J Bus Psychol (2012) 27:99–114
DOI 10.1007/s10869-011-9231-8

Detecting and Deterring Insufficient Effort Responding to Surveys

Jason L. Huang · Paul G. Curran · Jessica Keeney · Elizabeth M. Poposki · Richard P. DeShon

Published online: 31 May 2011

Abstract
Purpose Responses provided by unmotivated survey participants in a careless, haphazard, or random fashion can threaten the quality of data in psychological and organizational research. The purpose of this study was to summarize existing approaches to detect insufficient effort responding (IER) to low-stakes surveys and to comprehensively evaluate these approaches.
Design/Methodology/Approach In an experiment (Study 1) and a nonexperimental survey (Study 2), 725 undergraduates responded to a personality survey online.
Findings Study 1 examined the presentation of warnings to respondents as a means of deterrence and showed the relative effectiveness of four indices for detecting IE responses: response time, long string, psychometric antonyms, and individual reliability coefficients. Study 2 demonstrated that the detection indices measured the same underlying construct and showed the improvement of psychometric properties (item interrelatedness, facet dimensionality, and factor structure) after removing IE respondents identified by each index. Three approaches (response time, psychometric antonyms, and individual reliability) with high specificity and moderate sensitivity were recommended as candidates for future application in survey research.
Implications The identification of effective IER indices may help researchers ensure the quality of their low-stake survey data.
Originality/value This study is a first attempt to comprehensively evaluate IER detection methods using both experimental and nonexperimental designs. Results from both studies corroborated each other in suggesting the three more effective approaches. This study also provided convergent validity evidence regarding various indices for IER.

Keywords Careless responding · Random responding · Inconsistent responding · Online surveys · Data screening

An earlier version of this manuscript was presented at the Annual Conference of Academy of Management, Anaheim, CA, August 2008.

J. L. Huang (✉)
Department of Psychology, Wayne State University, Detroit, MI 48202, USA
e-mail: jasonhuang@wayne.edu

P. G. Curran · J. Keeney · R. P. DeShon
Department of Psychology, Michigan State University, East Lansing, MI 48824, USA
e-mail: curranp1@msu.edu

J. Keeney
e-mail: jkeeney@msu.edu

R. P. DeShon
e-mail: deshon@msu.edu

E. M. Poposki
Department of Psychology, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA
e-mail: epoposki@iupui.edu

Datasets in social science research are prone to contain numerous errors representing inaccurate responses, inaccurate coding, or inaccurate computation. It is widely recommended that researchers screen the data to identify and correct these inaccurate observations before their use in modeling and hypothesis testing procedures (Babbie 2001; Hartwig and Dearing 1979; Kline 2009; Smith et al. 1986; Tukey 1977; Wilkinson and the Task Force on Statistical Inferences 1999). Fortunately, a number of techniques exist to facilitate the treatment of simple data errors,

such as out of range data and duplicate cases (e.g., DiLalla and Dollinger 2006; O'Rourke 2000), as well as more complex data problems, such as outliers and missing values (e.g., Stevens 1984; van Ginkel and van der Ark 2005).

Data obtained using online or traditional paper-and-pencil surveys are particularly susceptible to a subtle yet insidious threat to data quality resulting from participants who are not sufficiently motivated to provide accurate responses. Often labeled as random, careless, or inconsistent responding (McGrath et al. 2010); this type of unmotivated response behavior can arise in any number of different contexts, including employee surveys, customer surveys, training evaluations, and so on. The identification and removal of inconsistent responding may enhance the criterion-related validity of measures (McGrath et al. 2010).

Logic can certainly be used to detect some of these problematic responses. For example, participants who claim to be in a significant relationship longer than they have been alive are relatively easy to detect (Wilkinson and the Task Force on Statistical Inferences 1999). Other methods have been developed such as using items that have apparent correct/incorrect answers (Hough et al. 1990) or examining within-person correlations between items with opposite scoring (Seo and Barrett 2007). Researchers and practitioners have expressed a growing interest in using certain indices for careless responding to ensure the quality of their survey data (e.g., Behrend et al. 2011; Curran et al. 2010; Johnson 2005; Meade and Craig 2011). However, various detection approaches have yet to receive simultaneous evaluation, and thus little is known on the relative effectiveness of these approaches.

The purpose of this study is threefold. First, after summarizing existing approaches in the literature, we conduct a comprehensive empirical evaluation of the methods that survey administrators can employ to deal with responses provided with insufficient effort. Second, we estimate the extent to which inclusion of unmotivated responses can affect the psychometric properties of measures. Third, we examine the extent to which different indices tap into an underlying construct. We conducted two survey studies to address these purposes in sequence.

Insufficient Effort Responding

Various labels have been used in the literature to refer to the phenomenon whereby participants are unmotivated to complete a survey measure as instructed. The term "random responding" is used in the majority of studies (e.g., Charter 1994; Pinsoneault 2007; Thompson 1975), despite the lack of evidence showing that unmotivated respondents select response options randomly. On the contrary, some respondents may opt to endorse the same response option repeatedly, leading to a nonrandom response pattern. Another group of labels focus on respondents' need to read, understand, and respond to item content, including content-independent responding (Evans and Dinning 1983), non-contingent responding (Marsh 1987), and content nonresponsivity (Nichols et al. 1989). Yet another group of labels are premised on the congruency of responses, such as inconsistent responding (Greene 1978) and variable responding (Bruehl et al. 1998). Finally, careless responding (Haertzen and Hill 1963) suggests occasional misresponding because of inattentiveness.

To provide a comprehensive depiction of the phenomenon of interest, we propose the label of *insufficient effort responding* (IER), defined as a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses. Insufficient effort responding underscores the cause of the response behavior without presupposing specific patterns or outcomes. Thus, IER includes random endorsement of response options as well as nonrandom repeated endorsement of the same response option. IER may vary in its intentionality—ranging, for example, from inadvertent misinterpretation of negatively keyed items to intentional disregard for item content, concealing one's true opinion. The emphasis on the lack of effort from the respondents also differentiates IER from effortful response distortion, such as positive/negative impression management (McGrath et al. 2010).

Review of Existing Approaches

We organize the literature review into four broad categories of approaches: (i) infrequency, (ii) inconsistency, (iii) pattern, and (iv) response time.

Infrequency Approach

The infrequency approach uses items on which most attentive respondents, if not all, will provide the same response. For example, Beach (1989) created items that would have the same responses from all individuals who read and understand the wording (e.g., "I was born on February 30th."). Endorsement of any of these improbable factual statements indicates IER. In a similar vein, Green and Stutzman (1986) included job-irrelevant bogus tasks in the context of a job analysis questionnaire.

Although studies have shown the infrequency scales' effectiveness at detecting random responses (e.g., Baer et al. 1999; Bagby et al. 1991), the infrequency approach may not be appropriate to detect IER, because infrequency scales can confound IER with impression management and faking (e.g., Butcher et al. 1989, 1992; Rosse et al. 1999).

Inconsistency Approach

The inconsistency approach assumes that unmotivated respondents provide inconsistent responses. This approach typically uses matched item pairs and compares the response on one item to the response on the other item (Pinsoneault 1998). Item pairs are created in three ways, including (i) direct item repetition, (ii) rational selection, and (iii) empirical selection.

Researchers have incorporated repeated items into surveys to measure inconsistency. Buechley and Ball (1952) used 16 pairs of identical repeated items (*Tr* scale) from the 566-item MMPI, with a large number of items between them. Similarly, Wilson et al. (1990) included repeated task statements in job analysis questionnaires and detected the respondents who provided inconsistent endorsement of task statements. Lucas and Baird (2005) also recommend survey researchers to design very similar questions in different places of a questionnaire to check against IER.

Empirical methods are also used to select item pairs for inconsistency scales (e.g., Bruehl et al. 1998; Greene 1978). For example, Schinka et al. (1997) selected ten-item pairs into an Inconsistency Scale (INC) for the NEO Personality Inventory-Revised (NEO-PI-R; Costa and McCrae 1992) using an arbitrary cutoff of $r > .40$. They found that the mean INC score was significantly higher for computer-generated random responses than for normal responses.

Inconsistency scales have generally been effective at identifying random responses generated by participants with partial or no access to the questionnaire (Baer et al. 1997, 1999; Berry et al. 1991; Wetter et al. 1992), as well as random responses generated by computer algorithm (Archer and Elkins 1999; Morey and Hopwood 2004). However, research using normal instruction has yielded mixed results in terms of the scales' effectiveness (e.g., Archer et al. 1998; Kurtz and Parrish 2001; Piedmont et al. 2000).

Johnson (2005) applied two variants of the inconsistency approach to detect careless respondents to online surveys, including (i) Goldberg's psychometric antonyms and (ii) Jackson's (1976) individual reliability. Whereas psychometric antonyms are selected empirically, individual reliability relies on item pairs from the same scales. Instead of using the sum of absolute differences between pairs of items, both methods split the pairs of items into two halves and use the correlation between the two halves to indicate normal responding. The effectiveness of these two methods has yet to be evaluated.

Response Pattern Approach

The response pattern approach relies on the pattern of response options endorsed by the respondent to indicate IER. Costa and McCrae (2008) suggested using a long string of the same response option (e.g., responding 15 "Strongly Agree" in a row) to indicate IER. From a sample of 983 normal respondents thought to be cooperative, they recommended a cutoff for each response option for the NEO-PI-R. When the number of consecutive identical responses exceeds the associated cutoff, the protocol can be suspected for IER. The effectiveness of the response pattern approach remains unstudied.

Response Time Approach

The response time approach assumes shortened response time for IER than for normal responding because of the absence of cognitive processing. Response time has been used in computer-based low-stakes educational testing, where students complete academic performance tests without concerns for the results. Using computer-based assessments from 472 new college students, Wise and Kong (2005) found that response time converged significantly with self-reported effort and person-fit statistics in indicating response effort. They also showed that the average test score for IER cases detected by response time was similar to the score predicted by chance and much lower than the average score from normal responding. Wise and DeMars (2006) further showed that the incorporation of response time in the item response theory (IRT) model yielded more accurate proficiency estimates.

Unanswered Questions

The review of IER approaches leaves three unanswered questions. First, aside from the infrequency approach, the effectiveness of the other detection approaches has yet to be evaluated with an appropriate design. For the inconsistency approach, most evaluation studies were based on the assumption that unmotivated respondents provide *random* responses. It remains unknown whether this approach is effective in detecting insufficient effort responses in general. On the other hand, the response pattern and response time approaches have not been investigated in the survey literature. Second, survey administrators have no knowledge as to which approach is most effective in detecting IER. Comparison of the relative effectiveness of detection approaches requires examination of detection indices within a single study. Third, all of the approaches reviewed in the literature deal with IER after it has occurred, and it remains unknown whether survey administrators can deter IER. Although we could find no study that deterred IER, warnings have been shown to effectively reduce positive response distortion in personality measures (Dwight and Donovan 2003). Thus, warning participants about the consequences of IER is an unexplored yet potentially useful means of dealing with IER.

To address these three questions, we examined the detection approaches under different survey conditions. Researchers have evaluated validity scales by experimentally inducing the response set of interest, such as random responding (e.g., Berry et al. 1991). Likewise, we employed this analog paradigm and manipulated survey conditions to induce IER in Study 1. We first evaluate the effectiveness of each index by comparing index scores between the IER conditions and the control condition—if an index is effective, it results in significantly higher rate of IER detection in the IER conditions than in the non-IER condition. We also assess the effect of warning by comparing IER index scores between the warning condition and the non-warning condition. If the warning is effective, it results in lower rate of IER detection by various indices in the warning condition than in the non-warning condition. In addition, we hypothesize that the IER index scores significantly correlate with each other, providing evidence for convergent validity. Finally, after applying cutoffs, the relative effectiveness of all indices can help identify IER indices that perform better than others.

Study 1

Method

Participants

The sample comprised 380 undergraduate students at a large Midwestern university (74% female, mean age = 21 years). A subset of respondents ($N = 39$) were students of one of the present authors who volunteered to participate and were thought to be highly motivated to respond accurately and follow directions. The remaining respondents were recruited from the psychology participant pool in the same or next semester. On completion of the study, all respondents were shown a debriefing statement online and received partial course credit.

Measure

The survey instrument consisted of 300 items from the International Personality Item Pool (IPIP; Goldberg 1999). All items were administered on a 5-point Likert scale from 1 (*very inaccurate*) to 5 (*very accurate*). The IPIP-NEO items are reliable measures of 30 personality facets (e.g., Cheerfulness, Trust, Anxiety) under the Big Five framework (Goldberg 1999).

Operationalization of Detection Approaches

Inconsistency Approach The inconsistency approach was applied as *psychometric antonym* and *individual reliability*, as described in Johnson (2005). To obtain psychometric antonyms, inter-item correlations were first computed for the entire sample, and 30 unique pairs of items with the highest negative correlation were selected. The correlation between the 30 pairs of items for a normal respondent was expected to be highly negative. The correlation was then reversed, with a lower score indicating a higher probability of IER.

Individual reliability is based on the premise that items on the same scale are expected to correlate with each other for each individual. Items on the same scales were first separated into odd-numbered and even-numbered halves, and half-scale scores were calculated. A correlation was computed between all 30 odd-numbered half-scale scores and all 30 even-numbered half-scale scores within each individual. The correlation was then corrected for decreased length using the Spearman–Brown formula. Low individual reliability indicated IER.

Response Pattern Approach We operationalized the response pattern approach with the *long string index*. A Java program was written to examine if the string of the same response option in a protocol exceeded predetermined cutoffs. Costa and McCrae (2008) found that none of their 983 cooperative participants selected the same response option more than 6, 9, 10, 14, and 9 times for the response options from *strongly disagree* to *strongly agree*, respectively. They recommended using these cutoffs to detect IER. We used the same values as initial cutoffs to categorize responses. To be consistent with the other indices, the long string index was scored 0 when IER was suspected to be present and 1 otherwise.

Response Time Approach We used *page time* for the response time approach. Page time is the time between the initiation and submission of each survey page online. The survey system recorded page time for each respondent. Extremely short page time signaled IER.

Survey Arrangement

The survey was administered in two halves, with a survey instruction preceding each half. To balance the IPIP items in the two halves, we divided each facet into two 5-item parts and randomly distributed them into either the first or the second half of the survey. Each half of the survey contained six Web pages, with each page containing 25 IPIP items and one extra item that was excluded from this study.[1]

[1] The extra item on each page, named check item, represented a failed attempt to improve on the infrequency approach. Each check item instructed participants to select a particular response option, e.g.,

Table 1 Experimental conditions (warning by IER) and sample size in Study 1

Cell number	Condition for the first half (150 items)	Condition for the second half (150 items)
Cell 1 ($n = 39$)	Warning	Warning
Cell 2 ($n = 57$)	Warning	Cautionary IER
Cell 3 ($n = 55$)	Warning	Outright IER
Cell 4 ($n = 84$)	Normal instruction	Normal instruction
Cell 5 ($n = 64$)	Normal instruction	Cautionary IER
Cell 6 ($n = 81$)	Normal instruction	Outright IER

Manipulations

To induce the response sets, we adopted a modified $2 \times 2 \times 3$ mixed design with *warning* (warning vs. normal) and *IER conditions* (control, IER with caution, and outright IER) as between-subject factors and *time* (first vs. second half of survey) as a within-subject factor. The conditions are summarized in Table 1.

The following scheme was used to assign participants into specific cells: The students from the class of one of the present authors ($N = 39$) were assigned to Cell 1, whereas another 112 respondents from the same semester were randomly assigned to either Cell 2 or Cell 3. The purposeful assignment of motivated respondents into Cell 1 was used to check against potential false-positive identification. The remaining participants ($N = 229$) recruited from the following semester were assigned in a random fashion into Cells 4, 5, and 6.

Because there are two instructions for each individual for the two halves of the survey, for clarity of description, we refer to the answers an individual provided to either half of the survey as a "*protocol*" (see Kurtz and Parrish 2001).

Warning Versus Normal Instruction

Warning versus normal instruction manipulation appeared at the beginning of the first half of the survey. All participants received the instructional set that commonly precedes administration of the IPIP—what we refer to as *normal instruction* (e.g., "there are no correct or incorrect answers…Describe yourself as you honestly see yourself"). In Cells 1, 2, and 3, in addition to normal survey instruction, participants were warned that "sophisticated statistical control methods" would be used to check for validity of responses and that responding without much effort would result in loss of credits. Regardless of the experimental conditions, all students received the full amount of extra course credits for their participation.

Insufficient Effort Responding Conditions

IER between-subjects manipulations were employed at the beginning of the second half of the survey. Specifically, participants were instructed to (a) continue the instructions from the first half of the survey (Cells 1 and 4; labeled *Non-IER*), (b) respond without much effort but "pretend that you want your laziness in filling out this survey to remain undetected" (Cells 2 and 5; labeled *Cautionary IER*), or (c) respond without effort with no risk of penalty: "in fact, we request that you do so" (Cells 3 and 6; labeled *Outright IER*). The last two instructional sets (b and 0) were intended to induce two levels of IER—(b) simulated situations where there might be consequences for IER, whereas (c) resembled situations where there was no threat whatsoever.

Manipulation Check

On the completion of the entire experiment, participants in Cells 4, 5, and 6 were asked two additional open-ended questions that served as manipulation checks. Specifically, the two questions asked "What strategy did you use to fill out the first/second half of the survey?"

Results

Manipulation Check

We first examined the manipulation checks to ensure the quality of IER manipulations. The fourth author blindly coded participants' open-ended answers to the manipulation check as indicative of normal responding (coded as 0) or IER (coded as 1), as most of the responses were not specific enough to be sorted into the three IER manipulation conditions. Of the 229 responses for each half of the survey, 213 for the first half and 206 for the second could be coded. Interrater agreement was estimated to be 98.46%, after the third author blindly coded a random subset of 32 responses from each half. The mean of the coded answers indicated general absence of IER for the first half and presence of IER in the second half (for Cells 4, 5, and 6, respectively, Ms = .01, .07, and .01 for the first half and .14, .85, and .90 for the second half). Logistic regression with dummy-coded Cautionary IER and Outright IER

Footnote 1 continued
"Please select Moderately Inaccurate for this item", and selecting any other response category would indicate IER. We excluded check items from this study because—inconsistent with the survey instruction, the manipulation check, and the other indices—the check item index flagged an unusually high rate of IER, even in the group of motivated respondents in Cell 1 of Study 1. We suspect some respondents may have viewed the check items as a measure of personality rather than an instruction set. Additional analysis also revealed significant trait influence on responses to the check items, after controlling for the other IER indices.

Table 2 Scores on indices by IER condition over time in Study 1

Index	Survey	IER manipulation		
		Non-IER (Cells 1 and 4)	Cautionary IER (Cells 2 and 5)	Outright IER (Cells 3 and 6)
Page time (average)	First half	128.79 (61.47)	141.16 (80.25)	147.06 (75.83)
	Second half	124.26^{a} (61.47)	79.75^{b} (35.52)	77.39^{b} (56.17)
Psychometric antonym	First half	0.64^{a} (0.20)	0.61^{ab} (0.23)	0.54^{b} (0.24)
	Second half	0.54^{a} (0.19)	0.27^{b} (0.31)	0.15^{c} (0.26)
Individual reliability	First half	0.69 (0.25)	0.70 (0.22)	0.67 (0.30)
	Second half	0.72^{a} (0.19)	0.39^{b} (0.45)	0.12^{c} (0.54)
Long string reversed[A]	First half	0.98 (0.16)	0.97 (0.18)	0.93 (0.25)
	Second half	0.98^{a} (0.16)	0.87^{b} (0.34)	0.67^{c} (0.47)

Note SD in parenthesis. Higher scores indicate lower probability for IER. Pairwise comparison conducted within each half. On each row, means with different superscripts are significantly different from each other

[A] Comparisons based on results from binary logistic regression

conditions as predictors reveals no significant effect for the first half, $\chi^2(2) = 4.06$, $p = .13$, and significant effect for the second half, $\chi^2(2) = 112.17$, $p < .001$. As expected, no between-group differences existed before the IER manipulations. After the IER manipulations, participants in both IER conditions were more likely to engage in IER than those in the non-IER condition.

Indicators of IER

We computed scores on the four IER indices for each protocol. To be consistent with the existing use of psychometric antonyms and individual reliability, the scorings of all indices were such that *higher scores indicated lower probability for IER*. A 2 (time) × 3 (IER condition) mixed ANOVA[2] was performed on each continuous index to examine if scores changed over time within individuals and differed across IER conditions. Descriptive statistics and results for post hoc comparisons are presented in Table 2.

Page Time Index The effect of page time was first examined on a per page basis. A mixed ANOVA yielded significant within-subjects, $F(11,4147) = 35.96$, $p < .001$, and between-subjects effects, $F(2,377) = 3.35$, $p < .05$, as well as a significant interaction that qualified both main effects, $F(22,4147) = 7.44$, $p < .001$. Pairwise comparison between IER conditions on each page showed that none of the page time differed significantly across conditions from page 1 to page 6, and page time for cautionary IER and outright IER was lower than non-IER from page 7 to page 12. When average page time was computed for each protocol, the mixed ANOVA revealed the same pattern of results (see Table 2). To use the level of specificity afforded by page time in detecting IER, we decided to construct the page time index for each protocol using the average of the lowest two page time.

Psychometric Antonym Index Psychometric antonyms were identified based on protocols from the non-IER condition (Cells 1 and 4). On each half of the survey, 30 pairs of items that had the highest negative correlations were selected. The correlations ranged from −.76 to −.51 in the first half, and from −.70 to −.48 in the second half. A psychometric antonym score was computed for each protocol by correlating the two sets of items within person. Psychometric antonym scores could not be computed for 18 protocols in the second half because of lack of variance in responses (e.g., answering all 3 s). Rationally speaking, it would be very rare for a non-IER participant to respond to a large number of survey items with the exact same response category, so the lack of variance in response likely indicated IER. Indeed, all these cases were either from cautionary IER or outright IER condition. Because we did not hypothesize such occurrence a priori, we took the conservative approach to exclude them from analyses in Study 1.

The same mixed ANOVA revealed significant effects of time, $F(1,354) = 291.84$, $p < .001$, and IER condition, $F(2,354) = 52.30$, $p < .001$, as well as the time by IER condition interaction, $F(2,354) = 28.31$, $p < .001$. Pairwise comparison revealed that the non-IER condition yielded higher scores than the outright IER condition in the first half, and all conditions were significantly different from each other in the second half in the expected order.

Individual Reliability Index Individual reliability index was calculated for each protocol using the aforementioned procedure. Thirteen protocols in the IER conditions in the

[2] When severe departure from homogeneity of variance occurred, ANOVA results were verified using pairwise unequal variance *t*-tests. For all repeated ANOVA, the Greenhouse–Geisser adjusted *p* value was reported when Mauchly's W test for sphericity was significant.

Table 3 Correlations between indices of insufficient effort responding in Study 1

Index	Page time[a]	Psychometric antonym	Individual reliability	Long string reversed
Page time[a]	1.00	.55***	.59***	.24***
Psychometric antonym	.22***	1.00	.69***	.16**
Individual reliability	.36***	.62***	1.000	.28***
Long string reversed	.09	.15**	.15**	1.00

Correlations below the diagonal are from the first half of the survey ($n = 380$); Correlations above the diagonal are from the second half of the survey ($n = 367$–380 because of missing values for psychometric antonym and individual reliability)

** $p < .01$; *** $p < .001$

[a] Page time was log transformed before this analysis

second half did not show any variance for the computation of this index and thus were excluded from the analyses in Study 1.

The mixed ANOVA yielded significant effects for time, $F(1,364) = 143.14$, $p < .001$, IER condition, $F(2,364) = 38.89$, $p < .001$, and the interaction, $F(2,364) = 54.73$, $p < .001$. Through pairwise comparison, the interaction was decomposed: none of the IER conditions differed in the first half, and the scores differed in the expected order in the second half.

Long String Index A logistic regression with dummy-coded cautionary IER and outright IER with the long string index as outcome was not significant for the first half, $\chi^2(2) = 3.06$, $p = .22$ and was significant for the second half, $\chi^2(2) = 49.16$, $p < .001$. For responses in the second half, the predicted probability of IER/non-IER as indicated by the long string index differed significantly across all three conditions in the expected order. Matched-sample nonparametric McNemar tests revealed no difference in long string index across two halves of the survey for No IE condition, $\chi^2(1) = 1.00$, $p = 1.00$, and significant increase in both cautionary, $\chi^2(1) = 7.56$, $p = .006$, and outright IE conditions, $\chi^2(1) = 26.63$, $p < .001$.

Correlation Among Indices

Correlations among indices for each half of the survey were calculated and presented in Table 3. Overall, the results indicate that those who scored high on one of the indices tended to score high on the other indices, providing evidence for convergent validity.

Effect of Warning

To assess the effect of warning, we conducted an independent samples t test between warning conditions (Cells $1 \sim 3$) and normal conditions (Cells $4 \sim 6$) in the first half for each index (Table 4). Significant differences were found for all indices except page time. To remove the potential confound of high motivation in Cell 1, the same analyses were run comparing Cells $2 \sim 3$ with Cells $4 \sim 6$ and the same pattern of results emerged.

Cutoff Scores

After finding significant mean differences on IER index scores between normal and IER conditions, we proceeded to create three sets of cutoffs to classify protocols into normal versus IER categories: The first set of cutoffs was based on past researchers' suggestions as well as logic, and the second and third sets were derived empirically to hold

Table 4 Effect of warning on indices in the first half in Study 1

Index	Warning manipulation	*N*	*M*	*SD*	*t*	*p*	Cohen's *d*
Page time	Warning	151	104.68	30.66	1.27	.20	.13
	Normal	229	99.84	39.47			
Psychological antonym	Warning	151	0.65	0.19	3.68	<.001	.35
	Normal	229	0.57	0.25			
Individual reliability	Warning	151	0.75	0.15	4.67[b]	<.001	.43
	Normal	229	0.64	0.30			
Long string reversed[a]	Warning	151	0.99	0.12	2.60	.01	.25
	Normal	229	0.94	0.24			

Note Warning = Cells 1–3; Normal = Cells $4 \sim 6$. Higher scores indicate lower probability for IER

[a] Binary logistic regression indicated a similar result, $\chi^2(1) = 5.82$, $p = .02$

[b] Unequal variance t-test result was reported as Levene's test for equality of variances indicated heterogeneity of variance

constant the number of protocols detected in the normal condition to allow for a meaningful comparison across indices.

Existing/Rational Cutoffs We followed Johnson (2005) for the cutoffs for psychometric antonym coefficient (−.03, produced from a simulation of "24000 pseudo-random cases", p. 118), individual reliability (.30), and long string (6, 9, 10, 14, and 9, for each response category from "*very inaccurate*" to "*very accurate*"). All protocols that did not receive values for the psychometric antonym index or the individual reliability index were classified as IER. Because no existing cutoff was available for page time—a situation likely encountered by applied researchers and practitioners—we decided to use an educated guess as the starting value. Based on our understanding of the survey instructions and items, we determined that it was unlikely for participants to respond to survey items faster than the rate of 2 s per item, or 52 s per page. To avoid screening out a valid protocol simply because the respondent happened to respond quickly to a few items on a page, we adopted a more conservative cutoff and classified a protocol as IER when it contained two pages with less than 52 s per page.

To establish each index's effectiveness, we calculated two common diagnostic statistics, (i) specificity and (ii) sensitivity (Hsu 2002), for each index using each cutoff. For a diagnostic test for a particular attribute, specificity refers to "the proportion of people without the attribute who are correctly labeled by the test" (i.e., true-negative rate), and sensitivity refers to "the proportion of people who have the attribute who are detected by the test" (i.e., true-positive rate) (Streiner 2003, pp. 210–211). In the context of IER, *specificity* indicates the proportion of protocols correctly not flagged in the normal conditions (i.e., all cells in the first half and Cells 1 and 4 in the second half), whereas *sensitivity* indicates the proportion of protocols correctly flagged as IER in the IER conditions (i.e., Cells 2, 3, 5, and 6 in the second half). The specificity and sensitivity for the first set of cutoffs are presented in Table 5. The examination of cutoffs showed that the specificity of indices ranged from .94 to .99, whereas their sensitivity ranged from .20 to .46. Thus, the indices with this set of cutoffs picked up a fraction of normal protocols while identifying a moderate proportion of insufficient effort protocols.

Empirically Derived Cutoffs We proceeded to empirically develop two sets of cutoffs, controlling for specificity on .95 and .99, respectively. Applying these cutoffs would identify 5% and 1% of normal protocols as IER. Table 5 includes the sensitivity and specificity for these two sets of cutoffs. When setting the cutoff with 95% specificity, the page time index and the psychometric antonym index were able to identify about 50% of IER protocols, whereas the individual reliability index and the long string index identified 40% and 30% of IER protocols. When using the cutoff with 99% specificity, the page time index and the psychometric antonym index identified around 25% of IER protocols, whereas the individual reliability index and the long string index identified around 20%.

Table 5 Cut scores, sensitivity, and specificity in Study 1

Index	Cut score	Specificity	Sensitivity	Sensitivity (cautionary IER)	Sensitivity (outright IER)
Existing/rational cutoffs					
Page time	52	.96	.45	.35	.54
Long string[a]	6,9,10,14,9	.96	.24	.13	.33
Psychometric antonym	−0.03	.99	.20	.16	.23
Individual reliability	0.30	.94	.46	.33	.58
Cutoffs set for 95% specificity					
Page time	56	.95	.49	.36	.60
Long string[a]	7,7,12,10,8[b]	.96	.30	.25	.35
Psychometric antonym	0.15	.95	.45	.37	.52
Individual reliability	0.22	.95	.40	.28	.50
Cutoffs set for 99% specificity					
Page time	38	.99	.24	.17	.30
Long string[a]	11,8,25,16,10[b]	.99	.19	.12	.26
Psychometric antonym	0	.99	.25	.19	.30
Individual reliability	−0.22	.99	.16	.08	.24

[a] The cut scores reported represent the cut scores for each response category from 1 (*very inaccurate*) to 5 (*very accurate*)

[b] Each response category identified about equal number of normal responses as IER

Discussion

Results of Study 1 demonstrated the effectiveness of the IER indices. First, each index indicated more IER protocols in the IER conditions than in the non-IER conditions. Second, the indices showed convergent validity as they correlated significantly with one another in general. Third, the use of warning reduced the severity of IER as suggested by scores on three out of four indices. Finally, when the percentage of IER protocols detected from the non-IER condition was held constant across indices, the page time index and the psychometric antonym index identified the most protocols from the IER conditions.

Results from Study 1 provided the basis for further investigation regarding the following research questions: (a) Given their modest to moderate intercorrelations, do different indices capture something in common? (b) To what degree do the different indices overlap in their identification when cutoffs are applied? and (c) Can the exclusion of IER protocols lead to changes in observed psychometric properties? Whereas research questions (a) and (b) have not been investigated in the literature, research question (c) has been studied in scales with negatively keyed items through simulated data. Using exploratory factor analysis, Schmitt and Stults (1985) found that 10% simulated careless responses may artificially increase observed scale dimensionality by introducing a factor composed of negatively keyed items. A similar finding by Woods (2006) showed that 10% of simulated careless responses to negatively keyed items could lead to rejection of a one-factor model in confirmatory factor analysis (CFA). This study goes beyond those two studies by studying naturally occurring IER.

Study 2 was conducted under normal survey instructions to address these research questions. In addition, several studies have used participants' self-report of IER (Baer et al. 1997, 1999; Berry et al. 1992; Costa and McCrae 1997), which we explored as another potential IER index through the use of a postsurvey self-report measure of effort.

Study 2

Method

Participants

The sample comprised 345 undergraduate students from the same university (68% female, mean age = 24.9 years), recruited from the general psychology participant pool. Study 2 sample contained a larger proportion of non-traditional students than Study 1, with 59% of the sample ranging from 25 to 34 years old. None of the participants who participated in Study 1 was included in Study 2. Students participated in the survey online in exchange of extra course credit. We conducted Study 2 at the end of a regular semester, at a time that we suspected a slightly higher rate of IER because of a heightened need to obtain extra credits before the semester finished.

Measures

The survey format and instructions of Study 2 were the same as Cell 4 in Study 1 (no warning, normal instruction), with the following exception: After completing all survey items, respondents were informed that they had finished the main survey and were asked to fill out three optional items regarding their effort in responding *for each half* of the survey. The items include "I didn't pay much attention to what the questions actually meant," "I filled out the questions WITHOUT thinking about myself," and "I responded carelessly to the questions," administered on the same 5-point Likert scale as the rest of the items.

Results

Cronbach's alpha for the postsurvey self-report of IER was .79 and .85 for both halves. To be consistent with the other indices, scale scores were computed such that a higher score would indicate more effortful responding. All participants responded to this measure for Survey Half 1, $M = 1.69$, $SD = 0.69$, whereas 341 participants (99%) responded to the measure for Survey Half 2, $M = 1.75$, $SD = 0.73$.

We first computed scores on all IER indices in the same way described in Study 1, with the dichotomously scored long string index computed using the cutoff established by 95% specificity in Study 1. Informed by Study 1 findings, we assigned a value of -1 to four cases to indicate the likely IER behavior when the psychometric antonym or individual reliability index could not be computed because of a lack of variance in response. Correlations among indices are presented in Table 6. In general, the relationships among page time index, psychometric antonym index, and individual reliability index were higher than those among the other indices.

To ascertain whether the indices measured a common construct, we conducted an exploratory factor analysis with principal axis factoring extraction on all indices except the dichotomous long string index. For each half of the survey, both the Kaiser criterion and the scree plot pointed to a one-factor solution, with the factor explaining 59% and 58% of variance in the indices in the two halves. In each half of the survey, factor loadings for psychometric antonym index, individual reliability index, and page time

Table 6 Correlations between indices of insufficient effort responding in Study 2

Index	Page time[a]	Psychometric antonym	Individual reliability	Long string reversed[b]	Self-report effort
Page time[a]	1.00	.40	.56	.38	.28
Psychometric antonym	.41	1.00	.68	.19	.34
Individual reliability	.53	.69	1.00	.28	.30
Long string reversed[b]	.39	.36	.38	1.00	.18
Self-report effort	.32	.35	.37	.22	1.00

Note $N = 345$. All correlations are significant at $p < .001$. Correlations from the first half of the survey are below the diagonal. Correlations from the second half of the survey are above the diagonal

[a] Page time was log transformed before this analysis

[b] Long String was calculated using the 95% specificity cutoff and then reversed to be consistent with the other indices in scoring direction

index were all above .60, whereas the loading for post-survey self-report was above .40. The exploratory factor analysis results suggest that these four indices indeed captured a common underlying response style.

We examined the impact of potential IER on two observed psychometric properties of scales: (a) facet-level item interrelatedness, indicated by Cronbach's alpha (Cortina 1993); and (b) facet unidimensionality, indicated by the eigenvalues for the first and the second factor in exploratory factor analysis. In addition, we also explored the impact of IER on factor structure at the Big Five factor level, indicated by fit indices reported from CFA on each of the Big Five factors.

To prepare datasets for the analyses, we first identified IER using the 95% and 99% specificity cutoffs developed in Study 1. In addition, we used a score of 3 (corresponding to "*neither inaccurate nor accurate*") as the rational cutoff for self-report effort. If the protocol for either half of the survey provided by an individual was identified by an index, the respondent would be flagged as an IER case. For each cutoff associated with an index, a trimmed dataset was created such that all respondents detected with this cutoff were removed from this dataset. Thus, a total of 11 datasets served as input for the following analyses, including the full sample and ten trimmed samples. The degree of overlap in the individual cases identified by each combination of index and cutoff is presented in Table 7.

We examined the impact of IER on average item interrelatedness by comparing Cronbach's alphas for all 30 facets before and after removal of IER. In general, the removal of suspected IER cases resulted in a slight increase of Cronbach's alpha for the scales. We conducted paired sample *t* tests examining whether there was significant change in alphas between each trimmed sample and the full sample, treating each facet as a case. The increase of alpha because of the removal of suspected IER protocols was significant for all indices (Table 8).

Because Cronbach's alpha indicates item interrelatedness but not scale unidimensionality (Cortina 1993), we further conducted exploratory factor analyses with principal axis factoring extraction to examine the impact of IER on the unidimensionality of each facet. Eigenvalues were computed for the first and second factors for the full sample as well as each trimmed sample. For each factor analysis, a large eigenvalue for the first factor and a small eigenvalue for the second factor indicate unidimensionality. Overall, larger first-factor eigenvalues and smaller second-factor eigenvalues were obtained after the removal of suspected IER with 99% specificity (Table 8). Analysis on trimmed samples created with 95% specificity (available from the first author on request) yielded a similar pattern of findings on both Cronbach's alpha and the eigenvalues.

To illustrate that the removal of IER could lead to a different interpretation of scale dimensionality, we produced the scree plot for exploratory factor analysis for the vulnerability facet, before and after removal of IER as indicated by page time with 99% specificity cutoff value (Fig. 1). The scree plot on the full sample appeared to suggest the presence of two factors, whereas the plot after removal of IER indicates a single-factor solution. This example is consistent with the finding through simulation by Schmitt and Stults (1985).

Despite the overall supportive findings, the improvement in Cronbach's alpha and eigenvalues was smaller than what we had expected. A closer examination of the changes suggested that Cronbach's alpha and first-factor eigenvalues may not be enough to ensure the quality of one's data. As a case in point, we would like to direct attention to the impact of IER on the Cheerfulness Scale: removal of IER resulted in sizable *decrease* in both Cronbach's alpha and the first-factor eigenvalue. A further look into those responses detected by IER indices revealed that the decrease could be attributed to the confluence of two factors: (a) the scale contained eight positively keyed

Table 7 Agreement among IER indices in Study 2

Index	Page time 95%S	Long string 95%S	Psychometric antonym 95%S	Individual reliability 95%S	Page time 99%S	Long string 99%S	Psychometric antonym 99%S	Individual reliability 99%S	Self-report IER
Page time 95%S	–	23	35	45	47	20	35	33	21
Long string 95%S	85	–	25	25	14	47	21	13	10
Psychometric antonym 95%S	87	82	–	43	29	25	65	24	18
Individual reliability 95%S	91	85	43	–	33	21	51	33	20
Page time 99%S	94	86	29	91	–	21	33	39	19
Long string 99%S	88	93	25	88	92	–	24	20	18
Psychometric antonym 99%S	90	84	65	93	92	90	–	36	17
Individual reliability 99%S	92	86	24	92	96	93	93	–	18
Self-report IER	89	84	18	88	93	93	89	93	–

Note 95%S: Cutoff set with 95% specificity derived from Study 1. 99%S: Cutoff set with 99% specificity derived from Study 1. *Total agreement*, i.e., (cases both indices agreed upon)/(total number of cases), is presented below the diagonal. *Agreement in detection*, i.e., (IER detected by both indices)/(IER detected by either index), is presented above the diagonal

and only two negatively keyed items; and (b) some IE respondents chose the same response option repeatedly for all items. When most of the items were scored in the same direction, the repeated selection of the same response option regardless of item content *increases rather than decreases* the intercorrelation among items.[3]

Because Cronbach's alpha and eigenvalues focus on intercorrelation of items on the same facet, we next turn to the Big Five factor level to understand how inclusion of IER may affect intercorrelation among items on different facets of the same factor. We conducted CFA on each of the Big Five factors, comparing various fit indices obtained from the full sample and the trimmed samples. For example, the CFA for Conscientiousness contained six latent factors, corresponding to the six facets of Conscientiousness (achievement striving, cautiousness, dutifulness, orderliness, self-discipline, and self-efficacy) and each containing ten indicators (i.e., ten items per facet). Because we are unaware of any statistical test that compares model fit between a full sample and a trimmed sample, we provided a number of commonly used fit indices to allow for visual examination of change of fit. Note that the purpose of the series of CFA was to investigate the impact of IER rather than to evaluate the model fit per se. Thus, we did not compare the fit indices to rules of thumb for acceptable fit but rather focused the interpretation on the difference of fit indices before and after removal of IER. A visual examination of the indices led to the following observations: Across the CFA analysis examined for all Big Five factors, the trimmed samples based on the cutoffs for the page time, psychometric antonym, individual reliability, and long string indices yielded better fit than the full sample, whereas the trimmed sample based on the postsurvey self-report did not lead to a clear change of fit. Because the patterns of results were similar across Big Five factors, we present the results on Conscientiousness in Table 9 (results for the other four factors are available from the first author).

Discussion

Similar to Study 1, Study 2 showed that the IER indices correlated significantly with each other. Furthermore, the exploratory factor analysis showed that scores on four indices obtained from various approaches loaded on a

[3] We explored the correlation between (a) the extent to which each of the 30 scales contained unequal numbers of positively and negatively worded items (i.e., $|N$ positive $- N$ negative$|$) and (b) the increase in Cronbach's alpha on each scale after removal of suspect IER using the 99% specificity long string index. The result of $r = -.33$, $p = .07$, $N = 30$ suggests, albeit inconclusively, that IER in the form of long string responding had a stronger impact on scales with equal rather than unbalanced number of positively and negatively worded items.

Table 8 Psychometric properties before and after removal of IER with 99% specificity cutoffs

	Full sample	Samples created by removing IER detected by each index				
	Base line ($N = 345$)	Page time ($N = 327$)	Long string ($N = 323$)	Psychometric antonym ($N = 310$)	Individual reliability ($N = 331$)	Self-report ($N = 326$)
Cronbach's alpha						
Average	.788	.801	.796	.802	.799	.793
T test[a]	–	4.65***	3.64**	3.80***	4.96***	2.87**
Cohen's *d*	–	.85	.66	.69	.90	.52
Eigenvalue for the first factor						
Average	3.710	3.844	3.767	3.858	3.811	3.760
T test[a]	–	4.88***	2.77**	4.05***	4.67***	3.33**
Cohen's *d*	–	.89	.51	.74	.85	.61
Eigenvalue for the second factor						
Average	1.304	1.214	1.206	1.209	1.220	1.298
T test[a]	–	−4.03***	−4.70***	−3.45**	−4.60***	−0.94
Cohen's *d*	–	−.74	−.86	−.63	−.84	−.17

Note $N = 30$ (number of facets). * $p < .05$; ** $p < .01$; *** $p < .001$

[a] Paired samples *t* test comparing the estimates for the 30 facets obtained from each trimmed sample to those obtained from the full sample

common latent factor. The examination of the impact of IER on the psychometric properties of the measures showed a uniform pattern: After removal of suspect cases, the observed item interrelatedness and unidimensionality were improved.

General Discussion

Despite a clear relevance to psychological and organizational research, the identification of IER has remained the target of isolated investigations. This study provides a first step toward understanding this response style as well as various methods to detect it. Specifically, we addressed the issue of IER by (a) defining IER and summarizing different treatment approaches, (b) evaluating the effectiveness of four IER indices, and (c) examining the impact of IER on scale psychometric properties. Findings from Study 1 suggested that the IER indices of page time, psychometric antonym, and individual reliability were more effective at detecting IER. Results from Study 2 indicate that a common factor underlies all IER indices as well as a postsurvey self-report of IER. Furthermore, Study 2 showed significant improvement in scale psychometric properties after removing even a small number of suspected IER protocols.

Research and Practical Implications

The comprehensive evaluation of IER approaches yielded three candidates for future applications: (i) response time, (ii) psychometric antonym, and (iii) individual reliability. Results from Study 1 showed that each of these approaches was able to identify at least 40% of IER protocols while flagging only 5% of normal protocols. Further support for these three approaches came from Study 2, which showed that the three corresponding indices had the highest loadings on the first factor underlying the IER indices, suggesting that they are the strongest indicators of the IER response style.

The finding that the removal of a rather small proportion of IER protocols resulted in improved scale measurement properties has important implications for survey research. Screening data sets for IER may enable researchers to reduce measurement error, obtain better fitting models, and derive more accurate estimates of relationships between constructs. Given the time and effort spent in designing a research study and in subsequent data analysis and reporting, the investment in safeguarding against the nuisance from IER seems quite worthwhile.

Although the sensitivity of these three indices may appear unimpressive, we like to highlight the decision-making context, which falls under what Swets (1992) described as industrial quality-control approach, where the cutoff is set at the tail of the distribution to yield extremely few false positives (i.e., normal responses misidentified as IER). In fact, the sensitivity of these three indices was not much lower than some regularly administered examinations. For instance, the specificity and sensitivity of the clinical breast examination for breast cancer were .94 and .54, respectively (Barton et al. 1999), and the specificity and sensitivity of employee drug testing were once 98.6% and 38.2% in the mid-1980s (Gleason and Barnum 1991). As another comparison, when a selection test with a validity coefficient of .4 is applied to hire 50 employees

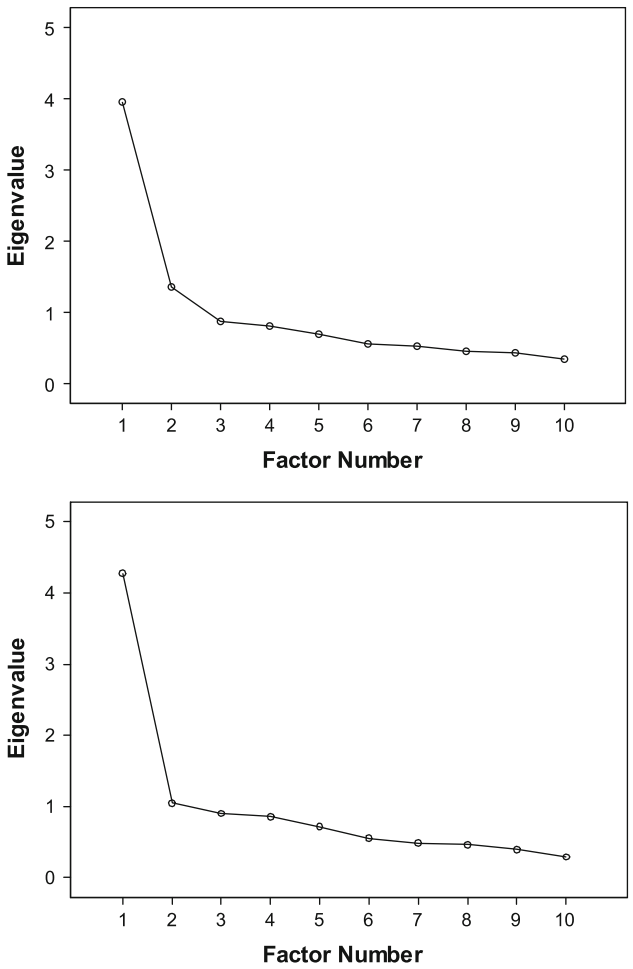

Fig. 1 Scree plot from exploratory factor analysis for vulnerability facet before and after removal of IER detected using page time 99% specificity cutoff. Full sample: before removal of IE responses ($N = 345$). Trimmed sample: after removal of IE responses ($N = 327$)

from 125 applicants with a base rate of .5, the specificity of the test to correctly identify unqualified applicants would be .73, whereas the sensitivity of the test to correctly identify qualified applicant would be .52 (computed based on Martin and Terris 1990, Table 1). As suggested in Study 1, taking a rational conservative approach to setting cutoffs can achieve a high specificity while detecting a reasonable proportion of IER protocols. Ultimately, the survey user, whether a researcher or a practitioner, needs to weigh the benefit of removing error from the data against the potential cost of incorrectly identifying valid protocols as IER.

Although encouraging participants to engage in the survey process takes higher priority, it is advisable to include some mechanism of IER detection when complete compliance from all participants cannot be ensured. For example, unmotivated incumbents may engage in IER when responding to surveys for human resource practices, such as surveys for job analysis (Morgeson and Campion 1997) and concurrent validation (Hough et al. 1990), despite the intentions of survey administrators. By including IER detection indices, survey administrators have a quality-control mechanism to guard against problematic data.

The decision to adopt any index may depend on the format and length of the survey as well as practicality, as none of the approaches examined in this study can apply to all survey situations. Establishing the effectiveness of the IER approaches enables practitioners to make an informed decision about the potential approaches to employ. For example, with the increasing use of online surveys, organizations may want to incorporate response time in the online survey design, which may be particularly helpful for shorter surveys that are not amenable to screening based on the inconsistency approach.

Limitations and Directions for Future Research

The results of this study may be of limited generalizability because of the particular sample and measures used. Indeed, the specific type of items, the format and length of the survey, and the online administration may all limit the

Table 9 Results from confirmatory factor analysis on conscientiousness factor in Study 2 on the full sample and trimmed samples

	Full sample	Page time 95%S	Long string 95%S	Psychometric antonym 95%S	Individual reliability 95%S	Page time 99%S	Long string 99%S	Psychometric antonym 99%S	Individual reliability 99%S	Self-report
RMSEA	.083	.070	.066	.069	.067	.068	.065	.067	.069	.083
SRMR	.100	.091	.089	.089	.088	.088	.086	.088	.090	.100
CFI	.92	.93	.93	.93	.94	.93	.93	.94	.93	.92
NFI	.87	.89	.88	.89	.89	.89	.88	.89	.88	.87
NNFI	.91	.93	.93	.93	.93	.93	.93	.93	.93	.91
IFI	.92	.93	.94	.93	.94	.93	.93	.94	.93	.92
RFI	.87	.88	.88	.88	.88	.88	.88	.88	.88	.87
AIC	5,935	4,510	4,130	4,332	4,296	4,531	4,292	4,313	4,640	5,784

Note 95%S: Cutoff set with 95% specificity derived from Study 1. 99%S: Cutoff set with 99% specificity derived from Study 1

generalizability. Nonetheless, this study contributes to the literature by offering a first attempt to comprehensively evaluate the different approaches to mitigate the presence of IER. The findings, albeit qualified by the uniqueness of the study design, laid a foundation for future investigations of the response phenomenon in other samples, measures, and formats.

The manipulation checks suggest that participants followed instructions, but it is still possible that some participants did not understand or follow instructions. However, to the extent that any deviation from standard instructions occurred, it can only lead to conservative estimates of sensitivity. Specifically, if some participants in the non-IER condition actually engaged in IER, then the empirical cutoffs based on 95% or 99% specificity were set too leniently, resulting in lowered estimates of sensitivity. Conversely, the occurrence of normal responding behavior in the IER conditions also suppresses the estimate for sensitivity.

The low levels of agreement among indices on the IER cases detected in the normal survey in Study 2 (Table 7) also point to a direction for future research. Although any disagreement can easily be attributed to detection error, a closer look at the mechanism by which each IER index functions yields an additional explanation: Each IER index may be slightly sensitive in detecting a specific *form* of IER. For instance, a distracted survey respondent may fill out the survey slowly and inconsistently. As a result, the inconsistency approach is more sensitive than the response time approach to detect such protocols. The different forms of IER are beyond the scope of this study and remain topics for future investigations.

Also worth discussing is the exploratory factor analysis finding that a single factor underlies three continuous IER indices and the postsurvey self-report measure. Although the variance explained by the factor was no more than 59%, the fact that these measures were obtained in very different ways lends strong support to the notion that these indices all capture an underlying "insufficient effort" response style. Although previous studies concluded that inconsistency measures may reflect the respondent's personality rather than response set (e.g., Goldberg and Kilkowski 1985; Kurtz and Parrish 2001), our findings suggest otherwise that the individual reliability and psychometric antonym indices primarily reflect response effort. The difference in findings could be because of different inconsistency indices used and different cutoff levels. More studies are needed to understand the nature of various inconsistency measures.

The identification of effective IER detection indices provides opportunities for future research. Researchers may begin to further examine how IER affects the relationship between variables. Interestingly, the serendipitous finding on Cheerfulness suggests that presence of IER, specifically in the form of long string responses, may increase the correlation between two scales, if the scales consist of most or all positively keyed items. Another venue worthy of pursuing is in the context of measurement invariance. Schmit and Ryan (1993) showed that measurement situations with different purposes or consequences may lead to different factor structure for the same instrument. It would be interesting to examine whether differential IER because of different measurement purposes or consequences may contribute to the difference in factor structure.

In conclusion, this study evaluated several alternatives for detecting and deterring IER to surveys, including inconsistency, response pattern, response time, and the use of warning. Rather than relying on faith that all participants will deliberate and provide reliable responses, practitioners and researchers are encouraged to examine the survey data at the individual level using one or more effective indices to filter insufficient effort responses.

Acknowledgments We thank Goran Kuljanin for collecting data for the two studies. We are grateful for the constructive comments from Neal Schmitt and Ann Marie Ryan on an earlier draft of this article.

References

Archer, R. P., & Elkins, D. E. (1999). Identification of random responding on the MMPI-A. *Journal of Personality Assessment, 73*, 407–421.

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

Babbie, E. (2001). *The practice of social research* (9th ed.). Belmont, CA: Wadsworth.

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.

Baer, R. A., Kroll, L. S., Rinaldo, J., & Ballenger, J. (1999). Detecting and discriminating between random responding and overreporting on the MMPI-A. *Journal of Personality Assessment, 72*, 308–320.

Bagby, R. M., Gillis, J. R., & Rogers, R. (1991). Effectiveness of the Millon Clinical Multiaxial Inventory Validity Index in the detection of random responding. *Psychological Assessment, 3*, 285–287.

Barton, M. B., Harris, R., & Fletcher, S. W. (1999). Does this patient have breast cancer? The screening clinical breast examination: should it be done? How? *JAMA, 282*, 1270–1280.

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied, 123*, 101–103.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*. doi:10.3758/s13428-011-0081-0

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: a meta-analysis. *Clinical Psychology Review, 11*, 585–598.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices:

validation using a self-report methodology. *Psychological Assessment, 4*, 340–345.

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.

Buechley, R., & Ball, H. (1952). A new test of "validity" for the group MMPI. *Journal of Consulting Psychology, 16*, 299–301.

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

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., et al. (1992). *MMPI–A: Minnesota Multiphasic Personality Inventory–Adolescent: manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Charter, R. A. (1994). Determining random responding for the Category, Speech-Sounds Perception, and Seashore Rhythm tests. *Journal of Clinical and Experimental Neuropsychology, 16*, 744–748.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

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

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The Sage handbook of personality theory and assessment: personality measurement and testing* (pp. 179–198). London: Sage.

Curran, P. G., Kotrba, L., & Denison, D. (2010, April). *Careless responding in surveys: applying traditional techniques to organizational settings*. Paper presented at the 25th annual conference of Society for Industrial and Organizational Psychology, Atlanta, GA.

DiLalla, D. L., & Dollinger, S. J. (2006). Cleaning up data and running preliminary analyses. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook: a guide for graduate students and research assistants* (2nd ed., pp. 241–253). Thousand Oaks, CA: Sage.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23.

Evans, R. G., & Dinning, W. D. (1983). Response consistency among high F scale scorers on the MMPI. *Journal of Clinical Psychology, 39*, 246–248.

Gleason, J. M., & Barnum, D. T. (1991). Predictive probabilities in employee drug-testing. *Risk, 2*, 3–18.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. D. Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

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

Green, S. B., & Stutzman, T. M. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543–564.

Greene, R. L. (1978). An empirically derived MMPI Carelessness Scale. *Journal of Clinical Psychology, 34*, 407–410.

Haertzen, C. A., & Hill, H. E. (1963). Assessing subjective effects of drugs: an index of carelessness and confusion for use with the Addiction Research Center Inventory (ARCI). *Journal of Clinical Psychology, 19*, 407–412.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Thousand Oaks, CA: Sage.

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

Hsu, L. M. (2002). Diagnostic validity statistics and the MCMI-III. *Psychological Assessment, 14*, 410–422.

Jackson, D. N. (1976). *The appraisal of personal reliability*. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

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

Kline, R. B. (2009). *Becoming a behavioral science researcher: a guide to producing research that matters*. New York: Guilford.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: the case of the NEO-PI-R. *Journal of Personality Assessment, 76*, 315–332.

Lucas, R. E., & Baird, B. M. (2005). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington, DC: American Psychological Association.

Marsh, H. W. (1987). *The Self-Description Questionnaire 1: manual and research monograph*. San Antonio, TX: Psychological Corporation.

Martin, S. L., & Terris, W. (1990). The four-cell classification table in personnel selection: a heuristic device gone awry. *The Industrial-Organizational Psychologist, 47*(3), 49–55.

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. (2011, April). *Identifying careless responses in survey data*. Paper presented at the 26th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

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

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627–655.

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239–250.

O'Rourke, T. (2000). Techniques for screening and cleaning data for analysis. *American Journal of Health Studies, 16*, 205–207.

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

Pinsoneault, T. B. (1998). A Variable Response Inconsistency Scale and a True Response Inconsistency Scale for the Jesness Inventory. *Psychological Assessment, 10*, 21–32.

Pinsoneault, T. B. (2007). Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols. *Psychological Assessment, 19*, 159–164.

Rosse, J. G., Levin, R. A., & Nowicki, M. D. (1999, April). *Assessing the impact of faking on job performance and counter-productive behaviors.* Paper presented at the 14th annual meeting of the Society for Industrial and Organizational Psychology, Atlanta.

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

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966–974.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: the result of careless respondents? *Applied Psychological Measurement, 9*, 367–373.

Seo, M. G., & Barrett, L. F. (2007). Being emotional during decision making–good or bad? An empirical investigation. *Academy of Management Journal, 50*, 923–940.

Smith, P. C., Budzeika, K. A., Edwards, N. A., Johnson, S. M., & Bearse, L. N. (1986). Guidelines for clean data: detection of common mistakes. *Journal of Applied Psychology, 71*, 457–460.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334–344.

Streiner, D. L. (2003). Diagnosing tests: using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*, 209–219.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.

Thompson, A. H. (1975). Random responding and the questionnaire measurement of psychoticism. *Social Behavior and Personality, 3*, 111–115.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van Ginkel, J. R., & van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement, 29*, 152–153.

Wetter, M. W., Baer, R. A., Berry, D. T. R., Smith, G. T., & Larsen, L. H. (1992). Sensitivity of MMPI-2 validity scales to random responding and malingering. *Psychological Assessment, 4*, 369–374.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594–604.

Wilson, M. A., Harvey, R. J., & Macy, B. A. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. *Journal of Applied Psychology, 75*, 158–163.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183.

Woods, C. M. (2006). Careless responding to reverse-worded items: implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189–194.