Article

SAGE Open
July-September 2014: 1–10

DOI: 10.1177/2158244014545964
sgo.sagepub.com
SAGE

The Conscientious Responders Scale: A New Tool for Discriminating Between Conscientious and Random Responders

Zdravko Marjanovic[1], C. Ward Struthers[2], Robert Cribbie[2], and Esther R. Greenglass[2]

Abstract

This investigation introduces a novel tool for identifying conscientious responders (CRs) and random responders (RRs) in psychological inventory data. The Conscientious Responders Scale (CRS) is a five-item validity measure that uses instructional items to identify responders. Because each item instructs responders exactly how to answer that particular item, each response can be scored as either correct or incorrect. Given the long odds of answering a CRS item correctly by chance alone on a 7-point scale (14.29%), we reasoned that RRs would answer most items incorrectly, whereas CRs would answer them correctly. This rationale was evaluated in two experiments in which CRs' CRS scores were compared against RRs' scores. As predicted, results showed large differences in CRS scores across responder groups. Moreover, the CRS correctly classified responders as either conscientious or random with greater than 93% accuracy. Implications for the reliability and effectiveness of the CRS are discussed.

Keywords

random responding, validity scale, personality, inventory, psychometric

When a self-report psychological inventory[1] is administered, the expectation is that respondents follow testing instructions and answer its items as honestly and accurately as possible. That is to say they respond *conscientiously* and thereby infuse their responses with meaning about their inner psychological workings. Unfortunately, not all of the data that respondents produce are generated consciously and, therefore, not all data are valid. Some individuals, for example, purposefully distort their responses to be perceived more positively or negatively than they really are. This is known as faking good and faking bad, respectfully (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989).

Random responding is another form of data distortion in which respondents endorse items indiscriminately. Random responders (RRs) answer items without regard for what they mean. For example, within a single inventory, a RR might answer "true" to an item like "I am taller than most people," but because he or she is answering indiscriminately, answer "false" to a similar item such as "I am tall." Resulting from a number of factors such as carelessness, fatigue, psychopathology, and low intelligence (Bentler, Jackson, & Messick, 1971), the prevalence of random responding in self-report data has been estimated to be up to 5% in non-disordered populations (Clark, Gironda, & Young, 2003; Pinsoneault, 2002), and 10% or more in disordered and forensic populations (Archer, Handel, Lynch, & Elkins, 2002; McNulty et al., 2003).

For clinical and applied psychologists, the presence of random data in their supposed valid data sets may lead them to make erroneous conclusions, diagnoses, and/or predictions about their clients (Ben-Porath & Waller, 1992). Bruehl, Lofland, Sherman, and Carlson (1998) showed this possibility in a clever study using a widely used pain inventory. They concluded that if the measure was administered to a group of RRs in a clinical setting and their random responding went unidentified, 35% of them would be classified as having elevated levels of interpersonal distress and another 35% as being highly adaptive copers. For researchers, random responding poses different problems. Primarily, it increases measurement error, making it more difficult to identify significant relations when they are present in data. In other words, it increases the likelihood of making Type II errors and otherwise jeopardizes the validity of one's results (Osborne & Blanchard, 2011). In a recent study, Credé (2010) showed that even low rates of random responding (e.g., 5%) can have meaningful moderating effects on the

[1]Thompson Rivers University, Kamloops, British Columbia, Canada
[2]York University, Toronto, Ontario, Canada

Corresponding Author:
Zdravko Marjanovic, Department of Psychology, Thompson Rivers University, 900 McGill Road, Kamloops, British Columbia, Canada V2C 0C8.
Email: zmarjanovic@tru.ca

size and direction of correlations, even increasing the likelihood of making Type I errors.

Old Approaches to Detecting Random Responding

Historically, there have been two types of validity scales that have been effective at identifying RRs in self-report inventory data: infrequency scales and inconsistency scales. An infrequency scale is composed of absurd item content—items that are endorsed so infrequently (e.g., <10% of conscientious responders [CRs] answer "true" to the item "I drink 10 glasses of milk a day") that it is reasonable to interpret responses in the infrequent direction as highly unusual. If a respondent endorses too many infrequency items in the unusual direction, it strongly indicates the presence of random responding and, therefore, an invalid test profile. An exemplar of an infrequency scale is the Minnesota Multiphasic Personality Inventory–2's (MMPI-2) *F* Scale (Butcher et al., 1989).

The logic behind an inconsistency scale is that if individuals are paying attention to item content, they should respond consistently to items that are semantically similar and inconsistently to items that are semantically dissimilar. For example, a person who answers "true" to an item like "I am a happy person" should also answer "true" to the item "I am happy." An inconsistency scale is created by identifying pairs of items that are usually answered in the same way, usually correlating above .90 (e.g., Butcher et al., 1989). Given this, if a person responds inconsistently to several such item pairs, it strongly indicates the likelihood of random responding and that the responder's data are invalid. Exemplars of inconsistency scales are the Variable Response Inconsistency (VRIN) and True Response Inconsistency (TRIN) scales of the MMPI-2 (Butcher et al., 1989).

Although effective, researchers rarely use or develop these scales due to their extensive costs. First, embedding an infrequency scale in a questionnaire makes it longer to administer, making it costly in terms of fatigue for the test taker and labor intensive for the test administrator. More importantly, because researchers must expect that a small proportion of CRs answer these items truthfully but in the infrequent direction (e.g., some individuals really do drink 10 glasses of milk a day), extensive normative testing is first necessary to establish base rates of infrequent responding in samples of CRs. In addition, normative testing is required to establish base rates for different categories of responders (e.g., disordered, non-disordered) and the various settings in which these scales are likely to be administered (e.g., vocational, forensic). For example, in a disordered population, one should expect that mean infrequency scale scores are higher than they are in a non-disordered population (see Archer et al., 2002; McNulty et al., 2003). This is partly due to the fact that in a disordered population, CRs are more likely to endorsee infrequency items in the infrequent direction. An item like "I talk to dead people" may be affirmatively endorsed because the responder actually believes he or she is communicating with the dead, not because they are answering indiscriminately. This has been a long-standing criticism of infrequency scales—that infrequency scale scores can sometimes confound random responding with psychopathology (Arbisi & Ben-Porath, 1995).

In contrast, inconsistency scales have the advantage not having to embed additional items in an inventory because they are made up of an inventory's existing items. They do, however, still require an extensive amount of research to create and validate. First, a psychologist must identify a pool of highly correlated items within a measure from which to create its inconsistency scale. Subsequently, the psychologist must generate normative data to determine the optimal cutoff score that most effectively discriminates between CR and RR. Similar to the above concerns, some cutoff scores may be more appropriate to a particular responder population and setting than others. Consequently, the entire process is unappealingly laborious, time-consuming, and complex.

Unfortunately, apart from these costly types of infrequency- and inconsistency-validity scales, there are currently no practical means for psychologists to differentiate between CR and RR in self-report inventory data (see Meade & Craig, 2012, for an evaluation of item-based and statistically based indices). Researchers and applied psychologists alike therefore stand to benefit from a simple, reliable, and flexible tool that can effectively identify RRs in inventory data without all of the above mess associated with traditional validity scales. The development and evaluation of such a tool was the primary goal of this investigation.

Using Instructional Item Content to Identify Random Responding

The tool we developed for this investigation is called the Conscientious Responders Scale (CRS; see the appendix), which is a five-item variant of a traditional validity scale. The main advantage of the CRS over standard infrequency and inconsistency scales is that it does not require extensive normative testing to establish its cutoff scores.[2] This is because the CRS is made up of instructional item content. Instructional item content directs responders how to answer each particular item (e.g., CRS item 3, "To respond to this question, please choose option number five, 'slightly agree'"); thus, unlike typical psychological inventory items, there is an objectively correct response for every item. Each correct response is given a score of 1 and incorrect response a score of 0. A CRS total score is generated by summing up all of a respondent's correct responses. Thus, scores range from 0 (all incorrect responses) to 5 (reflecting all correct responses).

Depending on how many response options a responder has to choose from, the likelihood that a RR will answer an item correctly by chance alone can be estimated a priori using probability theory. In our investigation, we used a 7-point response-option scale for all measures. Consequently, the probability of a RR answering a CRS item correctly was 14.29% (i.e., $1/k$, where k is the number of response options). Because the probability of answering several items correctly is great deal lower, a high CRS score was very likely the result of conscientious responding. This raised the following question: How high a CRS score is necessary to reliably discriminate between CR and RR?

Calculating A Priori CRS Cutoff Scores

Using the binomial distribution, we were able to determine that only 2.33% of RRs would be able achieve a CRS score of 3 or higher by chance alone (percentage answering 3 correct = 2.14% + 4 correct = 0.18% + 5 correct = 0.01%), whereas 15.19% would be able to answer 2 or more correctly by chance alone (percentage answering 2 correct = 12.86%). In fact, most RRs would only be able to generate a CRS score of 0 (46.25%) or 1 (38.56%), answering almost no items correctly. Using the ubiquitous critical probability value of $p < .05$ as our guide, we settled on a 2/3 cutoff score for the purposes of this investigation. That is, because fewer than 5% of RRs can be expected to achieve CRS scores of 3, 4, or 5 by chance alone, we are confident that these scores will be reflective of conscientious responding. Thus, responders with a CRS score of 3 or higher will be labeled "conscientious responders." In contrast, responders with CRS scores of 2 or lower (i.e., 0, 1, or 2), which are statistically indistinguishable from scores generated by random responding, will be labeled "random responders."[3]

The Present Investigation: Purpose, Design, and Strategy

The purpose of this investigation was to evaluate the effectiveness of the CRS for discriminating between CR and RR in self-report inventory data. We evaluated our new measure in two identical experiments in which a single 89-item questionnaire was administered to university students either using a paper-and-pencil format (Experiment 1) or over the Internet (Experiment 2). The question of interest was whether CRS scores could reliably discriminate between RR and CR across these two widely used means for data collections.

According to recent studies, Internet collected data are equivalent to data collected through traditional methods, such as paper-and-pencil data questionnaires, in terms of psychometric properties such as factor structures, inter-scale correlations, means, and standard deviations (Johnson, 2005). Importantly, Internet data have not yet been heavily scrutinized in terms of data distortion tendencies such as random responding. In one such study, Pettit (2002) found that paper-and-pencil responders actually produced slightly higher rates of random responses than their Internet counterparts. In that study, however, the Internet participants were all self-selected and therefore probably highly motivated to participate from the outset. The extent to which the average undergraduate student completes questionnaires conscientiously has long been a source of doubt and controversy among psychologists (Sears, 1986). Because students are often compelled to participate in psychological research as a means to fulfill program requirements, the typical student is probably less motivated to participate conscientiously than we expect him or her to be. Given that the Internet provides participants with greater anonymity than traditional forms of data collection, unmotivated students may take advantage and produce more random responding than they normally would on a paper-and-pencil questionnaire. Experiment 2 was conducted to replicate the findings of Experiment 1 in a typical online administration of a psychological questionnaire. We also included a traditionally developed infrequency scale in our questionnaire to serve as a comparative measure against which we could evaluate the effectiveness of the CRS.

Data were collected in both experiments using the *analog* design (e.g., Clark et al., 2003). Participants were randomly assigned to one of two experimental conditions. In the *CR* condition, participants were instructed to answer items as honestly and accurately as possible. In the *RR* condition, participants were given a questionnaire with no actual item content, only blank response options, and instructed to complete the response-option sheet as randomly as possible.

In each experiment, the statistical strategy for evaluating CRS was threefold. First, group differences on CRS and Pettit Random Responding Scale (PRRS) scores would be assessed using independent-samples t tests. Hypothesis 1 states that participants assigned to the CR group will produce significantly greater scores on the CRS and PRRS than participants in the RR group. Second, because the CRS and PRRS purportedly measure the same construct (i.e., conscientious responding), we conducted a correlation analysis to reveal whether these measures were indeed related. Hypothesis 2 states that the CRS will be strongly and positively correlated with the PRRS. Finally, in order that the CRS proves its worth as a tool for identifying CR and RR, Hypothesis 3 states that it will correctly label participants in the CR group as "conscientious responders" and participants in the RR condition as "random responders" at an average classification accuracy rate of ≥80% (cf. Clark et al., 2003). This analysis was also conducted on the PRRS for the purposes of comparison.

Method

Participants

Experiment 1. A total of 68 participants were recruited from a second-year psychology class in exchange for being entered into a draw for $50 (CAD). In total, 33 students were randomly assigned to the CR condition and 35 were randomly assigned to the RR condition. Three participants were removed from the CR sample due to missing data. The final total sample (N = 65; CR = 30, RR = 35) consisted of 46 women and 19 men, with a mean age of 24.22 years (SD = 5.61). As we expected to find a "large" effect of CRS scores across responder groups (Cohen's $d \geq .80$), this sample size was more than adequate to detect statistical significance and avoid Type II errors.

Experiment 2. A total of 412 participants were recruited from an undergraduate research participant pool in exchange for course credit. Thirty-two participants (7.8%) were removed due to missing data. The final total sample (N = 380, CR = 191 and RR = 189) consisted of 120 men, 258 women, and 2 individuals who did not report their sex. The mean age of the sample was 20.67 years (SD = 3.62).

Measures

The same questionnaire was administered in both experiments. In addition to the CRS and PRRS, all of the measures included were selected based on acceptable levels of internal consistency and breadth of content, such as perfectionism and ethics. The subject matter and validity of each scale was irrelevant to their selection. All of the questionnaire's 89 items, including the CRS and PRRS items, were presented in a scrambled, random order. All items were answered on a 7-point Likert-type scale, ranging from 1 = *strongly disagree* to 7 = *strongly agree*, with 4 = *neither agree nor disagree* at the midpoint. The questionnaire consisted of the following measures.

Self-Esteem Scale (SES). The 10-item SES is a widely used self-report measure of trait self-esteem (Rosenberg, 1965). It has acceptable internal consistency across a variety of samples and has been extensively used in psychological research (Blascovich & Tomaka, 1993). Higher scores reflect greater levels of trait self-esteem. A sample item is "I wish I could have more respect for myself."

Right-Wing Authoritarianism–Short Form (SRWA). The 14-item SRWA (Manganelli Rattazzi, Bobbio, & Canova, 2007) was created by factoring Altemeyer's (1996) 30-item RWA scale into two subscales measuring Authoritarian Aggression and Submission (SRWA-A) and Conservatism (SRWA-C). Each subscale has acceptable reliability and correlates highly with the original 30-item RWA scale (Bobbio, Canova, & Manganelli, 2010). Higher scores on either subscale reflect greater levels of RWA. A sample item from the authoritarian aggression and submission scale is "The situation in our country is getting so serious, the strongest methods would be justified if they eliminated the troublemakers and got us back to our true path."

Multidimensional Perfectionism Scale (MPS). The 35-item MPS (Frost, Marten, Lahart, & Rosenblate, 1990) is a scale that assesses six factors of trait perfectionism. Its subscales are Concern Over Mistakes (MPS-CM), Organization (MPS-O), Parental Criticism (MPS-PC), Personal Standards (MPS-PS), Doubts (MPS-D), and Parental Expectations (MPS-PE). The MPS has acceptable internal consistencies, with estimates for its subscales ranging between .73 and .93 (Frost et al., 1990), and has been used extensively in perfectionism research (Parker & Adkins, 1995). Higher scores reflect greater levels of perfectionism on all subscales. A sample item from the personal standards subscale is "I set higher goals than most people."

Ethics Position Questionnaire (EPQ). The 20-item EPQ (Forsyth, 1980) measures the philosophical framework from which individuals justify their decisions and behaviors. It contains two subscales, Idealism (EPQ-I) and Relativism (EPQ-R), which previous research has shown to be internally consistent measures (e.g., Davis, Andersen, & Curtis, 2001). Higher scores reflect greater levels of both ethical idealism and relativism. A sample item from the relativism subscale is "Whether a lie is judged to be moral or immoral depends upon the circumstances surrounding the action."

PRRS. The PRRS is a 10-item infrequency scale containing all absurd item content (Pettit, 1999, 2002). In the original scale, items endorsed in the infrequent (statistically unusual) direction are scored as 1s, whereas items endorsed in the frequent direction are scored as 0s. We reversed this scoring system so that higher PRRS sum scores reflect greater conscientious responding, not random responding. The original measure's cutoff score had also to be altered because of this scaling change, such that the original cutoff score of 2/3 was changed to 7/8. Put another way, only responders who achieved a high score of 8, 9, or 10 were labeled "conscientious responder." Low scorers (i.e., 7 and less) were labeled "random responder."

The original scale was validated on a large Internet sample using a dichotomous response scale and its psychometric properties are acceptable (Pettit, 1999, 2002). Given that a 7-point scale was used in this investigation, the likelihood of a RR answering an item correctly by chance alone was reduced from 50% to 14.29%, theoretically making the PRRS' 7/8 cutoff score a more difficult standard for a RR to meet. A sample PRRS item is "Sometimes I feel warm or cool," to which answering anything but "strongly agree" is an empirically infrequent response and is assigned a score of 0.

Table 1. Descriptive Statistics Across Responder Groups and Experiments..

	Experiment 1					Experiment 2				
	Responder group					Responder group				
	CR group (n = 30)			RR group (n = 35)		CR group (n = 191)			RR group (n = 189)	
Measure	*M*	*SD*	α	*M*	*SD*	*M*	*SD*	α	*M*	*SD*
CRS	4.53	1.43	.98	0.74	0.70	4.39	1.27	.84	0.86	0.81
PRRS	8.70	1.26	.41	0.94	1.08	6.76	2.50	.80	0.94	1.09
SES	5.63	1.33	.91	3.95	0.50	5.26	0.98	.88	3.96	0.50
SRWA-A	3.16	1.40	.86	4.04	0.75	3.77	1.17	.86	4.14	0.71
SRWA-C	2.90	1.17	.81	4.10	0.76	3.26	0.98	.74	3.83	0.74
MPS-CM	2.65	1.05	.76	4.10	0.67	3.37	1.07	.86	4.16	0.67
MPS-PS	4.69	1.43	.85	3.97	0.60	4.80	0.91	.80	4.11	0.74
MPS-PE	3.76	1.74	.83	4.31	0.79	4.48	1.29	.85	4.06	0.87
MPS-PC	2.47	1.57	.87	4.27	0.75	3.22	1.40	.83	4.06	0.93
MPS-D	3.28	1.49	.73	3.91	0.82	3.66	1.12	.67	4.21	0.94
MPS-O	5.33	1.75	.93	4.18	0.71	5.24	1.12	.93	4.06	0.81
EPQ-I	4.64	1.32	.75	4.10	0.51	4.97	0.73	.72	4.04	0.69
EPQ-R	4.35	1.23	.77	3.98	0.67	4.54	0.71	.67	4.12	0.67

Note. Cronbach's alpha (α) was calculated in the unconfirmed conscientious responder group only. CR = conscientious responder; RR = random responder; CRS = Conscientious Responders Scale; PRRS = Pettit Random Responding Scale; SES = Self-Esteem Scale; SRWA = Right-Wing Authoritarianism–Short Form (A = Authoritarian Aggression and Submission; C = Conservatism); MPS = Multidimensional Perfectionism Scale (CM = Concern Over Mistakes; PS = Personal Standards; PE = Parental Expectations; PC = Parental Criticism; D = Doubts; O = Organization); EPQ = Ethics Positions Questionnaire (I = Idealism; R = Relativism).

CRS. The CRS is a variant of a traditional infrequency scale that relies on instructional item content to identify CRs and RRs (see the appendix). The CRS is made up of five items that direct the responder exactly how to answer that particular item, such that each item has only one possible correct response. Thus, the number of the measure's items, as well as the number of response options, can be used to generate effective cutoff scores using probability theory. For the purposes of this investigation, we adopted the $p < .05$ critical value as our standard. Using the binomial distribution, we calculated that fewer than 5% of RRs would be able to achieve a CRS score of 3, 4, or 5, and were most likely to achieve a score of 0, 1, or 2. Consequently, 2/3 became our cutoff score to discriminate between CR and RR. Higher scorers (3 and above) were labeled "conscientious responder" and low scorers (2 and below) were labeled "random responder."

Procedure

Whether recruited in a second-year psychology class to complete an in-class paper-and-pencil questionnaire (Experiment 1) or from an undergraduate research participant pool to complete an online questionnaire of the same length (Experiment 2), participants were randomly assigned to complete one of two versions of the 89-item questionnaire. Participants assigned to the CR group received standard questionnaire instructions with some additional language that prepared them for the instructional nature of the CRS items: "Some of the items will ask you to answer them in a particular way . . ." Their questionnaire contained all of the measures listed above. In contrast, participants assigned to the RR group received a questionnaire with no items inside, only a 7-point Likert-type scale they had to endorse for each missing item. These participants were instructed to "respond on the scales below as randomly as possible, but do this in such a way that it will not be apparent that this is what you did." Thus, in an effort to simulate reality, participants tried not to make their random responses so obvious that they would be easily identified by a visual inspection of the data. These instructions have similarly been used in other validity scale studies that used the analog design (e.g., Clark et al., 2003). In the debriefing, no CR participants reported having any difficulty understanding the instructions or completing the items in the questionnaire.

Results

Descriptive statistics for all of the measures across responder groups and experiments are presented in Table 1. Our first analysis involved conducting independent-samples *t* tests to test Hypothesis 1. In Experiment 1, as predicted, the CR

Table 2. CRS Scores Across Responder Groups and Experiments.

	Experiment 1						Experiment 2					
	CR group (n = 30)			RR group (n = 35)			CR group (n = 191)			RR group (n = 189)		
CRS score	Score frequency	Score percentage	Cumulative percentage	Score frequency	Score percentage	Cumulative percentage	Score frequency	Score percentage	Score frequency	Score percentage	Score frequency	Score percentage
5	27	90.00	90.00	0	0.00	0.00	138	72.25	72.25	0	0.00	0.00
4	0	0.00	93.33	0	0.00	0.00	29	15.18	87.43	0	0.00	0.00
3	0	0.00	93.33	1	2.86	2.86	5	2.62	90.05	6	3.17	3.17
2	0	0.00	93.33	2	5.71	8.57	3	1.57	91.62	32	16.93	20.11
1	1	3.33	93.33	19	54.29	62.86	11	5.76	97.38	80	42.33	62.43
0	2	6.67	100	13	37.14	100.00	5	2.62	100	71	37.57	100.00

Note. CRS scores = larger scores (i.e., 5, 4, and 3) reflect a greater rate of conscientious responding. Cumulative percentage = % of sample to score at or above the CRS score. CRS = Conscientious Responders Scale; CR = conscientious responder; RR = random responder.

group produced significantly larger CRS scores (M = 4.53, SD = 1.43) and PRRS scores (M = 8.70, SD = 1.26) than their RR counterparts (M = 0.74, SD = 0.70, and M = 0.94, SD = 1.08, respectively), $t(63) = 13.86$, $p < .001$, $d = 3.37$, and $t(63) = 26.86$, $p < .001$, $d = 6.61$, respectively. In Experiment 2, results were much the same. The CR group produced significantly larger CRS scores (M = 4.39, SD = 1.27) and PRRS scores (M = 6.76, SD = 2.50) than the RR group (M = 0.86, SD = 0.81, and M = 0.94, SD = 1.09, respectively), $t(378) = 32.32$, $p < .001$, $d = 3.31$, and $t(378) = 29.34$, $p < .001$, $d = 3.02$, respectively. That is, across experiments, CRs were much more likely to answer CRS items correctly and PRRS items in the frequent direction than were RRs. Hypothesis 1 was therefore fully supported.

We next calculated zero-order correlations to test Hypothesis 2, that the CRS and PRRS would be strongly positively related because they both purportedly measure the same construct. This hypothesis was also fully supported by the data in Experiment 1, $r(63) = .87$, $p < .001$, and Experiment 2, $r(378) = .80$, $p < .001$. In general, if responders scored highly on the CRS, they were also very likely to have scored highly on the PRRS.

In the final stage of the analysis, we examined CRS scores in the RR and CR groups to assess the effectiveness of our theoretically derived, a priori cutoff score (see Table 2 for CRS scores across responder groups and experiments). In the RR group in Experiment 1 (n = 35), as expected, 0 participants answered four or all five CRS items correctly, and only 1 participant answered three items correctly by chance alone. Thus, as expected, fewer than 5% of RRs were capable of producing a CRS score of 3, 4, or 5. In contrast, in the CR group, 27 participants produced CRS scores of 5, answering all CRS items correctly. The remaining 3 participants' scores fell below the 2/3 cutoff at 1, 0, and 0, answering nearly all of the items incorrectly. Altogether, our a priori 2/3 cutoff produced a classification accuracy rate (i.e., number of correctly labeled responders to their responders groups/total n of that responders group) of 90.00% in the CR group (making 3 errors of 30), 97.14% in the RR group (making 1 error of 35), and 93.85% averaged across both groups, in full support of Hypothesis 3. The PRRS similarly correctly labeled 26 of 30 CR responders as "conscientious" (86.67%), 35 of 35 RR responders as "random" (100%), and produced an overall classification accuracy rate of 93.85%.

In Experiment 2, results were highly similar for the CRS. Altogether, our a priori 2/3 cutoff produced a classification accuracy rate of 90.05% in the CR group (making 19 errors of 191), 96.83% in the RR group (making 6 errors of 189), and 93.42% averaged across both groups, in full support of Hypothesis 3. In contrast, results were substantially worse for the PRRS. The PRRS correctly labeled 47.12% in the CR group as "conscientious responders" (101 errors of 191), 100% in the RR group as "random responders" (0 errors of 189), and achieved a 73.42% classification accuracy averaged across both groups. Thus, the CRS produced a similar result, again exceeding the ≥80% classification accuracy criterion, whereas the PRRS failed to meet that standard.

Given that the PRRS results were so jarringly different across the experiments, we reasoned that the problem was likely due to the imposition of the a priori 7/8 cutoff score on the data. Although it suited the Experiment 1 data just fine, in Experiment 2 it was too conservative, leading to too many CR participants being labeled as "random responders." To explore this hypothesis further, we conducted binary logistic regression analyses for each of the CRS and PRRS measures in both sets of experimental data. Specifically, we sought to examine whether empirically derived cutoffs generated by the logistic regression models would be different and more effective than the a priori cutoff scores we imposed on the data.

In both sets of data, two binary logistic regressions were conducted in which the criterion variable (responder group: RR or CR) was regressed on the either the CRS or PRRS as the predictor variable. As expected, results of all four regressions were significant. For the CRS, results showed that 2/3 was the best empirically based cutoff to accurately differentiate between CR and RR—the same as the theoretically derived cutoff. In contrast, results from

the PRRS logistic regressions were significant, but showed that the a priori 7/8 cutoff was not the optimal cutoff in either data set. In Experiment 1, 4/5 was shown to be a better empirical cutoff (correctly classifying 100% of CR participants and 100% of RR participants), whereas in Experiment 2, the best cutoff was 3/4 (correctly classifying 90% of CR participants and 97.88% of RR participants, producing an average accuracy rate of 93.93%). The smaller 3/4 cutoff was better in Experiment 2 because CR participants in that study produced a lower mean score than in Experiment 1.

In sum, these logistic regression data showed that an effective CRS cutoff score can be generated a priori using probability theory and applied reliably across data sets. In contrast, an effective a priori PRRS cutoff cannot be reliably applied across data sets. Rather, an empirical cutoff score needs to be generated for every data set it is used to optimize its discriminative power.

Discussion

The purpose of this investigation was to evaluate the effectiveness of a novel tool for identifying CR and RR in self-report inventory data. The CRS is a five-item variant of a traditional validity scale, which uses instructional item content and theoretically derived cutoff scores as its means to identify responders. Because CRs are assumed to follow testing instructions diligently, answering items as honestly and accurately as possible, we expected them to answer all of the CRS items correctly. In contrast, because RRs answer items indiscriminately, we expected them to account for all of the incorrect responses in the data, and only a very small proportion of items answered correctly that were due to chance. In our questionnaire, we used a 7-point scale; thus, the chance of a RR answering an item correctly was 14.29%. Given this rationale, we hypothesized that CRs would produce CRS total scores near the ceiling of the scale's range (i.e., 5) and RRs near the floor (i.e., 0). The large gap in expected scale scores would make it easy to discriminate between individual cases of conscientious and random responding. The other unique advantage of instructional items over the traditional variety used in infrequency scales stems from the fact that they can be objectively scored as correct or incorrect. Because of this difference, effective cutoff scores can be generated using probability theory and therefore eliminate the need for extensive normative testing. This would save test developers the laborious task of having to validate every single validity measure they create, and also allow test administrators the flexibility of being able change the CRS format depending on their particular testing requirements (e.g., by increasing or decreasing the number of its items or the size of its response-option scale). These were the main ideas behind the CRS when we designed it. This investigation was conducted to assess whether these lofty speculations were realistic.

Overall, the findings of this investigation were positive for the discriminative power and validity of the CRS. As predicted, CRs produced significantly larger CRS scores than RRs across experiments and these group differences were large in magnitude. The PRRS, a traditionally developed infrequency scale, which was administered alongside the CRS for comparative purposes, correlated positively and strongly with the CRS. For both measures, CRs produced scores toward the ceiling of the measures' scale range (5 for the CRS and 10 for the PRRS), whereas RRs produced mean scores near the scale floors (0 for both measures). Because the PRRS contains twice the number of items the CRS has, the average difference between the CR and RR groups' sum scores was larger for the PRRS and this produced a larger effect size. In Experiment 2, this difference was largely eliminated because CRs produced PRRS scores nearer the middle of the measure's scoring range.

The implication of this consistency is positive for the CRS. Given that across testing situations one can reliably expect CRs to produce a score near the ceiling and RRs near the floor, an a priori cutoff will be consistently effective at identifying responders. In these data, the theoretically derived 2/3 cutoff accurately discriminated between CR and RR about 93% of the time. Additional analyses with binary logistic regression models showed that the theoretically derived cutoff was identical to empirically derived cutoff scores from both sets of data. This agreement boosts the validity of our probability-based approach to generating cutoff scores.

Results for the PRRS were good, but less positive. Like the CRS, the PRRS produced large group differences between CR and RR, making distinguishing between them a fairly easy task. However, unlike the CRS, the size of the group difference in PRRS scores was inconsistent across studies. Moreover, the optimal empirical cutoff score changed from Experiment 1 to Experiment 2 and in neither study agreed with the a priori cutoff score of 7/8. The consequence of this was nicely demonstrated by the dramatic loss of discriminative power across studies. In Experiment 1, the a priori PRRS cutoff score produced an average classification accuracy of 93%, whereas in Experiment 2, its accuracy fell to just above 73%, failing to even meet the 80% classification standard. Consequently, one has to seriously doubt the utility of an a priori PRRS cutoff score, like the one we used in this investigation. The PRRS worked best using empirically based cutoffs. This means that after collecting PRRS data, an administrator should generate an equally large set of random data and run statistical analyses to identify the best empirical cutoff score. In sum, the effort required to use the PRRS effectively is far greater than it is to effectively use the CRS. With the CRS, a theoretically derived cutoff score can work reliably and effectively in a greater variety of testing scenarios. One has only to tally responders' scores and then assign them their appropriate responder labels. No normative testing is required.

Limitations and Future Research

Although the findings of this investigation were straightforward, producing large CRS score differences across responder conditions and nearly identical results across experiments, the utility of the CRS should be further evaluated using a wider variety of study designs and settings in which inventories are commonly used. For example, the CRS effective in forensic and psychiatric settings, where rates of random responding are highest (Archer et al., 2002; McNulty et al., 2003), may be somewhat different than it was in this investigation with student, non-disordered samples. Given that the CRS only requires respondents to follow simple instructions, at this point we believe that the CRS is safe for use in non-disordered samples and for research purposes. Caution should be exercised when using it outside of these groups or for individual assessment. In addition, the classification accuracy of the CRS should be examined at finer gradients of random responding, for example, in identifying responders who engage in random responding in only 25% of a questionnaire's items versus 100% of them. Research on the prevalence of random responding suggests that this form of intermittent random responding may account for the bulk of all random responding cases, as most responders admit to responding randomly to at least some of a questionnaire's items, but few report doing it to all of them (e.g., Baer, Ballenger, Berry, & Wetter, 1997).

Also, the classification accuracy of the CRS should be evaluated against multiple standards of comparison, such as the highly regarded validity scales of the MMPI series (the *F* Scale, and VRIN and TRIN scales; Butcher et al., 1989), and perhaps the completion times of online-administered questionnaires. In an online question-naire, for example, it is reasonable to assume that a 100-item inventory completed in 1 min is not the result of conscientious responding. We predict that responders who produce these and the types of abnormal responding patterns would also produce very low CRS scores.

Finally, there is an off chance that embedding validity scales like the CRS or PRRS in a questionnaire may exacerbate random responding by lowering the questionnaire's face validity. Face validity is defined as the extent to which item content seems appropriate for the purposes of a given testing situation (Holden & Jackson, 1979). For example, when assessing sadism, an item like "I enjoy hurting others" has higher face validity than the item "I would enjoy the occupation of a butcher." Because the CRS items instruct one how to respond, which is very different from what people expect to find in an inventory, their odd nature may sap the motivation of some responders to participate conscientiously. Perhaps, even the notion of being told what to do may motivate some individuals to respond incompliantly in a fit of psychological reactance (Miron & Brehm, 2006). With infrequency scales, their item content may seem so absurd in some cases that responders may feel ridiculous and put off in having to respond to them, which similarly may sap their motivation to act conscientiously. As far as we are aware, this hypothesis that random responding scales exacerbate random responding has not been experimentally examined and perhaps should be pursued in future research. We speculate, however, that given the prevalence and utility of some low-face-validity inventories (e.g., the MMPI series), if there was an effect here to find, it would be negligible in size and fully compensated by the positive effects of validity scales (i.e., being able to discriminate between CRs and RRs).

Conclusion

Due to the many costs associated with random responding scales, basic and applied psychologists rarely use them when administering inventories. This investigation aimed to remedy this situation with the introduction and preliminary validation of the CRS, a five-item measure that uses instructional item content to achieve this goal. Results across two experiments were compelling in that effect sizes were large, results were consistent across samples, and the CRS was accurate in classifying responders about 93% of the time. Simply put, by embedding the CRS items randomly throughout a questionnaire, researchers can use endorsements of the CRS items as reliable indicators of whether data were generated conscientiously and should be retained or whether they were produced randomly and should be deleted.

Appendix

Conscientious Responders Scale (CRS)

1. To answer this question, please choose option number four, "neither agree nor disagree."
2. Choose the first option—"strongly disagree"—in answering this question.
3. To respond to this question, please choose option number five, "slightly agree."
4. Please answer this question by choosing option number two, "disagree."
5. In response to this question, please choose option number three, "slightly disagree."

Note. In this investigation, the CRS was administered using a 7-point Likert-type scale. To use the CRS effectively, embed its items randomly throughout the length of a questionnaire, not all in a row or cluster. To prevent responders from being surprised or confused by the instructional nature of the CRS items, we added a line to our questionnaire's instructions that warned, "Some of the items will ask you to answer them in a particular way . . ."

Acknowledgment

We thank Cathy Faye, Lisa Fiksenbaum, and David Flora for their reviews of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

Notes

1. Psychological inventories contain items that cannot be answered in an objectively correct or incorrect manner, unlike educational or intelligence test items (e.g., 2 + 2 = ?). Personality-test item responses merely reflect the extent to which respondents agree or disagree with an item's content or how much it is true (e.g., I like to read).
2. Importantly, changes can be made to the number of the Conscientious Responders Scale (CRS) items or the size of its response-option scale without affecting its effectiveness at identifying conscientious responders. The flexibility of the CRS, or of any validity scale developed with instructional item content, stems from the fact that effective cutoff scores can be generated with ease using probability theory. This makes it considerably simpler to suit the CRS to one's needs and more likely to be used by testers.
3. Although we cannot be entirely sure that low CRS scorers are in fact random responders, we can be confident that these responders are unable or unwilling to follow simple testing directions and that alone is enough to invalidate their data.

References

Altemeyer, B. (1996). *The authoritarian specter*. Cambridge, MA: Harvard University Press.

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

Archer, R. P., Handel, R. W., Lynch, K. D., & Elkins, D. E. (2002). MMPI-A validity scale uses and limitation in detecting varying levels of random responding. *Journal of Personality Assessment*, *78*, 417-431.

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

Ben-Porath, Y. S., & Waller, N. G. (1992). Five big issues in clinical personality assessment: A rejoinder to Costa and McCrae. *Psychological Assessment*, *4*, 23-25.

Bentler, P. M., Jackson, D. J., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76*, 186-204.

Blascovich, J., & Tomaka, J. (1993). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (3rd ed., pp. 115-160). Ann Arbor, MI: Institute for Social Research.

Bobbio, A., Canova, L., & Manganelli, A. M. (2010). Conservative ideology, economic conservatism, and causal attributions for poverty and wealth. *Current Psychology*, *29*, 222-234.

Bruehl, S., Lofland, K. R., Sherman, J. J., & Carlson, C. R. (1998). The variable responding scale for detection of random responding on the Multidimensional Pain Inventory. *Psychological Assessment*, *10*, 3-9.

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

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment*, *15*, 223-234.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*, 596-612.

Davis, M. A., Andersen, M. G., & Curtis, M. B. (2001). Measuring ethical ideology in business ethics: A critical analysis of the Ethics Position Questionnaire. *Journal of Business Ethics*, *32*, 35-53.

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cogntive Therapy and Research*, *14*, 449-468.

Forsyth, D. R. (1980). A taxonomy of ethical ideologies. *Journal of Personality and Social Psychology*, *39*, 175-184.

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology*, *47*, 459-468.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103-129.

Manganelli Rattazzi, A. M., Bobbio, A., & Canova, L. (2007). A short version of the Right-Wing Authoritarianism (RWA) Scale. *Personality and Individual Differences*, *43*, 1223-1234.

McNulty, J. L., Forbey, J. D., Graham, J. R., Ben-Porath, Y. S., Black, M. S., Anderson, S. V., & Burlew, A. K. (2003). MMPI-2 validity scale characteristics in a correctional sample. *Assessment*, *10*, 288-298.

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

Miron, A. M., & Brehm, J. W. (2006). Reactance theory—40 years later. *Zeitschrift für Sozialpsychologie*, *37*, 9-18.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science results. *Frontiers in Psychology*, *1*, Article 220. doi:10.3389/fpsyg.2010.00220

Parker, W. D., & Adkins, K. K. (1995). A psychometric examination of the Multidimensional Perfectionism Scale. *Journal of Psychopathology and Behavioral Assessment*, *17*, 323-334.

Pettit, F. A. (1999). *Response sets in World Wide Web and paper-and-pencil personality questionnaires* (Unpublished doctoral dissertation). York University, Toronto, Ontario, Canada.

Pettit, F. A. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments, & Computers*, *34*, 50-54.

Pinsoneault, T. B. (2002). A Variable Response Inconsistency scale and a True Response Inconsistency scale for the Millon Adolescent Clinical Inventory. *Personality Assessment*, *14*, 320-330.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515-530.

Author Biographies

Zdravko Marjanovic is a social and personality psychologist at Thompson Rivers University in British Columbia, Canada. His research spans the areas of prosocial behavior, reactions to economic crisis, and psychological measurement.

C. Ward Struthers, Robert Cribbie, and **Esther R. Greenglass** are all professors of psychology at York University in Ontario, Canada. Their work includes various topics such as attributions and forgiveness, statistical analysis of psychological data, and coping amid crisis, respectively.