differs from one pair to another (e.g., one item pair reflects anxiety, one reflects drug use). As a result the scale does not capture any specific construct (i.e., anxiety or drug use) other than response inconsistency within each pair. *T*-scores of 73 or higher indicate that responses within item pairs were highly inconsistent (and occurred in less than 4% of both the community and clinical normative samples), and therefore such profiles should be considered invalid (Morey, 1991). As a result, a cut-score of 73 *T* was used in this study and compared to a cut-score of 80 *T* as proposed by Edens & Ruiz (2005) for use with correctional samples.

Psychopathic Personality Inventory (PPI)

The PPI (Lilienfeld & Andrews, 1996) is a 187-item self-report measure intended to assess traits of psychopathic personality. Items are scored on a four-point Likert scale ("false,"

DOI: 10.1002/bsl

"mostly false," "mostly true," "true"). Although not developed specifically for use with forensic and correctional populations, a considerable amount of construct validation research on this measure has relied on offender samples (e.g., Poythress et al., 2010). The PPI yields a total score representing global psychopathy, and eight subscale scores measuring various facets of this personality pattern (e.g., Fearlessness, Social Potency, Egocentricity). PPI total scores exhibit high internal consistency, test-retest reliability, convergent validity with other measures of psychopathy and discriminant validity from measures of depression, negative emotionality, and schizotypy (Benning, Patrick, Hicks, Blonigen, & Krueger, 2003; Blonigen, Carlson, Krueger, & Patrick, 2003; Lilienfeld & Andrews, 1996; Poythress et al., 2010).

The PPI contains three validity scales designed to assess three different response styles. The Unlikely Virtues (UV; 14 items) and Deviant Responding (DR; 10 items) scales are used as indicators of social desirability/positive impression management, and malingering/"faking bad," respectively. The Variable Response Inconsistency scale (VRIN; 40 item pairs with similar content) was modeled after the Variable Response Inconsistency scale of the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008). It is intended to reflect random or careless responding and therefore helps to determine whether a PPI profile can be interpreted. The VRIN scale contains item pairs that were found to be highly correlated ($r \geq 0.30$) during the development of the instrument. For the purposes of this study we used a cut-score of 39 on VRIN to identify inconsistent responding, as this is the cut-score recommended in the PPI-R manual, and the PPI has the same number of item pairs comprising this validity scale as does the PPI-R.

Psychopathic Personality Inventory-Revised

The PPI-R (Lilienfeld & Widows, 2005) is the revised version of the PPI. It was updated to lower the required reading level to grade 4 and to eliminate culturally specific items and items with poor psychometric properties. The current version contains 154 items, and yields a total score, as well as eight content and three factor scores. There is evidence for the convergent and discriminant validity of the PPI-R with respect to Levenson's Self-Report Psychopathy Scale (LSRP; Levenson, Kiehl, & Fitzpatrick, 1995), and the Self-Report Psychopathy Scale-II (SPR-II, Hare, 1985). Support for the construct validity of the PPI-R is also evident in associations between this scale and theoretically relevant subscales from measures such as the PAI (e.g., Antisocial Features, Aggression, and Dominance) and the NEO- Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992; e.g., Negative Relationships with Agreeableness and Conscientiousness; DeMauro & Leung, 2005; Lilienfeld & Widows, 2005) and the Brief Psychiatric Rating Scale (Edens & McDermott, 2010).

Like the PPI, the PPI-R contains three validity scales: Virtuous Responding (VR; 13 items), Deviant Responding (DR; 10 items), and Inconsistent Responding scales (IR; which has 15- or 40-item pair versions), referred to as VRIN on the PPI-R's predecessor, the PPI. Scores of 15 or 16 on IR15 were obtained in less than 5% of the normative sample, and raise concerns about the validity of those profiles. Similarly, scores between 39 and 44 on IR40 were obtained in less than 5% of the normative sample, and indicate questionable profile validity (Lilienfeld & Widows, 2005). With this in mind, for the purposes of analyses reported here, we used cut-scores of 15 and 39 on the IR15 and IR40, respectively.

DOI: 10.1002/bsl

Two normative samples were used to provide preliminary validation for the PPI-R: community/college ($N = 985$) and offender samples ($N = 154$). In addition, data from these samples provided support for the reliability of the measure. For instance, Lilienfeld and Widows (2005) reported satisfactory internal consistencies for the total score and content scales – α ranging from 0.78 to 0.92 in the community/college sample and from 0.71 to 0.84 in the offender sample. Further, test-retest correlations for the total and factor scores as well as content scales of the PPI-R based on a subset of the normative sample ranged from 0.82 to 0.95 over the course of 19.94 days on average (range 12–45 days; $n = 51$).

Data Analyses

For both samples, two types of analysis were used to determine whether the validity scales of the PAI and the PPI/PPI-R reflecting inconsistent or random responding (i.e., ICN and VRIN/IR respectively) perform comparably. First, we identified PAI and PPI/PPI-R profiles for which the inconsistent responding scale scores were above recommended cut-offs, and evaluated the extent to which the identified profiles overlapped across measures. For the PAI, we used *T* scores on ICN above two recommended cut-offs (i.e., 73 *T* as recommended by the PAI manual and 80 *T* for correctional samples as suggested by Edens & Ruiz, 2005). We also examined the percentage of PAI versus PPI/PPI-R profiles for which the scores on the inconsistent responding scales fell above the recommended cut-offs. Further, we computed zero-order correlations between the VRIN/IR and ICN scales. Finally, chi-squared analyses were conducted to examine the level of agreement between the VRIN/IR and ICN scale cut-off scores.

RESULTS

Study 1

A limited number of questionnaires filled out by the 1,741 participants in this study could not be scored due to excessive missing data as defined in the Method section. As a result, 1,642 (94.3%) PAIs and 1,607 (92.3%) PPIs were successfully scored. Of them, based solely on the ICN scale ($T = 73$ cut-score) of the PAI, 1,589 (97%) of participants produced valid profiles. By comparison, 1,342 (84%) of the participants produced a valid profile based on the VRIN scale of the PPI. When the PAI and the PPI were used in conjunction, 1,264 (i.e., 84%) of the participants produced valid profiles. Specifically, although the PPI identified 265 invalid cases (i.e., 16%), and the PAI identified 53 invalid cases (i.e., $< 3\%$), only 14 of those cases were identified as invalid by both measures. Finally, there was a significant, but small, correlation between the ICN and VRIN scales ($r = 0.185$, $p < 0.001$). Table 1 presents a 2×2 table of the number of individuals who exhibited valid and invalid profiles on the PAI and PPI.

Although this information is based on PAI cut-scores developed for community samples, for purposes of comparison we also used correctional sample cut-scores suggested for the PAI ICN scale (i.e., 80 *T*; see Edens & Ruiz, 2005). Based on this criterion, there were seven invalid PAI profiles, only one of which was also identified as invalid by the PPI.

Supplemental analyses were conducted to determine a lower ICN *T*-score, at which the PAI and the PPI identified a similar percentage of invalid profiles. Results revealed ICN scores of 61 *T* or higher in 21% of the profiles, and scores of 62 *T* or higher in 13% of the

DOI: 10.1002/bsl

Table 1. Correspondence between the validity scales of the Personality Assessment Inventory (PAI) and the Psychopathic Personality Inventory (PPI) (Sample 1)

	PPI (VRI)		
Measure	Valid	Invalid	Total
PAI (ICN)			
Valid	1264	240	1504
Invalid	35	14	49
Total	1299	254	1553

profiles, which was close to the percentage of invalid profiles identified by the PPI. Nevertheless, the overlap between those profiles was only 22% at 61 *T* or higher, and 25% at 62 *T* or higher, indicating that even when the PAI and PPI identify roughly the same proportion of profiles as invalid, they identify largely different groups of individuals as having responded inconsistently.

Most individuals who produced invalid profiles on either the PAI or the PPI were male (82.2%), African American (50.7%), and had no high school diploma (42.8%). Their mean age was 30.26 (SD = 6.78). There was a significant difference in ethnicity and education between participants who produced valid and invalid profiles. African American participants were more likely than others to produce invalid profiles, χ^2 (4, 1565) = 59.56, $p < 0.001$, as were participants without a high school diploma, χ^2 (5, 1564) = 26.425, $p < 0.001$. These findings were partially consistent with the overall demographic characteristics of the sample, majority of which consisted of Caucasian males without a high school diploma.

Study 2

In Sample 2, all 101 participants completed the PAI and PPI-R, although one PAI protocol could not be scored due to excessive missing data, as defined in the Method section. Similar to Sample 1, of the 100 participants whose profiles contained enough responses to be scored, 98 (98%) produced valid profiles based solely on the ICN scale of the PAI using 73 *T* as the cut-off. Also similar to Sample 1, the PPI-R identified a greater proportion of this sample as invalid. Based on the IR scales of the PPI-R, of the 101 participants, 83 (82%) produced valid profiles on IR15, and 76 (75%) participants produced valid profiles on the IR40. When the invalid response rules for the PAI (ICN scale) and the PPI-R (IR40 scale) were used in conjunction, 74 participants (74%) in total produced valid profiles. In terms of identifying inconsistent or random responding based on the ICN and IR40 scales used in conjunction, 25 participants (25%) were identified by IR40, only one of whom was also identified by ICN. The ICN alone identified two invalid profiles in total. The results were similar when IR15 was used in lieu of IR40: even though the IR15 alone identified 18 participants (18%) who responded inconsistently, only one was identified when ICN and IR15 were used in conjunction. Finally, the correlations between the ICN, on the one hand, and the IR15 and IR40 scales, on the other, were weak and nonsignificant (i.e., $r = 0.079$ and $r = -0.006$ respectively). Given that the IR40 scale contains all of the IR15 items, those two scales were significantly correlated, as expected (i.e., $r = 0.829$, $p \leq 0.001$). Tables 2 and 3 present the number of individuals who produced valid and invalid profiles based on the PAI (ICN) and PPI-R (IR15 and IR40) in this sample.

DOI: 10.1002/bsl

Table 2. Correspondence between the validity scales of the Personality Assessment Inventory (PAI) and the Psychopathic Personality Inventory-Revised (PPI-R) (IR15) (Sample 2)

	PPI-R (IR15)		
Measure	Valid	Invalid	Total
PAI (ICN)			
Valid	81	17	98
Invalid	1	1	2
Total	82	18	100

When the cut-score for the ICN scale of the PAI for correctional samples (i.e., 80 *T*) was used, there was only one invalid PAI profile, which was also identified as invalid by the PPI-R. Further, similar to Study 1, supplemental analyses were performed to determine a lower ICN *T*-score, at which the PPI-R and the PAI identify a similar percentage of invalid profiles. Results revealed that 24% of the profiles obtained an ICN score of 60 *T* or higher, which was comparable to the percentage of invalid profiles identified by the IR40. Of note, even when cut-scores were manipulated to identify a comparable proportion of invalid protocols across instruments, there was only a 32% overlap between the IR40 and the ICN scales in terms of identifying inconsistent responders.

Individuals who produced invalid profiles on either the PAI or the PPI-R (IR40) were equally split by gender, and most were Caucasian (65.4%), and had no high school diploma (46.2%). Their mean age was 31.31 years (SD = 8.36). These characteristics were reflective of the overall demographic characteristic of the sample. With respect to invalid profiles on either the PAI or the PPI-R (IR15), the results were somewhat different: here most invalid scorers were male (i.e., 63.2%), 68.4% were Caucasian, their mean age was 33.47 years (SD = 8.57), and individuals with high school diplomas or an equivalent (i.e., 42.1%) slightly outnumbered those without (i.e., 36.8%). Nevertheless, given the small number of participants who fell into this group (i.e., 19), these results should be considered preliminary. In both instances (i.e., while using IR40 or IR15 in the analyses), however, individuals who produced invalid profiles did not differ significantly from those who produced valid profiles in terms of demographic characteristics.

DISCUSSION

To date, almost no published research has examined the "consistency" of inconsistency scales embedded in multi-scale self-report personality inventories in terms of their convergence in

Table 3. Correspondence between the validity scales of the Personality Assessment Inventory (PAI) and the Psychopathic Personality Inventory-Revised (PPI-R) (IR40) (Sample 2)

	PPI-R (IR40)		
Measure	Valid	Invalid	Total
PAI (ICN)			
Valid	74	24	98
Invalid	1	1	2
Total	75	25	100

DOI: 10.1002/bsl

identifying protocols deemed invalid because of non-systematic response distortion. In this study, we aimed to determine whether the PAI and PPI/PPI-R identify the same individuals as producing invalid profiles. Results from both samples demonstrated that the PPI and PPI-R identified a substantially and significantly higher number of invalid profiles than did the PAI. In addition, there was surprisingly little correspondence between the inconsistency scales of the PAI and those of either edition of the PPI: only 14 of all invalid profiles were identified as invalid in Sample 1 by both the PAI and the PPI (i.e., $< 3\%$), and only one person in Sample 2 was identified as responding inconsistently by both the PAI and PPI-R (i.e., $< 1\%$).

Our analyses also revealed that although the demographic make-up of those who produced invalid profiles was reflective of the overall demographic make-up of the respective sample, there were cross-sample differences. Specifically, there were no demographic differences in valid and invalid respondents in Sample 2; in Sample 1 individuals of African American descent and less educated individuals produced a higher number of invalid profiles. Given African Americans were over-represented among the group of invalid responders, further research is needed to replicate these demographic differences and to ascertain whether they may be attributable to other variables, such as educational or reading level.

The inconsistency across our inconsistency scales may stem from various sources. Idiosyncratic factors, such as participants being inconsistent in providing random/inconsistent responses over time or across measures, may have played a role. Although participants who respond randomly to questionnaires presumably do so across most or all measures, it appears that they provided such responses more on one of these measures (i.e., the PAI or PPI/PPI-R, but not both). In fact, there is reason to be concerned that the level of carelessness may not even be consistent within one test, as evidenced by the development of various types of "back random responding" scales, such as the MMPI-2 F back [F(b)] scale (see MMPI-2, Butcher et al., 1989; see also Morey & Hopwood, 2004). Further, fatigue effects could have adversely affected the validity of profiles in Sample 2; given that the PAI and PPI-R were counter-balanced; however, we were unable to investigate this possibility. That methodological feature was not present in Sample 1, in which the order of administration of the PAI and PPI (both of which were the first measures to be administered among a battery of tests in their respective sessions) was used to control for fatigue effects.

Our results suggest that the recommended cut-scores for the PPI/PPI-R have a considerably lower threshold than the PAI for detecting inconsistent response sets or styles among offender samples. This finding could reflect the difference in the length of the inconsistent responding scales on the PAI and the PPI-R. Because the validity scales on the PPI/PPI-R are four times as long as the corresponding PAI scale, the ICN scale may not be long enough to yield a stable assessment of random responding. Nevertheless, in Sample 2 the PPI-R identified more invalid profiles than the PAI even when the shorter of the two inconsistent responding scales (i.e., 15-item scale) on the PPI-R was used. It is also possible that the PPI/PPI-R tend to over-identify inconsistent response styles, or that the PAI tends to under-identify such styles. Alternately, it may the case the neither scale is more "accurate" in its detection of inconsistent responding, but that these measures differ substantially in their tolerance for their variability in responses to similar items. In other words, the meaning of items included in item pairs on one measure may be more ambiguous than those included on the other, therefore allowing more room for interpretation of their meaning. This could in turn lead to higher rates of inconsistent responses without such inconsistency being intentional. Alternately, because the item content of the PPI and PPI-R

DOI: 10.1002/bsl

item pairs is narrower in scope (i.e., all items relate to psychopathic traits) than the content of the PAI item pairs, the PAI may provide better breadth in terms of content domains and be less likely to be distorted by inconsistent responding specific to one domain of personality functioning (e.g.,. psychopathy).

Our findings hold potential implications for research and clinical practice in correctional settings. First and foremost, one cannot presume that an examinee who produced an ostensibly invalid profile on one measure because of inconsistency necessarily engaged in a similar degree of distortion on other measures. Our findings suggest a marked inconsistency between the inconsistency scales of two commonly used measures, and raise the question of whether one would find similar inconsistency across other widely used inconsistency scales, such as the Variable Response Inconsistency Scales of the MMPI-2 (Butcher et al., 1989) or MPQ (Tellegen & Waller, 2008). Our findings also raise the intriguing question of whether these differences reflect simply a difference between those measures resulting from item selection for their respective validity scales, or whether the validity of their validity scales is questionable. Clearly, the reasons for these differences will require further investigation, including an examination of differences in item content and thresholds for ascertaining invalid profiles.

Our two-part study was marked by several limitations. Because individuals who exhibited acute psychotic symptoms or mental retardation were excluded from participation, the range of responses on the validity scales may have been somewhat reduced. However, based on the distribution of scores, that exclusion did not appear to exert a significant impact on the overall outcome of scores. Further, given that this is one of a very few studies that has reported on the validity of validity scales of self-report questionnaires, it remains to be seen whether the results can be generalized to different samples, settings, and measures. Overall, the PAI and PPI/PPI-R displayed little convergence in whom they identified as responding inconsistently, an unanticipated finding that raises significant concerns in applied and research settings. We hope that our preliminary investigation encourages other researchers to examine the extent to which our findings extend to validity indicators embedded in other widely used measures of personality and psychopathology.

REFERENCES

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., Buffington-Vollum, J., Stredny, R., & Handel, R. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 84–94.

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

Benning, S. D., Patrick, C. J., Hicks, B. M., Blonigen, D. M., & Krueger, R. F. (2003). Factor structure of the Psychopathic Personality Inventory: Validity and implications for clinical assessment. *Psychological Assessment*, *15(3)*, 340–350.

Blonigen, D. M., Carlson, S. R., Krueger, R. F., & Patrick, C. J. (2003). A twin study of self-reported psychopathic personality traits. *Personality and Individual Differences*, *35*, 179–197.

Boyle, G., & Kavan, M. (1995). Review of the Personality Assessment Inventory. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* [Electronic version]. Retrieved May 20, 2006, from EBSCO Mental Measurements Yearbook database.

Briere, J. (1995). *TSI: Trauma Symptom Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

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

DOI: 10.1002/bsl

Butcher, J., Graham, J., Ben-Porath, Y., Tellegen, A., Dahlstrom, W., & Kaemmer, B. (2001). *MMPI-2 manual for administration, scoring and interpretation* (revised ed.). Minneapolis: University of Minnesota Press.

Cooke, D., Hart, S., & Logan, C. (2005). Comprehensive Assessment of Psychopathic Personality Disorder: Institutional Rating Scale, Version 1.1. [Unpublished manual].

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

DeMauro, G. E., & Leung, S. A. (2005). Review of the Psychopathic Personality Inventory-Revised. In K. F. Geisinger, R. A. Spies, J. F. Carlson, and B. S. Plake (Eds.), *The seventeenth mental measurements yearbook*. [Electronic version]. Retrieved July 28, 2009, from EBSCO Mental Measurements Yearbook database.

Douglas, K. S., Hart, S. D., & Kropp, P. R. (2001). Validity of the Personality Assessment Inventory for forensic assessments. *International Journal of Offender Therapy and Comparative Criminology*, *45*, 183–197.

Edens, J. F., & McDermott, B. E. (2010). Examining the construct validity of the Psychopathic Personality Inventory–Revised: Preferential correlates of Fearless Dominance and Self-Centered Impulsivity. *Psychological Assessment*, *22*, 32–42.

Edens, J. F., & Ruiz, M. A. (2005). *PAI interpretive report for correctional settings (PAI-CS) – Professional manual*. Lutz, FL: PAR.

Edens, J. F., & Ruiz, M. A. (2006). On the validity of validity scales: The importance of defensive responding in the prediction of institutional misconduct. *Psychological Assessment*, *18*, 220–224.

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, *53*, 7–16.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a non-institutionalized population. *Journal of Personality and Social Psychology*, *68*, 151–158.

Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment*, *66*, 488–524.

Lilienfeld, S., & Fowler, K. (2006). The self-report assessment of psychopathy: Problems, pitfalls, and promises. In C. J. Patrick (Ed.), *Handbook of the psychopathy* (pp. 107–132). New York, NY US: Guilford Press.

Lilienfeld, S. O., & Widows, M. R. (2005). *The Psychopathic Personality Inventory-Revised (PPI-R) professional manual*. Lutz, FL: PAR.

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

McGrath, R., Mitchell, M., Kim, B., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*, 450–470.

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

Morey, L. C. (2000). The challenge of construct validity in the assessment of psychopathology. In R. D. Goffin & E. Edward (Eds), *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy* (pp. 357–388). New York, NY: Kluwer Academic/Plenum Publishers.

Morey, L. C. (2007). *The Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Morey, L. C., & Hopwood, C. (2004). Efficiency of a strategy for detecting back random responding on the Personality Assessment Inventory. *Psychological Assessment*, *16*, 197–200.

Morey, L. C., & Hopwood, C. (2007). *Casebook for the Personality Assessment Inventory (PAI): A structural summary approach*. Lutz, FL: Psychological Assessment Resources.

Mullen, K. L., & Edens, J. F. (2008). A case law survey of the Personality Assessment Inventory: Examining its role in civil and criminal trials. *Journal of Personality Assessment*, *90*, 300–303.

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

Poythress, N. G., Lilienfeld, S. O., Skeem, J. L., Douglas, K. S., Edens, J. F., Epstein, M., & Patrick, C. J. (2010). Using the PCL-R to help estimate the validity of two self-report measures of psychopathy with offenders. *Assessment*, *17*, 206–219.

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms professional manual*. Odessa, FL: Psychological Assessment Resources.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews & D. H. Sakolfie (Eds.), *The Sage handbook of personality theory and assessment (Vol. 2): Personality measurement and testing* (pp. 261–292). London: Sage.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*, 243–262.

DOI: 10.1002/bsl