Assessment
1–9
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191116641507
asm.sagepub.com
SAGE

Identifying Careless Responding With the Psychopathic Personality Inventory–Revised Validity Scales

David K. Marcus[1], Abere Sawaqdeh Church[1], Debra O'Connell[1], and Scott O. Lilienfeld[2]

Abstract

The Psychopathic Personality Inventory–Revised (PPI-R) includes validity scales that assess Deviant Responding (DR), Virtuous Responding, and Inconsistent Responding. We examined the utility of these scales for identifying careless responding using data from two online studies that examined correlates of psychopathy in college students (Sample 1: *N* = 583; Sample 2: *N* = 454). Compared with those below the cut scores, those above the cut on the DR scale yielded consistently lower validity coefficients when PPI-R scores were correlated with corresponding scales from the Triarchic Psychopathy Measure. The other three PPI-R validity scales yielded weaker and less consistent results. Participants who completed the studies in an inordinately brief amount of time scored significantly higher on the DR and Virtuous Responding scales than other participants. Based on the findings from the current studies, researchers collecting PPI-R data online should consider identifying and perhaps screening out respondents with elevated scores on the DR scale.

Keywords

psychopathy, Psychopathic Personality Inventory–Revised, validity scales, response bias, careless responding

Response bias can contribute to error variance in psychological assessment (Ben-Porath, 2003). Depending on the circumstances, individuals may respond in a manner that exaggerates their strengths and minimizes their flaws (positive impression management) or that exaggerates their pathology (negative impression management), or they may respond carelessly or randomly. Consequently, many psychological instruments include validity scales in an attempt to detect response bias. In research settings, careless and rushed responding may be a greater threat to test validity than is impression management, especially when responses are anonymous or confidential and respondents' primary motivations are to receive course credit or payment. This threat may be exacerbated in online data collection, which is typically unmonitored by research assistants. The current study examined the utility of the Psychopathic Personality Inventory–Revised (PPI-R; Lilienfeld & Widows, 2005) validity scales for detecting careless responding in online research settings.

One method for evaluating whether validity scales can effectively identify careless responding is to examine whether using scores on the validity scale to screen cases increases the criterion validity of substantive/content scales from the assessment instrument (McGrath, Mitchell, Kim, & Hough, 2010). A validity scale may provide useful information because it either *moderates* the association between the substantive scale and the criterion or *suppresses* scores on the substantive scale.[1] In the case of moderation, the association between the substantive scale and the criterion should decrease as scores on the validity scale increase, and scores above a cut point on the validity scale may indicate that scores on the substantive scale are uninterpretable. Thus, moderation is likely when a validity scale detects careless responding (the focus of the current studies), because someone responding carelessly on the validity scale is likely to also respond carelessly on the substantive scale, attenuating the associations between the substantive scale and criterion measures.

Unlike most other commonly used self-report measures of psychopathy, such as the Levenson Self-Report Psychopathy Scale (Levenson, Kiehl, & Fitzpatrick, 1995) and the Self-Report Psychopathy Scale–III (Paulhus, Neumann, & Hare, in press), the PPI-R (Lilienfeld & Widows, 2005) includes validity scales designed to detect response bias. The Deviant Responding (DR) scale aims to

[1]Washington State University, Pullman, WA, USA
[2]Emory University, Atlanta, GA, USA

Corresponding Author:
David K. Marcus, Department of Psychology, Washington State University, PO Box 644820, Pullman, WA 99164-4820, USA.
Email: david.marcus@wsu.edu

assess overreporting, malingering, carelessness, or attempts to sabotage the test. This scale is composed of bizarre or highly unlikely symptoms and behaviors that may superficially resemble signs or symptoms of psychopathology (e.g., "I sometimes forget my name"). The Virtuous Responding (VR) scale assesses underreporting or positive impression management (e.g., "I have never wished harm on someone else"). The Inconsistent Responding (IR) scales (the IR-15 is a 15-item short form and the IR-40 is a 40-item long form) are composed of pairs of items that are moderately correlated (e.g., "Sometimes I do dangerous things on a dare" with "I am a daredevil"). Highly inconsistent responses within these pairs of items suggest that the respondent was careless, was responding randomly, or did not understand the questions. Although only the IR scales were explicitly developed to identify inconsistent or careless responding, all of these PPI-R validity scales may potentially detect careless responding. Thus, careless or random responders may elevate the DR scale not because they are trying to appear pathological but because they are more likely to provide affirmative answers to implausible items than are respondents who have carefully read the items; a similar process may occur for the VR scale, which consists of items assessing extremely unlikely personal virtues. Although the items that comprise these validity scales appear to be face-valid and the PPI-R manual provides rough guidelines for how to interpret high scores on the DR, VR, and IR scales, the research examining their validity for detecting aberrant response styles or sets has been limited.

Two studies used a within-subjects simulation design with college samples to examine the validity scales of the original version of the PPI (Lilienfeld & Andrews, 1996). The DR scale demonstrated excellent diagnostic efficiency (area under the curve = .98) for distinguishing when participants feigned a psychosis compared with when they responded honestly (Edens, Buffington, & Tomicic, 2000). Edens, Buffington, Tomicic, and Riley (2001) found that respondents scored higher on the Unlikely Virtues (UV) scale (a precursor to the PPI-R VR scale derived from the Multidimensional Personality Questionnaire; Tellegen, in press) when they were instructed to "fake good" than when they responded under standard conditions. However, using the UV scale to distinguish feigned from standard protocols resulted in numerous misclassifications. In a between-subjects simulation study that included both a student sample and a forensic psychiatric sample, respondents instructed to overreport symptoms scored higher on the DR scale of the PPI-R than did those instructed to respond honestly, and those instructed to underreport symptoms scored higher on the VR scale than did those responding honestly (Anderson, Sellbom, Wygant, & Edens, 2013). Consistent with the moderation approach to assessing validity scales, they also found that the associations between the substantive PPI-R scales and related measures of psychopathy were lower in the overreporting and underreporting groups than in the honest groups. In a study of offenders who were not instructed to simulate pathology, Watts et al. (2016) failed to find consistent evidence that a set of validity indicators that included the PPI UV and DR scales moderated the associations between self-report psychopathy scores and non–self-report correlates relevant to psychopathy.

Nikolova, Hendry, Douglas, Edens, and Lilienfeld (2012) found little correspondence between the inconsistent protocols identified by the Inconsistency (ICN) scale of the Personality Assessment Inventory (Morey, 2007) and the scales that assess IR on the PPI (Variable Response Inconsistency scale) or the PPI-R (IR-15 and IR-40) in data from two correctional samples. Furthermore, the correlations between the ICN scale and each of the IR scales were very low and not statistically significant. A small study reported in the PPI-R manual, however, provides provisional support for the PPI-R IR scales. Nine students who were instructed to respond extremely quickly to the PPI-R scored significantly higher on the IR-15 and IR-40 scales than 11 students instructed to respond normally (Lilienfeld & Widows, 2005). However, the average IR scores in the speeded group (11 on the IR-15 and 33 on the IR-40) were still below the cut scores recommended by the manual (15 and 39, respectively). This finding suggests that the IR scales may be influenced by careless responding but may yield false negatives when attempting to identify invalid protocols.

Overall, the studies of the PPI/PPI-R validity scales suggest that the DR scale is capable of detecting simulated symptom exaggeration and that the VR scale may (perhaps to a lesser degree) be capable of detecting simulated positive impression management. However, these scales may not moderate the associations between substantive scales and external correlates when respondents do not have an incentive to provide deceptive responses to items or to complete questionnaires quickly. Less attention has been devoted to detecting careless responding, with only one very small, non–peer-reviewed study finding that students instructed to rush scored higher on the IR scale than those in the standard condition. No simulation studies involving the PPI/PPI-R validity scales have instructed participants to feign careless responding.

The Current Study

When psychopathy measures are used as part of a forensic evaluation, there are likely to be concerns about positive or negative impression management because respondents may have an incentive to underreport psychopathic personality traits or perhaps feign mental illness by overreporting other forms of psychopathology. In contrast, research participants typically have little incentive to engage in impression management, but they may be inclined to complete the questionnaire quickly to receive payment or course credit, which makes careless responding a potential threat to test validity. For example, careless responding is likely to elevate total PPI-R scores. The mean total score for young adults on the PPI-R ranges from roughly 275 to 300 (Lilienfeld & Widows, 2005), but because the PPI-R total score (which does not include the VR and DR scale items) consists of 131 items answered on a scale of 1 to 4, the mean score for a randomly completed protocol would be 327.5 (131 × 2.5). Thus, random responding would be especially problematic in studies that use the PPI-R to identify high-psychopathy individuals.

Although the PPI-R was developed for use with both college/community samples and forensic samples, most of the published research using the PPI-R has actually been conducted with college/community samples. Specifically, a PsycInfo search of studies that used the PPI-R yielded 69 articles that reported the findings from 77 independent samples. Only 14 of these 77 samples were forensic (48 were college and 15 were community). Furthermore, PPI-R data from 22 of the college samples and 2 of the community samples were collected online. The rate of careless responding may be especially high when questionnaire responses are collected online for course credit. For example, Meade and Craig (2012) found that at least 10% of undergraduates in an online study for course credit were deemed to be careless responders. Yet only 10 of the 24 PPI-R studies that collected data online from college students used the PPI-R validity scales to assess whether the PPI-R protocols were valid, perhaps because of the limited research on the utility of the scales for detecting careless responding and the lack of clear recommendations for appropriate cut scores. Given that studies using the PPI-R with college and community sample substantially contribute to the research literature on psychopathy and that these studies are increasingly being conducted online, knowledge regarding if and how the PPI-R validity scales can detect careless responders should enhance the quality of this research.

The current study used two methods to ascertain the ability of the PPI-R validity scales to identify response bias in two samples of college students who received course credit for participating. First, we examined whether scores on the PPI-R validity scales moderated the association between the PPI-R substantive scales and corresponding scales from the Triarchic Psychopathy Measure (TriPM; Patrick, 2010), which was designed to assess broadly comparable constructs as those in the PPI-R. Second, extending the approach used by Lilienfeld and Widows (2005), we examined the relations between time to completion and scores on the validity scales. However, unlike Lilienfeld and Widows, rather than manipulating the speed at which students completed the PPI-R, we examined whether the PPI-R validity scales discriminated between participants who completed the study in an inordinately brief period of time and those who completed the study in a reasonable amount of time (Meade & Craig, 2012).

Method

Participants

The participants were students at a public university in the Northwest United States who participated in two online studies examining the external correlates of psychopathic personality traits (Sample 1: $N = 583$; Sample 2: $N = 454$). Participants were recruited from undergraduate psychology courses and received research credits for participating. They were required to be fluent in English. In both studies, the participants ranged in age from 18 to 49 years (Sample 1: $M = 20.21$, $SD = 3.68$; Sample 2: $M = 20.14$, $SD = 3.58$). Women comprised the majority of the participants in both Sample 1 (73.8%) and Sample 2 (70.3%). Most participants reported being non-Hispanic Caucasian (73.4% and 72.7% in Sample 1 and 2, respectively), followed by Asian (8.9%, 8.4%), Hispanic (7.4%, 8.8%), and African American (4.8%, 3.7%).

Materials and Procedures

Both samples were collected using the Qualtrics online survey software (Qualtrics Labs, 2009), which provides start and end times for each respondent. Sample 1 was collected from a study that focused on psychopathy and sexually coercive behavior (O'Connell & Marcus, 2016), and Sample 2 from a study that focused on psychopathy and impulsivity. In Sample 1, the PPI-R was the first questionnaire that the participants completed, and in Sample 2, it was the second.

Psychopathic Personality Inventory–Revised. The PPI-R is a 154-item self-report scale that has been standardized and validated with both a mixed college student and community sample and an offender sample. It includes eight content scales, but most factor analyses (e.g., Benning, Patrick, Hicks, Blonigen, & Krueger, 2003) of these scales yield two factors labeled "Self-Centered Impulsivity" and "Fearless Dominance" (FD), with the Coldheartedness scale not loading highly on either factor. For the validity scales, the DR scale consists of 10 items and the VR scale includes 13 items. The IR-15 consists of 15 pairs of correlated items, and 40 pairs of correlated items comprise the IR-40. Cronbach's alphas for these scales are provided in Table 1, along with the values reported in the PPI-R manual by Lilienfeld and Widows (2005) for their community/college sample.

Triarchic Psychopathy Measure. The TriPM is a 58-item self-report scale consisting of three subscales: Boldness (19

Table 1. Internal Consistency (α) Coefficients for the PPI-R Factor and Validity Scales.

Scale	Sample 1	Sample 2	PPI-R manual
Self-Centered Impulsivity	.93	.92	
Fearless Dominance	.88	.88	
Coldheartedness	.81	.82	.78
Deviant Responding	.85	.83	.52
Virtuous Responding	.66	.64	.72
Inconsistent Responding 15	.44	.50	.33
Inconsistent Responding 40	.64	.66	.53

Note. PPI-R = Psychopathic Personality Inventory–Revised. The PPI-R manual values are from the community/college sample (Lilienfeld & Widows, 2005).

items), Disinhibition (20 items), and Meanness (19 items). The Boldness scale corresponds closely to FD from the PPI-R, Disinhibition corresponds closely to Self-Centered Impulsivity, and Meanness corresponds broadly to Coldheartedness, although it places somewhat greater emphasis on cruelty and sadism than on a paucity of feelings of interpersonal connectedness and other social emotions (e.g., Drislane, Patrick, & Arsal, 2014). In the current samples, all three scales demonstrated acceptable internal consistency (Boldness α = .75/.77, Disinhibition α = .86/.90, Meanness α = .90/.90, in Samples 1 and 2, respectively).

Results

Validity Scores and Psychometric Validity

Although the PPI-R manual (Lilienfeld & Widows, 2005) does not set specific cut scores for the DR or VR scales, the manual suggests that *t* scores above 65 should be considered suspect, which for community/college young adults would be raw scores of 17 and 34 for the DR and VR scales, respectively. Alternatively, Meade and Craig (2012) found that approximately 10% of their sample of undergraduates completing an online study responded carelessly, which, based on the distribution of scores in the current samples, would suggest provisional cut scores of 23 for DR and 35 for VR. Finally, based on their simulation studies with college students, Anderson et al. (2013) recommended higher cut scores of 25 and 38 for the DR and VR scales, respectively, to maximize classification accuracy. Consequently, we conducted the analyses examining the utility of all three sets of DR and VR cut scores.

Regardless of which of the three DR cut scores were used, the correlations between the PPI-R substantive scales and the corresponding scales from the TriPM were significantly greater in five of the six comparisons, for those scoring below the cut compared to those scoring above the cut. However, the magnitude of the differences between the correlations above and below the cut scores were, with only one exception, largest when 23 was used as the cut score, and this score has the advantage of rejecting fewer cases than the more stringent cut score of 17 (Table 2). In contrast, the VR scale generally did not moderate the associations between corresponding PPI-R and TriPM scales regardless of which cut score was used, with only one significant difference between corresponding correlations in the expected direction (FD by Boldness in Study 1; Table 2).

The PPI-R manual reports that scores of 15 or higher on the IR-15 scale and scores of 39 or higher on the IR-40 scale are "atypical," so 15 and 39 were used as cut scores for these scales. Scores on the IR-15 generally did not moderate the associations between the PPI-R and TriPM substantive scales, with only one significant decrease in validity. The IR-40 performed slightly better, with two significant differences between corresponding correlations (Table 2).[2]

Validity Scores and Time to Complete the Studies

Each of the studies included the PPI-R, the TriPM, and two other measures, resulting in 290 questionnaire items in Sample 1, and 293 items for Sample 2. Each study took roughly 30 minutes to complete on average, with median times to completion of 34.5 minutes (Study 1) and 31 minutes (Study 2). Participants who completed the studies in fewer than 15 minutes were at the 11th and 8.5th percentile for Samples 1 and 2, respectively. Consequently, it is likely that most participants who completed the studies in fewer than 15 minutes (thus averaging 3 seconds or less per question) responded in a careless, random, or haphazard manner. In both samples, participants who completed the studies in fewer than 15 minutes scored significantly higher on both the DR and VR scales. However, the effect sizes for the DR scales were considerably larger than for the VR scales. Unexpectedly, there were no significant differences between these groups for either IR scale in either sample (Table 3).

Discussion

Despite the frequency with which the PPI and the PPI-R have been used to study psychopathy (a PsycINFO search of the PPI or PPI-R yielded over 400 citations since 1996, with 279 records since 2010), only five published studies have examined the ability of the PPI/PPI-R validity scales to detect response bias. In the current study, we examined the utility of the PPI-R validity scales for identifying careless responding in two samples of college students using two distinct methods. We examined whether scores on the validity scales moderated the established associations between the PPI-R substantive scales and the TriPM, a closely related measure of psychopathic personality traits. Second, using a more direct measure of behavior, we examined whether scores on the PPI-R validity scales were associated with completing these online studies in an

Table 2. Correlations Between PPI-R Factor Scores and Corresponding TriPM Scales by Validity Scale Cut Scores.

	Sample 1				Sample 2			
Scale	*n*	FD × B	SCI × D	C × M	*n*	FD × B	SCI × D	C × M
DR								
<17	458	.78***	.73***	.54***	344	.79***	.66***	.47***
>16	113	.52***	.36***	.34***	108	.41***	.12	.39***
Z		4.4***	5.2***	2.3*		5.7***	6.0***	0.9
<23	514	.77***	.73***	.55***	407	.77***	.67***	.43***
>22	57	.34**	.11	−.02	45	.15	−.51***	.53***
Z		4.7***	5.7***	4.4***		5.4***	8.5***	−.84
<25	538	.76***	.75***	.57***	424	.76***	.67***	.53***
>24	33	.25	.11	.06	28	.08	−.59***	.64***
Z		4.0***	4.6***	3.1**		4.5***	7.2***	−0.8
VR								
<34	473	.77***	.73***	.61***	365	.74***	.60***	.59***
>33	98	.61***	.76***	.54***	87	.77***	.71***	.62***
Z		2.8**	−0.7	0.9		−0.6	−1.6	−0.4
<35	509	.76***	.75***	.60***	393	.74***	.59***	.59***
<34	62	.64***	.65***	.50***	59	.78***	.74***	.58***
Z		1.8	1.4	1.1		−0.6	−1.9	0.1
<38	545	.76***	.74***	.59***	437	.74***	.61***	.59***
>37	26	.71***	.67***	.53**	15	.92***	.85***	.33
Z		0.5	0.7	0.4		−2.2*	−1.9	1.1
IR-15								
<15	533	.76***	.75***	.60***	406	.77***	.61***	.59***
>14	38	.66***	.59***	.47**	46	.39**	.64***	.52***
Z		1.2	1.7	1.1		3.8***	−0.3	0.6
IR-40								
<39	534	.76***	.75***	.60***	419	.77***	.62***	.60***
>38	37	.59***	.52***	.43**	33	.41*	.64***	.42*
Z		1.8	2.2**	1.3		3.1**	−0.1	1.3

Note. PPI-R = Psychopathic Personality Inventory–Revised; TriPM = Triarchic Psychopathy Measure; FD = Fearless Dominance; B = Boldness; SCI = Self-Centered Impulsivity; D = Disinhibition; C = Coldheartedness; M = Meanness; DR = Deviant Responding; VR = Virtuous Responding; IR-15 = Inconsistent Responding (15 items); IR-40 = Inconsistent Responding (40 items). *Z* tests for the difference in the magnitudes of the correlations between the corresponding PPI-R and TriPM scales for those above and below the cut score for each validity scale.
**p* < .05. ***p* < .01. ****p* < .001.

Table 3. Mean PPI-R Validity Scale Score Differences Between Participants Who Completed the Studies in Under 15 Minutes or Over 15 Minutes.

	Sample 1					Sample 2				
Scale	*n*	<15	>15	*t*(*df*)	*g*	*n*	<15	>15	*t*(*df*)	*g*
DR	45/539	23.27 (4.69)	13.41 (3.83)	16.39*** (582)	2.52	39/416	22.90 (4.33)	13.69 (3.98)	13.72*** (453)	2.29
VR	44/539	31.55 (3.76)	28.65 (4.95)	4.78*** (55.98)	0.59	38/416	31.99 (3.66)	28.89 (4.93)	4.83*** (50.15)	0.64
IR-15	44/539	10.32 (5.56)	8.77 (3.44)	1.82 (45.72)	0.43	38/416	10.53 (5.02)	9.30 (3.77)	1.47 (40.91)	0.32
IR-40	44/539	28.11 (12.52)	26.30 (7.51)	0.95 (45.56)	0.23	38/416	26.08 (10.82)	27.03 (7.96)	−0.53 (40.74)	−0.12

Note. df = degrees of freedom; PPI-R = Psychopathic Personality Inventory–Revised; DR = Deviant Responding; VR = Virtuous Responding; IR-15 = Inconsistent Responding (15 items); IR-40 = Inconsistent Responding (40 items). In the *n* columns, the first number refers to the number of participants who completed the study in under 15 minutes, and the second value is those who took longer than 15 minutes. When Levene's test for equality of variances was significant, *t* tests were computed assuming unequal variances.
****p* < .001.

unrealistically brief amount of time (see also Lilienfeld & Widows, 2005). This approach to evaluating validity scales was made possible by the use of survey programs that unobtrusively note when participants start and end a study.

Across both samples and both methodologies, there was consistent evidence that the DR scale is sensitive to careless responding. Scores on the DR scale moderated the validity of the substantive PPI-R scales, with high scorers on the DR scale yielding attenuated correlations between the corresponding scales of the PPI and the TriPM. Participants who completed the studies in fewer than 15 minutes also scored considerably higher ($g > 2.0$) on the DR scales than those who completed the studies in a more credible amount of time. Even though there were also significant differences between those completing the studies in under or over 15 minutes on the VR scale, the extent of these differences were considerably smaller ($g \approx 0.6$) than on the DR scales. Furthermore, the VR score moderated the association only between one of six substantive correlations that we examined (FD by Boldness in Sample 1), and even there, the effect size was considerably smaller than for DR.

It appears that the DR cut score of 23, which corresponded to the 90th percentile in our samples and is consistent with Meade and Craig's (2012) findings regarding the rate of careless responding in online college student samples, may have functioned best. It generally yielded larger differences between corresponding correlations than the more stringent cut score recommended by the PPI-R manual (Lilienfeld & Widows, 2005) or the higher cut score yielded by Anderson et al.'s (2013) simulation study. Given that Anderson et al. created extreme groups by instructing one group to overreport, it follows that the optimal cut score in that study would be higher than in the current study. Overreporting participants in the Anderson et al. study were explicitly instructed to endorse items indicative of psychopathology (e.g., the items on the DR scale). In contrast, careless responders, like those in the current study, tend to endorse DR items by chance. In other words, individuals who have an incentive to overreport are likely to obtain higher DR scores than those who elevate this scale due to carelessness, so the cut score for detecting careless responding would be expected to be lower than the cut score for detecting overreporting. In fact, because the DR scale consists of 10 items scored on a scale of 1 to 4, the average DR scale score from a random protocol would be 25, which would mean that a cut score of 25 would miss half of random responder protocols. We generated 50,000 sets of random data and found (as would be expected) that a score of 25 on the DR scale is at the 50th percentile, whereas a score of 23 is at the 25th percentile. Although a cut score of 23 will still result in some random responders being included in subsequent analyses, this 2-point difference will halve the number of false negatives. Because false negatives (i.e., including invalid protocols in the data analysis) are more likely to result in inaccurate parameter estimates than are false positives (i.e., excluding valid profiles from the data analysis) in a large data set, we encourage researchers to consider rejecting protocols with elevated DR scores, or at least to conduct subsidiary sensitivity analyses excluding these protocols to ensure that they do not unduly affect the overall findings. The current findings are preliminary and suggest that a cut score of 23 worked well with our samples. Future research may find that different DR cut scores work best depending on the nature of the sample, methodology (e.g., online vs. in-person, proctored vs. unproctored), survey length/participant demands, base rates of careless or otherwise aberrant responding, and incentives for participation.

Unlike the current study, the only previous study to examine moderation failed to find evidence that the PPI validity scales moderated the associations between the PPI substantive scales and features associated with psychopathy (Watts et al., 2016). Methodological differences between the two studies may explain these apparently inconsistent results. Most important, Watts et al. (2016) used data individually collected from prison inmates who were monitored by research assistants. These inmates not only had no incentive to rush but also may have been inclined to take their time and respond carefully rather having to quickly return to their regular duties or cells. In contrast, some of the students in our samples may have been inclined to complete the assignment as quickly as possible to receive their course credit. Second, it is easier to detect moderation when the predictor and criterion variables are highly correlated (Morey, 2012). In the current study, the PPI-R factor scores and the corresponding TriPM scales were highly correlated, in contrast to the Watts et al. article where many of the correlations between the predictors and the criterion variables were small (although it should be noted that in the Watts et al. analyses, moderation was not found even when these correlates were medium to large in magnitude). Finally, for Watt's et al.'s moderation analyses, the scores on the validity scales were analyzed as continuous variables. It is possible that the moderation effects for these scales are curvilinear. For example, low and medium scores on the DR may not indicate compromised validity, but high scores might. Conversely, beyond a certain score, even higher scores may not indicate a less valid protocol. Watts et al. did conduct a supplementary analysis that examined curvilinear moderation by squaring the scores on the validity scales, which also failed to find moderation. Despite the methodological differences between the studies, it may be worthwhile to examine if using a cut score of 23 on the DR scale moderates the association between the PPI substantive scales and the criterion variables that were most highly correlated with the PPI in the Watts et al. data.

Considering that the IR scales were specifically designed to detect careless or inconsistent responding, their performance in these two sets of analyses was less impressive than expected. Neither the IR-15 nor the IR-40 differentiated between those who completed the studies in under or over 15 minutes. Although the IR-40 appears to be a more successful moderator of the association between the PPI-R substantive scales and those of the TriPM than the IR-15 or the VR scale, it moderated only two of the six associations that we examined. In contrast, the DR scale moderated five of the six associations. Furthermore, the IR-40 identified fewer invalid protocols than the DR scale (with a cut score of 23), and even when the IR-40 cut score was lowered to correspond to the 90th percentile on this scale in the current samples, it did not function any more effectively.

Despite the respective aims of the IR and DR scales, it appears that the DR scale is better at identifying careless responding than the IR scales in research settings. A likely explanation why the DR scale better identified careless responding than the IR scales concerns the construction of IR scales. Each item of the IR scales is the absolute value of the difference between two corresponding items, which means that the scores vary between 0 and 3 and are not evenly distributed: Four pairs of scores can yield a 0, six combinations can yield a 1, four combinations can yield a 2, and only 2 combinations can yield a 3. Therefore, the average item score when randomly responding to the IR scale is 1.25, not 1.5. According to the PPI-R manual (Lilienfeld & Widows, 2005), the correlations between each pair of linked items on the IR-15 ranged from r = .48 to .66 in their college/community standardization sample. On the one hand, this is an impressive degree of agreement for single items, but it also means that even among non-careless responders, nonidentical responses to paired items will not be uncommon. Random responses to the IR-40 would average 50, and based on our simulation of 50,000 sets of random data, a score of 39 or higher would occur in 96% of random protocols. However, given the moderate correlations among the paired IR items, a score of 39 might also occur by chance in a valid protocol, which could explain why those protocols above 39 on the IR-40 scale still yielded reasonably high correlations between the PPI-R and corresponding TriPM scales. Note that the correlations between the PPI-R and corresponding TriPM scales were similar in magnitude for those below the cut on the DR or IR scales but that they appeared to be considerably higher for those above the cut on the IR scales than they were for those above the cut on the DR scale (Table 2).[3] The limited success of the IR scales in the current study lends support to Nikolova et al.'s (2012) concerns about the validity of IR scales and their call for additional research on the IR scales of other commonly used assessment instruments. IR scales may be capable of achieving adequate validity only if the item pairs that comprise the scale have challengingly high interitem correlations. It is also possible that ICN scales work better for scales with dichotomous items (e.g., Minnesota Multiphasic Personality Inventory–2), in which an inconsistency would typically indicate a direct contradiction, than for Likert-type items, in which an inconsistency could reflect either a difference in magnitude or a subtle distinction between two slightly different situations, attitudes, emotions, or cognitions.

Limitations and Future Directions

Although the findings from the two samples were generally consistent, this study was not without limitations. The percentage of the samples scoring above the cut scores for the DR and VR scales was considerably higher than in the PPI-R community/college normative samples (Lilienfeld & Widows, 2005). The use of online data collection and the aforementioned incentives system, which did not reward careful responding, may have accounted for the elevated DR and VR scales in the current studies. Studies using in-person data collection will be necessary to determine whether the current findings generalize beyond online samples, and whether an alternative DR cut score might be more optimal for in-person studies. Additionally, the 15-minute cutoff for the time to completion analyses was an educated guess based on the number of items in the studies and the average time it took participants to complete these studies. It is possible that some respondents who were very fast readers carefully completed the studies in under 15 minutes, but to the extent that this was the case, it should have attenuated the group differences on the PPI-R validity scales. Although more participants exceeded the cut score of 23 on the DR scale than took less than 15 minutes to complete the study, the Qualtrics software tracked how long participants had the study open but not the actual amount of time they spent working on it. Thus, some careless responders may have erroneously appeared to complete the questionnaires in a reasonable amount of time. Finally, because neither study was originally designed to assess the validity of the PPI-R validity scales, the placement of the PPI-R in the first and second positions in the questionnaire protocol may not have been ideal. Boredom or fatigue may have led some participants to become more careless toward the end of the studies (where the TriPM was located), despite providing valid scores on the PPI-R validity scales.

The criterion measure in the current study was also a self-report measure, raising the risk that shared response bias inflated the apparent validity of these PPI-R validity indicators (McGrath et al., 2010). However, this issue of shared method variance is more of a problem for detecting impression management using the suppression method than for detecting careless responding using moderation analyses.[4] These latter analyses should not be problematic because participants who responded carelessly on the PPI-R were also likely to respond carelessly on the TriPM (and on the DR scale), so to the extent that the DR scale assesses careless responding, this would have attenuated correlations between the PPI-R and the TriPM for high scorers on the DR scale. It is difficult to imagine a credible alternative explanation apart from careless responding (or misunderstanding the items across all of the scales) for how high scores on the DR scale could attenuate the correlations between the corresponding PPI-R and TriPM scales. Regardless, studies examining whether these validity scores moderate the association between the substantive PPI-R scales and other psychopathy-related behaviors using different modes of assessment (e.g., startle response, exploitive behavior, impulsive behavior) would provide additional evidence regarding the validity of these validity scales (although because the associations between psychopathy and these external correlates are modest, it will probably require very large samples to detect moderation).

The current analyses do not address whether the DR and VR scales are valid measures of over- or underreporting. As a complement to simulation studies, one option would be for researchers to analyze whether PPI-R validity scales from protocols collected in applied settings (e.g., forensic evaluations) moderate the validity of the PPI-R substantive scales. Although it is still possible that such moderation may be driven by careless responding, evidence of reduced validity among high scorers on these scales when respondents have stronger motivations to over- or underreport would suggest that these scales are assessing impression management and not just carelessness. For example, Edens and Ruiz (2006) found that the interaction between the Positive Impression Management and the Antisocial Features scales of the Personality Assessment Inventory predicted institutional misconduct in male inmates. An ecologically valid study of the ability of the PPI-R DR and VR scales to detect over- and underreporting may require the collection of data from a wide range of forensic assessments.

Overall, we found that the IR scales are not effective for identifying careless responding in college samples, which is consistent with Nikolova et al.'s (2012) findings with correctional samples. Consequently, researchers and clinicians who use the PPI-R will not only have to rely on the DR scale to identify careless responders but will also be faced with the challenge of discerning whether high scores on the DR scale reflect careless reporting or overreporting. Very high scores on the DR scale are unlikely to be due to chance: In our Monte Carlo data, only 20% of random protocols yielded DR scores of 28 or higher, which was the mean score for the overreporting group in the Anderson et al. (2013) simulation study. For DR scores in the mid-20s, investigators may have to rely on the context of the assessment to infer the likely source of the elevation (e.g., "Is there an incentive to overreport pathology?" "Is there evidence that the participant responded quickly and carelessly?"), or include supplementary validity scales from other instruments to inform this determination. Regardless, the results of the current study strongly indicate that researchers attend to the DR scale when using the PPI-R to study psychopathic personality traits in college students. The extent to which these findings are generalizable to other samples, including psychiatric samples, merits further investigation.

Declaration of Conflicting Interests

The author(s) declared the following potential conflict of interest with respect to the research, authorship, and/or publication of this article: Dr. Scott O. Lilienfeld is the codeveloper of the Psychopathic Personality Inventory–Revised, and receives royalties for the sale of this measure, which is published by Psychological Assessment Resources in Lutz, Florida.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In the case of suppression, higher scores on the validity scale would suggest that the respondent is underreporting (or overreporting) and that a correction should be added to the substantive scale to increase its accuracy (e.g., the K correction on the Minnesota Multiphasic Personality Inventory–2; Butcher et al., 2001).
2. A cut score of 37, which was the 90th percentile on the IR-40, was also examined, but it functioned no better than a cut score of 39.
3. When a cut score of 45 on the IR-40 is used, the correlations between the corresponding PPI-R and TriPM scales drop considerably for those above the cut, but this cut score is so stringent that it captures very few protocols (16 in Sample 1 and 10 in Sample 2). Therefore, it is likely to miss careless protocols and have limited practical value.
4. Suppression analyses require a multimethod approach because if respondents underreported on the PPI-R, they probably also underreported on the TriPM, so there would be no way to detect suppression. Thus, examination of suppression, which would be most relevant to evaluating the ability of VR scale to detect positive impression management, would require another measure of psychopathic traits not based on self-report (e.g., peer ratings).

References

Anderson, J. L., Sellbom, M., Wygant, D. B., & Edens, J. F. (2013). Examining the necessity for and utility of the Psychopathic Personality Inventory–Revised (PPI-R) validity scales. *Law and Human Behavior*, *37*, 312-320.

Benning, S. D., Patrick, C. J., Hicks, B. M., Blonigen, D. M., & Krueger, R. F. (2003). Factor structure of the psychopathic personality inventory (PPI): Validity and implications for clinical assessment. *Psychological Assessment*, *15*, 340-350.

Ben-Porath, Y. (2003). Assessing personality and psychopathology with self-report inventories. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 553-577). Hoboken, NJ: John Wiley.

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

Drislane, L. E., Patrick, C. J., & Arsal, G. (2014). Clarifying the content coverage of differing psychopathy inventories through reference to the Triarchic Psychopathy Measure. *Psychological Assessment*, *26*, 350-362.

Edens, J. F., Buffington, J. K., & Tomicic, T. L. (2000). An investigation of the relationship between psychopathic traits and malingering on the Psychopathic Personality Inventory. *Assessment*, *7*, 281-296.

Edens, J. F., Buffington, J. K., Tomicic, T. L., & Riley, B. D. (2001). Effects of positive impression management on the Psychopathic Personality Inventory. *Law and Human Behavior*, *25*, 235-256.

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.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, *68*, 151-158. doi:10.1037/0022-3514.68.1.151

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. O., & Widows, M. R. (2005). *Psychopathic Personality Inventory Revised (PPI-R): Professional manual*. Lutz, FL: Psychological Assessment Resources.

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

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455.

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

Morey, L. C. (2012). Detection of response bias in applied assessment: Comment on McGrath et al. (2010). *Psychological Injury and Law*, *5*, 1-9.

Nikolova, N. L., Hendry, M. C., Douglas, K. S., Edens, J. F., & Lilienfeld, S. O. (2012). The inconsistency of inconsistency scales: A comparison of two widely used measures. *Behavioral Sciences & the Law*, *30*, 16-27.

O'Connell, D., & Marcus, D. K. (2016). Psychopathic personality traits predict positive attitudes toward sexually predatory behaviors in college men and women. *Personality and Individual Differences*, *94*, 372-376.

Patrick, C. J. (2010). *Triarchic psychopathy measure (TriPM)*. Retrieved from https://www.phenxtoolkit.org/index.php?pageLink=browse.protocoldetails&id=121601

Paulhus, D. L., Neumann, C. S., & Hare, R. D. (in press). Manual for the Self-Report Psychopathy Scale. Toronto, Ontario, Canada: Multi-Health Systems.

Qualtrics Labs. (2009). *Qualtrics Research Suite* (Version 12,018) [Computer software]. Provo, UT: Author.

Tellegen, A. (in press). *Manual for the Multidimensional Personality Questionnaire*. Minneapolis: University of Minnesota Press.

Watts, A. L., Lilienfeld, S. O., Edens, J. F., Douglas, K. S., Skeem, J. L., Verschuere, B., & LoPilato, A. C. (2016). Does response distortion statistically affect the relations between self-report psychopathy measures and external criteria? *Psychological Assessment*, *28*, 294-306.