Psychological Methods
2012, Vol. 17, No. 3, 437–455

© 2012 American Psychological Association
1082-989X/12/$12.00 DOI: 10.1037/a0028085

Identifying Careless Responses in Survey Data

Adam W. Meade and S. Bartholomew Craig
North Carolina State University

When data are collected via anonymous Internet surveys, particularly under conditions of obligatory participation (such as with student samples), data quality can be a concern. However, little guidance exists in the published literature regarding techniques for detecting careless responses. Previously several potential approaches have been suggested for identifying careless respondents via indices computed from the data, yet almost no prior work has examined the relationships among these indicators or the types of data patterns identified by each. In 2 studies, we examined several methods for identifying careless responses, including (a) special items designed to detect careless response, (b) response consistency indices formed from responses to typical survey items, (c) multivariate outlier analysis, (d) response time, and (e) self-reported diligence. Results indicated that there are two distinct patterns of careless response (random and nonrandom) and that different indices are needed to identify these different response patterns. We also found that approximately 10%–12% of undergraduates completing a lengthy survey for course credit were identified as careless responders. In Study 2, we simulated data with known random response patterns to determine the efficacy of several indicators of careless response. We found that the nature of the data strongly influenced the efficacy of the indices to identify careless responses. Recommendations include using identified rather than anonymous responses, incorporating instructed response items before data collection, as well as computing consistency indices and multivariate outlier analysis to ensure high-quality data.

Keywords: careless response, random response, data screening, data cleaning, mixture modeling

In any type of research based on survey responses, inattentive or careless responses are a concern. While historically the base rate of careless or inattentive responding has been assumed to be low (Johnson, 2005), there is reason to believe that careless responding may be of concern in contemporary Internet-based survey research, particularly with student samples. While there are many advantages to Internet-based data collection, the lack of environmental control could lead to a decrease in data quality (Buchanan, 2000; Johnson, 2005). Accordingly, it is important for researchers to be able to screen such data for careless, partially random, or otherwise inattentive responses. Such data could lead to spurious within-group variability and lower reliability (Clark, Gironda, & Young, 2003), which, in turn, will attenuate correlations and potentially create Type II errors in hypothesis testing. Student samples contributing data via web-based forms under unproctored conditions may be particularly prone to substantial levels of inattentive or partially random response.

The studies described here make several significant contributions to the literature. First, we provide the first investigation of a comprehensive set of methods for screening for careless responding and provide an understanding of the way in which they relate to one another and observed data responses. Second, we examine the latent class structure of a typical undergraduate respondent sample using both latent profile analysis and factor mixture modeling, revealing data patterns common among careless responders. Third, we provide an estimate to the prevalence of careless responding in undergraduate survey research samples. Fourth, we examine the role of instruction sets (anonymous vs. identified) as well as the efficacy of response time and self-report indicators of careless responding.

What Is Careless Responding, and Why Does It Matter?

There are several aspects to data "quality." In this study, we focus on raw data, provided directly by respondents, that do not accurately reflect respondents' true levels of the constructs purportedly being measured. While there are multiple reasons that respondents may provide inaccurate responses to survey questions, Nichols, Greene, and Schmolck (1989) delineate two types of problematic response. The first they term *content responsive faking*, which has two hallmarks: (a) responses are influenced by the item content but (b) are not completely accurate. Content responsive faking can be further delineated into purposeful faking (e.g., malingering, a common concern on clinical instruments such as the Minnesota Multiphasic Personality Inventory [MMPI-2]; Berry, Baer, & Harris, 1991; Rogers, Sewell, Martin, & Vitacco, 2003) and socially desirable response of both intentional and nonintentional varieties (Paulhus, 1984).

The primary focus of the current studies is on Nichols et al.'s (1989) second category of response bias: *content nonresponsivity*, which is defined as responding without regard to item content. This would include responses that have been variously described as random response (Beach, 1989; Berry et al., 1992), careless

This article was published Online First April 16, 2012.

Adam W. Meade and S. Bartholomew Craig, Department of Psychology, North Carolina State University.

Correspondence concerning this article should be addressed to Adam W. Meade, Department of Psychology, North Carolina State University, Campus Box 7650, Raleigh, NC 27695-7650. E-mail: awmeade@ncsu.edu

responding (Curran, Kotrba, & Denison, 2010), and protocol invalidity (Johnson, 2005). There are multiple data patterns that can result from such a situation. For example, some persons may randomly choose from all response options on a scale. Others may employ a nonrandom pattern, such as giving many consecutive items a response of "4," or repeating a pattern of "1, 2, 3, 4, 5. . .." We prefer the terms *inattentive* or *careless* response, rather than *random* response, as the resultant data may be decidedly nonrandom.

Concerns with such respondents are not new. Clinical measures such as the MMPI-2 have long contained special scales intended to detect purposefully deceptive responses (e.g., MMPI-2 Lie scale; Berry et al., 1992; Berry, Wetter, et al., 1991) and a lack of consistency on items to which attentive respondents tend to answer in a similar way (e.g., MMPI-2 Variable Response Inconsistency [VRIN] and True Response Inconsistency [TRIN] scales). Similar scales have been developed for personality measures, such as the NEO Personality Inventory (NEO-PI; Kurtz & Parrish, 2001; Schinka, Kinder, & Kremer, 1997).

There are several reasons to be concerned about inattentive or careless responding. First and perhaps most intuitively, a "clean" data set is highly desirable and data screening to delete cases with inappropriate responses is commonly recommended as part of the data analytic process (e.g., Tabachnick & Fidell, 2007). Unfortunately, common recommendations typically entail only cursory data screening methods, such as univariate outlier analysis, the effectiveness of which is predicated on the assumption that careless or inattentive responses are rare or extreme in magnitude. Moreover, typical univariate outlier analysis is unable to detect careless responses in cases where respondents chose a common response option such as the middle response option.

A second reason to be concerned with careless responses is that they can have serious psychometric implications. This is particularly true when surveys are administered for scale development purposes, as item development is based largely on item intercorrelations (Hinkin, 1998). Random responses constitute error variance, which attenuates correlations, reduces internal consistency reliability estimates, and potentially results in erroneous factor analytic results. Nonrandom inattentive responses may have unpredictable effects on the correlations among items. In one of the few studies on this topic, Johnson (2005) illustrated how factor structures differed for subsamples identified as purposeful and careless respondents. Additionally, careless responding on reverse-coded items can contribute to the presence of so-called "method" factors, in which positively worded items for a given scale load onto one factor, while negatively worded items for the same scale load onto another (Woods, 2006). Woods (2006) found that a single-factor confirmatory model does not fit in such instances, when as little as 10%–20% of respondents are careless with reverse coded items (see also Huang, Curran, Keeney, Poposki, & DeShon, 2012).

Base Rate

Relatively few studies have examined the prevalence of inattentive response, and among those that have, prevalence estimates vary widely. Many of these differences in estimates can be attributed to the method used to assess careless response. For example, Johnson (2005) cites a base rate of 3.5% for careless response; however Johnson's study featured International Personality Item Pool (IPIP; Goldberg, 1999) respondents who voluntary sought out and completed the measure on the Internet. Such respondents are likely to be considerably more motivated than a typical university sample. Moreover, a rather stringent statistical consistency index criterion was required to have been met before a response was judged as inattentive. Similarly, Ehlers, Greene-Shortridge, Weekley, and Zajack (2009) estimated random responding to be around 5% in their sample of job applicants, who likely were motivated to respond diligently. Curran et al. (2010) examined three indicators of random response to a job satisfaction questionnaire and found prevalence around 5%, 20%, or 50% among a large sample of employee respondents, depending on the criteria by which the researchers defined inattentive response. The variance in their results highlights the importance of the indices chosen as indicators of careless response. Kurtz and Parish (2001) found random responding prevalence to be 10.6% with college students completing the Revised NEO-PI (NEO-PI-R; Costa & McCrae, 2008) for course credit. However, their results led them to question the efficacy of the index with which they classified persons as inattentive.

One commonality across all of these studies is that they used indices of consistency calculated by comparing responses from items in different locations in the survey. While such indicators are useful for identifying respondents that respond carelessly in a very pervasive way, we posit that very few people will respond in an outright random pattern across the length of an entire survey. We believe it is much more likely that respondents will only intermittently respond inattentively. Previous work supports this notion. For instance, Berry et al. (1992) found that across three studies, 50%–60% of college student respondents admitted via self-report to answering randomly on one or more MMPI-2 items. Similarly, Baer, Ballenger, Berry, and Wetter (1997) found that 73% of respondents to the MMPI-2 self-reported responding carelessly to one or more items. Berry et al. (1992) found that even among respondents completing the MMPI-2 as part of a job application, 52% self-reported responding inattentively to at least one item. However, in all three studies, the overall proportion of items for which respondents admitted to careless response was small. In sum, it appears that relatively few respondents provide truly random responses across the entire length of a survey. However, it would seem that occasional careless response to a limited number of items may be quite common.

Factors Affecting Careless Responding

Respondent Interest

It is clear that the engagement of the respondent in the response process is critical (Schwarz, 1999; Tourangeau, Rips, & Rasinski, 2000). When respondents have a genuine interest in the survey, for example a selection employment test or a feedback survey, careless responses is less likely (although other types of bias, such as self-presentation, may occur). This is especially true when survey participation is optional and disinterested persons can opt out. In the current research, we focus on perhaps the most common source of data in psychological research, college students (Gordon, Slade, & Schmitt, 1986). Survey research routinely makes use of such samples, especially for tasks such as large-sample testing of new measures. In most universities in the United States, students typically agree to participate in a research study in exchange for credit that meets some type of course obligation. While institutional review boards mandate that alternatives, such as writing a paper of suitable length, be made available, many students consider such alternatives to be even more burdensome and thus become members of a participant pool. Individual differences among members of such a population are considerable, and many participants are actively interested and engaged in the research process. However, others may be reluctant participants at best or may be resentful of the process (Schultz, 1969). While motivational incentives may sometimes be used, often the "incentive" is that participation in research meets an obligation that is not typical of college classes in other disciplines.

Additionally, while some lab-based studies can be quite cognitively engaging, survey participation tends to be passive, in that there is minimal interaction between the researcher and the participants. It seems somewhat unlikely that there would be strong intrinsic interest in completing surveys for most undergraduates given that feedback is rarely provided to the participant.

Survey Length

Longer surveys require more sustained effort from respondents and most modern personality tests are long by any measure. For instance, the NEO-PI-R (Costa & McCrae, 2008) contains 240 Likert-type items, the MMPI-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008) contains 338 true/false items, the 16PF (5th version; Cattell, Cattell, & Cattell, 1993) contains 185 multiple choice items, and the California Personality Inventory (Gough & Bradley, 1996) contains 434 true–false questions. It seems reasonable to expect respondent attention to wane over the course of such long surveys. Indeed, respondents are more likely to self-report responding randomly toward the middle or end of long survey measures (Baer et al., 1997; Berry et al., 1992). Even highly motivated samples such as job applicants may feel fatigue effects in long surveys (Berry et al., 1992).

Social Contact

As a social interaction, social norms govern participation in research. Dillman, Smyth, Christian, and Dillman's (2009) highly popular model of survey design (originally developed for mail surveys) stresses highly tailored communication designed to increase the social interaction between the researcher and the respondent. In university settings, survey research seldom has the hallmarks of multiple communications and other methods advocated by Dillman et al. Paper-and-pencil survey research in university settings typically takes place in a proctored setting involving direct communication between the researcher (or an assistant) and the respondents via face-to-face interaction. Internet-based surveys are not new but do represent a change in the level of social contact from previous paper-and-pencil surveys in the university setting. Johnson (2005) contends that the physical distance and lack of personalization introduced by online administration may result in less accountability and thus more undesirable response patterns. We believe this effect could be exacerbated by anonymity during the response process. Online anonymity has been shown to reduce personal accountability leading to a greater potential for negative behavior, such as posting negative comments in a discussion group (Douglas & McGarty, 2001; Lee, 2006).

Environmental Distraction

Perhaps the largest disadvantage of Internet surveys is the lack of a controlled setting. Online surveys require only Internet access—the environment is otherwise completely uncontrolled. Even motivated and conscientious respondents may encounter distractions or put themselves in an environment prone to distraction (television, etc.). Recent surveys (Carrier, Cheever, Rosen, Benitez, & Chang, 2009; Montgomery, 2007) support popular press accounts (Wallis, 2006) suggesting that younger generations are more likely to attempt to multitask than previous generations. While little survey-specific research exists, participants completing the survey under conditions of divided attention would seem to be more likely to provide inattentive responses given the well documented effects of divided attention on performance of cognitive and physical tasks (Spelke, Hirst, & Neisser, 1976).

Summary

Internet-based survey studies using undergraduates may see low intrinsic respondent interest, long measures, virtually no social exchange, and very little control over the environment. In terms of conditions likely to be conducive to collecting high-quality data, the situation is very poor (Buchanan, 2000).

Methods for Identifying Careless Responders

Methods of screening for careless response can be broken into roughly two types. The first type requires special items or scales to be inserted into the survey prior to administration. One version of these items are those that covertly attempt to index respondent care in response, or to flag those that are not carefully reading the item stem. Examples include social desirability (e.g., Paulhus, 2002) and lie scales (e.g., MMPI-2 Lie scale), nonsensical or "bogus" items (e.g., Beach, 1989), special scales designed to assess consistent responding (e.g., the MMPI-2 VRIN and TRIN scales), and instructed response items (e.g., "To monitor quality, please respond with a two for this item"). A second version includes self-report measures of response quality placed at the end of a survey.

The second broad type of screening can be described as post hoc in that these methods do not require specialized items but instead involve special analyses after data collection is complete. There are several indices that can be computed post hoc for identifying careless response. The first general type can be considered indices indexing response consistency. *Consistency indices* typically match items that are highly similar either based on their designed function (i.e., which construct the item was written to measure) or based on empirical correlations among items. A lack of consistent responding is then indicated by some type of deviation among responses to similar items or a within-person correlation across item pairs. Several such indices are available and are geared toward identifying respondents that do not respond consistently across similar items. The variations of each of these approaches are examined in the current study and are detailed in the method section.

Variants of the consistency approach, which we term *response pattern indices*, are intended to identify persons responding *too consistently* to items measuring theoretically distinct constructs. These indices are typically computed by examining the number of consecutive items for which a respondent has indicated the same response option. Presuming items are arbitrarily ordered, excessive utilization of a single response option can be considered an index of careless responding.

A second general class of indices are *outlier indices*. While univariate outlier analysis may have some utility for identifying extreme cases, multivariate approaches such as Mahalanobis distance are much more appropriate, as they consider the pattern of responses across a series of items. Thus, a series of responses for a given scale's items may appear quite improbable via the distance measures, even though each individual item response may not.

An additional approach is to examine study response time. Typically this approach posits a nonlinear relationship between response time and response quality such that very fast responses are assumed to be careless in nature, yet once some threshold is identified, response times above the threshold may or may not be considered careless depending on other indicators.

Another strategy is to try to prevent inattentive responses in the first place via instruction sets. Survey research tends to provide anonymity for respondents on the premise that anonymity will afford respondents the freedom to be honest when asked questions with potentially socially desirable response options. Ironically, however, such instructions may also result in less accountability for respondents. As such, it is possible that forcing respondents to respond in an identified manner would lead to fewer careless responses. Thus, we asked:

> *Research Question 1:* Does manipulation of survey anonymity affect the prevalence of careless responding?

Understanding how each method of identifying careless respondents indexes various response patterns and how these indices relate to one another is fundamental to the study. Such an investigation is imperative prior to making any recommendations regarding how to best identify such respondents. Thus, we asked:

> *Research Question 2:* What are the correlations among, and factor structure of, careless response indices?

Using two approaches of examining the latent nature of response tendencies, we asked:

> *Research Question 3:* Are there different types (latent classes) of careless respondents, and if so, what data patterns underlie these different types of careless respondents?

A central concern among researchers is the prevalence of careless responding among data. Thus, we asked:

> *Research Question 4:* What portion of the sample can be said to be responding carelessly?

Several of the data screening methods we discuss are labor intensive and impractical for some purposes. For this reason we give special attention to two easily implemented potential measures of data quality:

> *Research Question 5:* Are self-report measures of data quality sufficient for data screening?

> *Research Question 6:* Can survey completion time be used as an indicator of data quality?

Finally, in order to recommend specific indices that can be used to identify careless response, we ask:

> *Research Question 7:* Of the different types of careless response measures, which are most sensitive to careless responding?

These questions were answered using a large undergraduate sample in Study 1. A follow-up study was also conducted in which simulated data with properties similar to those in Study 1 were created. A portion of the simulated sample was then replaced with totally random data or partially random data, and the indicators were evaluated with respect to their efficacy for identifying random data.

Study 1

Method

Participants. Participants were 438 respondents drawn from a participant pool composed primarily of students enrolled in introductory psychology courses at a large university in the southeastern United States. Although students were not required to participate in this particular study, they were required to spend approximately 3 hours during the semester participating in studies of their choice, selected from an online listing of current research projects. The university is a state-supported school with approximately 24,000 undergraduate students with average SAT Verbal + Quantitative scores of entering freshmen between 1,175 and 1,200. Demographic information was not collected for this study, but on the whole the participant pool tends to be approximately 60% female and predominantly Caucasian. In this study, we only utilized responses for participants that completed the entire study (88% of the sample), with a final $N = 386$.

Survey design and issues. Survey items were spread across 12 web pages, each clearly marked with the page number and total number of pages (e.g., page 3 of 12). Pages 1–10, with the exception of page 7, contained 50 items per page displayed in table form with the item content in the first column and a drop-down box of response options in the second. Page 11 contained an additional 31 items, and page 12 contained items asking respondents about the study experience (described below). Items on pages 1–11 used a 1–7 Likert-type response scale. Page 7 included a brief description of a series of e-mails intended for use as a recall data quality measure, which was later abandoned because of a low base-rate of accurate recall.

Procedure. Participants who signed up for the study were e-mailed a link to a website where a JavaScript routine randomly assigned them to one of three survey conditions.

Anonymous condition. The first condition was intended to represent a typical survey research condition in which respondents remain anonymous ($N = 147$).

Identified condition. In the second condition respondents were instructed "Your responses will be completely confidential, however, on each page of the survey you will be asked to enter your name so that we can later merge your responses across the different web pages." Each page included the following instructions at the bottom along with a blank text field "Please enter your name so that we may merge your responses across web pages." There were 120 participants in this condition.

Stern warning condition. The third condition identified respondents as described above but also provided a stern warning: "Your honest and thoughtful responses are important to us and to the study. Remember that your honesty and thoughtful responses are subject to [the university's] academic integrity policy." The bottom of each page included a text field for the respondent to type his or her name beside the statement: "I verify that I have carefully and honestly answered all questions on this page in accordance with [the university's] honor policy." There were 119 participants in this condition.

Measures

Personality. Given that some authors have argued that careless response is a relatively low-frequency event (Johnson, 2005), and others have found that such responses tend to occur later in a long survey (Berry et al., 1992), we sought to administer a long but realistic survey. Our primary measure was the 300 item International Personality Item Pool (IPIP; Goldberg, 1999), a freely available measure of the five-factor model of personality. Additionally, a 26-item measure of psychopathy (Levenson, Kiehl, & Fitzpatrick, 1995) and a 40-item measure of narcissism (Raskin & Terry, 1988) were included. There were no research questions related to these constructs, but they were included to be typical of the scales commonly found on long surveys.

Social desirability. Perhaps the most common type of special data screening scale is social desirability or "unlikely virtues" scales. Social desirability scales typically present the respondent with items whose endorsement would be considered virtuous but for which it is unlikely that many of the items would be true for any single individual. The current study included four social desirability measures. The first was the 33-item Marlowe-Crowne social desirability scale (Crowne & Marlowe, 1960). Although the original Marlowe-Crowne scale consisted of dichotomous true/false items, the same 7-point response format used for other items in the current survey was used with this scale. The second scale was the IPIP social desirability scale (Goldberg, 1999). The third and fourth indices were the self-deception (SD) and impression management (IM) subscales of an updated version of the Balanced Inventory of Desirable Responding scale (BIDR; Paulhus, 1984, 2002).

Bogus items. Another type of data screening method is the use of items with a clear correct answer (e.g., "I was born on February 30th"; Beach, 1989). If the respondent chooses the incorrect response, it is assumed that he or she was not attending to the content of the item. An advantage of such bogus items is that if the respondent chooses an incorrect response, there is little doubt that he or she is responding carelessly or dishonestly; thus, there is little chance of a false positive. However, false negatives are considerably more likely, particularly if all such items are coded in the same direction such that a consistent use of "agree" results in uniformly correct responses. Further, such items must be written so there is an unambiguous correct answer. Grice (1975) and Schwarz (1999) discuss conversational norms that dictate how respondents and researchers interact via a survey. In essence, participants are not expecting "trick" items on such a survey and may acquiesce to a bogus item if the item is ambiguously worded. The current study contained 10 bogus items, developed for this study, with one on each of the first 10 pages in a random position between the 30th and 45th (of 50) items (see Table 1 for list of bogus items and endorsement rates).

Self-reported study engagement. Seventeen items were written by the current authors to assess self-reported (SR) participant engagement with the study and were included on the final webpage. The Appendix presents the results of preliminary analyses that led to the formation of two subscales of engagement: SR Diligence (nine items, α = .83) and SR Interest (six items, α = .81).

Self-reported single item (SRSI) indicators. We also included a series of single-item measures at the end of the survey intended to allow respondents to indicate how much effort and attention they devoted to the study. Prior to these items, we included the following instruction set: "Lastly, it is vital to our study that we only include responses from people that devoted their full attention to this study. Otherwise years of effort (the

Table 1
Bogus Items Response Rates

Item	Str. D1	D2	Sl.D3	Neither A nor D4	Sl. A5	A6	Str. A7	% Flagged by item
1. I am using a computer currently. (R)	2	4	4	3	3	55	315	4
2. I have been to every country in the world.	299	59	10	7	4	4	0	7
3. I am enrolled in a Psychology course currently. (R)	2	6	5	7	3	53	310	7
4. I have never spoken to anyone who was listening.[a]	116	123	36	90	9	10	2	38[a]
5. I sleep less than one hour per night.	286	60	13	15	7	3	2	10
6. I do not understand a word of English.	301	48	7	12	5	11	2	10
7. I have never brushed my teeth.	322	33	6	12	8	4	1	8
8. I am paid biweekly by leprechauns.	269	38	13	43	8	4	11	20
9. All my friends are aliens.	270	47	15	28	8	7	11	18
10. All my friends say I would make a great poodle.	215	66	21	62	8	11	3	27

Note. D = disagree; A = agree; Str. = strongly; Sl. = slightly; R = Items flagged if (reverse coded) strongly disagree or disagree not chosen (except missing data).
[a] Item 4 was dropped as a Bogus Item based on frequent response.

researchers' and the time of other participants) could be wasted. *You will receive credit* for this study no matter what, however, please tell us how much effort you put forth towards this study." *SRSI Effort* was assessed as the response to the item "I put forth ____ effort towards this study" with response options of 1 = "almost no," 2 = "very little," 3 = "some," 4 = "quite a bit," and 5 = "a lot of."

Next, we included the text "Also, often there are several distractions present during studies (other people, TV, music, etc.). Please indicate how much attention you paid to this study. Again, you will receive credit no matter what. We appreciate your honesty!" *SRSI Attention* was then assessed as the response to the item "I gave this study ____ attention" with options 1 = "almost no," 2 = "very little of my," 3 = "some of my," 4 = "most of my," and 5 = "my full."

Last, we asked, "In your honest opinion, should we use your data in our analyses in this study?" with a 1 = "yes" or 0 = "no" response. This item was referred to as *SRSI UseMe*.

Indices of Careless Responding

Response time. We examined completion time for each of the 12 web pages that made up the survey. Although the system was not designed in such a way that participants could save their information and resume the study at a later date, several appeared to have bookmarked the page and returned to the study later. Others may have simply taken a break from the study. As a result, for some persons, the time spent on a given page included very long durations (e.g., thousands of minutes). Descriptive statistics revealed that between 95% and 98% of persons spent fewer than 15 min on each survey page. Additionally, it seemed that 15 min was more than sufficient time for a diligent participant to respond to 50 items. As a result, elapsed time values greater than 15 min were set as missing for 95 persons. These missing data were handled differently for different analyses as described later.

The survey administration system was unable to prevent multiple entries from the same respondent. As a result, for 120 respondents, there were multiple responses for the same person for a given page or pages. This is a pervasive issue with Internet-based surveys and is one for which there is little published guidance (Johnson, 2005). Eighty-four of these cases were completely identical; however, 36 respondents changed their responses between submissions. For these persons, we retained the first set of responses, as these were deemed most likely to be equivalent to responses from individuals who only completed each page once. As computations of elapsed time to complete the study could be misleading for persons with multiple page submissions, we did not compute response time for these 120 persons and treated those data as missing. Note that these 120 persons and the 95 with long response times are not mutually exclusive. Table 2 provides a summary of the indicators we examined.

Table 2
Summary of Data Screening Methods

Index	Description
Total minutes	Total time to complete survey
Sum of Bogus	Sum of nine dichotomously scored bogus items with clear correct/incorrect answers
Psy Antonyms	Within-person correlation across item pairs with strong negative correlation
Psy Synonyms	Within-person correlation across item pairs with strong positive correlation
Even Odd Cons.	Within-person correlation across subscales formed by even-odd split of unidimensional scales, with Spearman-Brown split-half formula applied
Avg LongString	Average of 10 LongString values. LongString is the maximum number of identical consecutive responses on a webpage
Max LongString	Maximum of 10 LongString values
Mahalanobis D	Multivariate distance between respondent's response vector and the vector of sample means
SRSI Use Me	Dichotomous self-reported single item yes/no response as to whether respondent feels his or her data should be used for analysis
SR Diligence	Mean of self-reported diligence scale
SRSI Attention	Self-reported single item attention to study
SRSI Effort	Self-reported single item effort expended on study

Note. Psy = psychometric; Cons. = consistency; SR = self-report; SI = single item.

Outlier analysis. Recent evidence suggests that Mahalanobis distance can be effective at identifying inattentive responses (Ehlers et al., 2009). Given the large size of the raw item-level data matrix, using all survey items simultaneously would have been computationally intensive. Instead, five Mahalanobis distance measures were computed for each respondent, one for each of the five broad personality factors (60 items per factor). The correlations among the five Mahalanobis distance measures were in excess of .78 ($p < .05$) and were averaged to a single Mahalanobis distance value.

Bogus items. As described earlier, each of the first 10 pages of the survey contained an item that could not possibly be true (or false, depending on the item). If participants indicated a response of either 6 (*agree*) or 7 (*strongly agree*) to true items, the bogus item was considered correct (scored as zero). Other responses were scored as erroneous (i.e., assigned a "1"). In all, there were 10 scored bogus items variables as well as an overall response quality indicator computed as the sum of the bogus item flag variables. However, initial analyses indicated that the bogus item contained on page four of the survey was not interpreted as literally as intended (see Table 1), so it was dropped from further consideration and the sum variable had a possible range from 0 to 9. This appears to be a context effect (Schwarz, 1999) in which, when embedded among personality items, "never listen" was interpreted much more figuratively than we had anticipated.

Consistency indices. Consistency indices can be formed by examining the differences among responses to items that are highly similar in content. Conversely, distance measures of items thought to be antonyms can be computed as well. Goldberg (2000, cited in Johnson, 2005) suggested a method called Psychometric Antonyms, in which correlations among all survey items are computed post hoc and 30 item pairs with the largest negative correlations are identified. The *Psychometric Antonyms* index is then computed as the within-person correlation across these 30 pairs of items. We used a similar index in the current study. However, rather than using 30 item pairs, we sought to ensure item pairs were truly opposite in meaning and only retained item pairs with a negative correlation stronger than –.60. As a result, this index included only five item pairs.

We also developed a similar index that we call the *Psychometric Synonyms* index, which was formed in the same way as the Psychometric Antonyms measure, except that item pairs with the largest positive correlations were used. As before, within-person correlations were computed across item pairs exceeding this +.60 threshold. There were 27 such pairs.

An additional index recommended by Jackson (1976, as cited in Johnson, 2005) was examined which we termed the *Even-Odd Consistency* measure. With this approach, unidimensional scales are divided using an even-odd split based on the order of appearance of the items. An even subscale and also an odd subscale score is then computed as the average response across subscale items. A within-person correlation is then computed based on the two sets of subscale scores for each scale. We formed subscales from all 30 IPIP facets (Goldberg, 1990) as well as the Psychopathy scale (Levenson et al., 1995) and the Narcissm scale (Raskin & Terry, 1988), as our analyses suggested all had relatively high (>.75) coefficient alpha values. Jackson also recommended that the measure be corrected using the Spearman-Brown split half prophecy formula. Note that very inconsistent response patterns can lead to a negative intraindividual correlation. When this occurs, the Spearman Brown prophecy formula can lead to correlation estimates in excess of negative 1.0. For such values, a value of –1.0 was assigned. This index is a more elaborate version of simply examining within-person variance on a unidimensional set of items.

Response pattern. Response patterns in which respondents consistently respond with the same answer (e.g., "5") can be identified via an approach recommended by Johnson (2005). This index, termed *LongString* is computed as the maximum number of consecutive items on a single page to which the respondent answered with the same response option. For instance, if the respondent indicated "2—disagree" for 12 items in a row but otherwise varied his or her response, 12 would be the LongString value for that respondent. While Johnson computed this index for each individual response option (e.g., the maximum number of consecutive responses of 1, 2, etc.), we simply computed the maximum number of items with consecutive response, regardless of the value of that response. A short Visual Basic for Applications program in Microsoft Excel was used to compute this index for each survey page. Additionally, an overall measure (*Avg LongString*) was computed as the average of the LongString variable, averaged across the nine webpages that included 50 items. A second overall index (*Max LongString*) was computed as the maximum LongString variable found on any of the webpages.

Results

Research Question (RQ) 1: Instructions. RQ 1 concerned whether instruction set could impact the quality of the data. There were small but significant differences for some of our variables across instruction set conditions. For instance, one-way analyses of variance (ANOVAs) indicated that there were significant main effects of instruction set on the number of bogus items erroneously endorsed and self-reported attention given to the study (see Table 3). Post hoc Tukey tests indicated significant differences between the anonymous and identified (no warning) conditions for the number of bogus items and self-reported attention variables with the identified condition having fewer bogus items flagged and greater self-reported attention. There were no differences across conditions for any of the other data quality indicators (see Table 3).

RQ 2: Construct validity of indices. RQ 2 concerned the correlations among the various methods of assessing response quality. If different indicators are highly correlated, then it is feasible for researchers to choose one of the less computationally intensive methods of data screening. For these analyses, participants across all three conditions were combined. Table 4 presents the within-condition (identified, anonymous, etc.) pooled correlations among the measures.

Table 3
Results of One-Way Analysis of Variance and Tukey Post Hoc Pairwise Tests by Study Condition

			M			
Indicator	*F*	*p*	Anonymous	Identified	Warning	Pooled *SD*
No. Bogus items	5.24	.006	1.50^{ab}	0.93^{a}	0.79^{b}	1.91
SRSI attention	3.46	.033	4.04^{a}	4.29^{a}	4.17	0.77
SR attitude	2.94	.054	4.04	4.06	3.75	1.12
SR diligence	0.28	.76	5.45	5.46	5.53	0.99
SRSI effort	0.70	.50	3.95	4.01	4.06	0.79
Psychometric Synonyms	1.34	.26	0.69	0.74	0.73	0.23
Psychometric Antonyms	0.96	.38	0.69	0.76	0.71	0.36
Even-Odd consistency	1.23	.29	0.69	0.72	0.74	0.28
Average LongString	0.66	.52	4.15	4.02	3.64	3.67
Max LongString	0.79	.46	7.10	6.18	6.18	7.01
Average Mahalanobis D	2.46	.09	61.05	55.01	56.44	23.45
Crown-Marlow	0.12	.89	4.01	4.03	4.04	0.56
BIDR-*SD*	0.39	.68	4.18	4.15	4.21	0.58
BIDR-IM	1.08	.34	3.88	3.75	3.79	0.69
IPIP *SD*	0.75	.47	4.39	4.31	4.40	0.64
Total Study Time	0.32	.72	49.65	50.72	51.73	15.93

Note. SR = self-report; SI = single item; BIDR = Balanced Inventory of Desirable Responding scale (Paulhus, 1984, 2002); IM = impression management; IPIP = International Personality Item Pool (Goldberg, 1999). *df* = 2, 383 for all analyses other than Total Study Time, where *df* = 2, 228. Superscripts indicate significant Tukey post-hoc pairwise comparison.

Table 4
Pooled Within-Condition Correlations Among Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Total Minutes	1.00															
2. Sum of Bogus	−0.22	1.00														
3. Psy Antonyms	0.13	−0.37	1.00													
4. Psy Synonyms	0.22	−0.65	0.44	1.00												
5. Even Odd consistency	0.27	−0.62	0.35	0.76	1.00											
6. LongString Avg	−0.12	0.33	−0.03	−0.24	−0.27	1.00										
7. LongString Max	−0.16	0.37	−0.06	−0.30	−0.27	0.83	1.00									
8. Mahalanobis D	−0.23	0.39	−0.22	−0.57	−0.57	0.10	0.22	1.00								
9. SR SI Use Me	0.19	−0.47	0.24	0.41	0.39	−0.25	−0.25	−0.32	1.00							
10. SR Diligence	0.22	−0.43	0.29	0.51	0.43	−0.16	−0.18	−0.24	0.43	1.00						
11. SR Attitude	0.01	−0.04	0.09	0.18	0.18	0.02	0.05	−0.08	0.08	0.39	1.00					
12. SR SI Attention	0.10	−0.37	0.20	0.42	0.38	−0.16	−0.17	−0.29	0.41	0.55	0.30	1.00				
13. SR SI Effort	0.31	−0.38	0.25	0.43	0.42	−0.16	−0.17	−0.27	0.44	0.63	0.28	0.62	1.00			
14. Crowne-Marlow	0.08	0.05	0.04	0.11	0.13	−0.05	−0.05	−0.20	−0.03	0.15	0.23	−0.05	0.11	1.00		
15. BIDR Self Dec.	0.07	−0.10	0.04	0.17	0.20	−0.01	−0.01	−0.25	0.01	0.11	0.12	0.07	0.09	0.45	1.00	
16. BIDR IM	0.09	0.06	0.03	0.11	0.11	0.02	−0.01	−0.18	−0.03	0.14	0.16	0.02	0.16	0.71	0.34	1.00
17. IPIP Social Desirability	0.08	−0.08	0.09	0.20	0.18	−0.04	−0.05	−0.19	0.03	0.22	0.15	0.07	0.21	0.65	0.36	0.81

Note. Psy = psychometric; SR = self-report; SI = single item; BIDR = Balanced Inventory of Desirable Responding scale (Paulhus, 1984, 2002); IM = impression management; IPIP = International Personality Item Pool (Goldberg, 1999). $N = 385$. Correlations > ~|.13| are significant at the $p < .01$ level. Standard deviations across imputed samples were small and are not reported for clarity.

While Table 4 is rich with information, there are three primary notable findings. First, the objective measures of data quality were only slightly to moderately correlated across index types. This is not surprising, given that consistency indices assess a very different type of careless response (random responding) than the LongString indicator. Thus, a single indicator of data quality is unlikely to be sufficient for full data screening.

Second, the self-reported indicators of data quality exhibited moderate correlations with one another, indicating that respondents view study interest as different from diligence. Moreover, the UseMe variable did not correlate higher than .5 with any other self-reported measure, which indicates a lack of consistency across respondents with regard to why they might believe their data to be suitable for inclusion in the study. Perhaps more important, the self-reported indices correlated at low or moderate levels with the other indices of careless responses, suggesting that self-report alone is not sufficient to identify careless responses.

Third, careless responding is clearly a different phenomenon than socially desirable responding. In some ways, it is the opposite. Responding in a socially desirable manner necessitates deliberate mental processing of an item and the appropriateness of response. Based on our findings, screening on social desirability would have no utility for removing careless respondents.

EFA. Response time data for 155 persons for which responses either took longer than 15 min on a given page or had multiple submissions were imputed using the Multiple Imputation procedure in SAS 9. Five imputed data sets were created with imputations based on all 30 facet-level personality scales. Imputations occurred at the individual page level and total study completion times were computed as the sum of individual page response times for both complete ($N = 231$) and imputed data ($N = 155$).

An exploratory factor analysis of Total Minutes, Sum of Bogus Items, Psychometric Antonyms, Psychometric Synonyms, Even Odd Consistency, Avg. LongString, Max LongString, Mahalanobis D, SRSI Use Me, SRSI Attention, SRSI Effort, and SR Diligence was conducted on the five replication imputed data sets. For each imputed data set, the within-condition (identified, anonymous, etc.) pooled covariance matrix was analyzed in order to remove the potential influence of mean differences across conditions. Principal axis factoring with squared multiple correlation prior communality estimates was specified as was promax oblique rotation in SAS 9. While the third eigenvalue was rather small (and less than 1.0) in each imputed data set, all five showed three clear interpretable factors underlying the indices and a clear break in the associated scree plots. The average first eigenvalues were 4.43 (imputation data set $SD = 0.01$), second eigenvalues of 1.37 ($SD < 0.01$), third eigenvalues of 0.79 ($SD < 0.01$), and fourth eigenvalues of 0.23 ($SD = 0.02$). The average of the rotated factor pattern matrix of loadings and factor correlations are presented in Table 5.

Examining the factor loadings, the three factors are clearly interpretable. The first contains the consistency indices (Psychometric Anonyms, Psychometric Synonyms, Even-Odd Consistency) as well as the Mahalanobis D multivariate outlier index and the sum of the scored bogus items. The second factor was clearly the four self-report indices, with the exception of the self-reported UseMe variable. The third factor was the two LongString indices. Response time failed to load onto a factor, potentially because of the hypothesized nonlinear relationship between time and response quality.

RQs 3 and 4: Latent classes. RQ 3 asked whether there were different types (latent classes) of careless respondents. We approached this question in two ways. First, we ran a latent profile analysis using our indicators of careless responding. Second, we utilized factor mixture modeling with covariates to establish latent class membership. These are detailed in the following.

Latent profile analysis (LPA). We conducted an LPA on the non-self-report indicators of response quality (Total Minutes, Sum of Bogus Items, Psychometric Antonyms, Psychometric Synonyms, Even Odd Consistency, Avg. LongString, Max LongString, and Mahalanobis D) using Mplus 5. Missing data for Total Minutes were handled via full information maximum likelihood per Mplus 5 defaults. We fit models with one to five classes; however, models with more than three latent classes generated estimation errors associated with a nonpositive definite matrix. Inspection of the output of those models revealed further segmentation of very small classes with no effect on the larger classes. Thus, models with fewer than three latent classes were preferred for parsimonious interpretation (cf. Clark et al., 2009) as well as for statistical reasons.

As suggested by Nylund, Asparouhov, and Muthén (2007) we relied primarily on the Bayesian information criterion (BIC) to judge the most appropriate number of classes, although all indices of model fit (e.g., Akaike information criterion [AIC], –2 log likelihood [LL], Adjusted BIC) indicated that the three class model fit better than the two class model, which, in turn, fit better than the one class model. The BIC values for the one-, two-, and three-class models were 12,964.3; 11,699.8; and 10798.2, respectively. The log likelihood, AIC, and Adjusted BIC showed very similar improvement with larger numbers of classes. The class sizes were 342 (89%) for Class 1, 35 (9%) for Class 2, and 9 (2%) for Class 3, while entropy for the model was .997, indicating a high degree of determination in classification.

The variable means associated with each class are presented in Table 6. As can be seen, Classes 2 and 3 responded much more quickly and had many more bogus items flagged than did Class 1. Class 2 responded considerably less consistently than did Class 1, as judged by the Psychometric Anonym and Synonym and Even-Odd Consistency measures. They also had larger multivariate outlier Mahalanobis D values than did Class 1. The defining hallmark of the small Class 3 was very large LongString values. The consistency indices for this class were between those of Classes 1 and 2, which is to be expected of respondents putting the same value for a large number of items.

On the whole, it appears that approximately 11% of the sample was responding in a careless way with the majority of those (9% of the total sample) responding in an inconsistent way utilizing many different response options, and a small minority (2% of the total sample) responding with the same response option for many consecutive items.

Table 5
Rotated Factor Loadings and Factor Correlations

Variable	Factor 1	Factor 2	Factor 3
Factor 1	—		
Factor 2	−.54 (.00)	—	
Factor 3	.32 (.00)	.03 (.00)	—
Total Minutes	.23 (.04)	.09 (.04)	−.06 (.04)
Sum of Bogus	**−.65** (.00)	−.04 (.00)	.16 (.01)
Psy Antonyms	**.47** (.00)	.06 (.00)	.13 (.00)
Psy Synonyms	**.85** (.00)	.04 (.00)	.02 (.00)
Even Odd Cons.	**.83** (.01)	.01 (.01)	−.01 (.00)
LongString Avg.	.07 (.00)	.00 (.01)	**.90** (.00)
LongString Max	−.03 (.00)	.06 (.00)	**.88** (.00)
Mahalanobis D	**−.71** (.00)	.10 (.00)	−.05 (.00)
SR SI Use Me	.28 (.00)	.31 (.00)	−.13 (.01)
SR SI Attention	.11 (.00)	**.71** (.00)	.00 (.00)
SR SI Effort	−.09 (.00)	**.51** (.01)	.13 (.00)
SR Diligence	.06 (.01)	**.68** (.00)	−.03 (.00)

Note. Psy = psychometric; Cons. = consistency; SR = self-report; SI = single item. Values averaged across five imputed data sets with standard deviation in parenthesis. Bold indicates loadings > .40.

Table 6
Latent Profile Analysis Averages of Observed Variables

Variable	Class 1	Class 2	Class 3
Class size	342 (89%)	35 (9%)	9 (2%)
Means			
Total Minutes	54.81	29.75	33.27
Sum of Bogus	0.58	4.99	5.67
Psy Antonyms	0.77	0.19	0.68
Psy Synonyms	0.78	0.19	0.30
Even Odd Cons.	0.79	0.06	0.23
LongString Avg	3.40	4.48	23.14
LongString Max	5.21	8.74	48.00
Mahalanobis D	53.09	100.20	69.84

Note. Psy = psychometric; Cons. = consistency.

Factor mixture model. Next we sought to identify careless responding via a factor mixture model. Factor mixture modeling combines latent classes with confirmatory factor analysis, allowing the specification of a factor structure for indicator variables; thus, latent class membership can be inferred directly from response tendencies on the indicators as well as latent variable means (see Lubke & Muthén, 2005). With a multitude of data from which to choose, we selected facet-score level data from the IPIP Agreeableness measure.[1] The single agreeableness factor was then regressed onto a latent categorical class membership variable. With factor mixture models, covariates can be specified that can also impact class membership. As we wanted to ensure that our latent classes represented careless response tendencies, rather than mean differences in agreeableness, we specified the Sum of Bogus Items, Psychometric Antonyms, Psychometric Synonyms, Even-Odd Consistency, Avg LongString, Max LongString, and Mahalanobis D indices as latent covariates. We did not include the UseMe variable, as the model could not achieve convergence with a dichotomous covariate. We likewise omitted the Total Minutes indicator based on the missing data and correlations that showed small relationships with the other variables (see Table 4). Figure 1 depicts the factor mixture model.

We hypothesized that some respondents were providing careless responses and therefore expected the relationship between the latent factor and the observed responses to be very small for such responses. For this reason, we allowed our latent classes to have different factor loadings and unique indicator variances. The indicator intercepts and factor variances were constrained to equality across groups (the former to achieve proper model identification and 1.0 for the latter parameter).

[1] A test of covariance matrices across the three study conditions of instruction sets for these variables indicated no difference: comparative fit index = 0.978, Tucker-Lewis index = –0.978, $\chi^2(30) = 42.09$, $p = .07$. For this reason, we combined data across the three conditions for these analyses in order to maximize sample size. We also investigated other personality scales, and we note that we had trouble reproducing latent classes that represented careless responding with some of the personality factors.

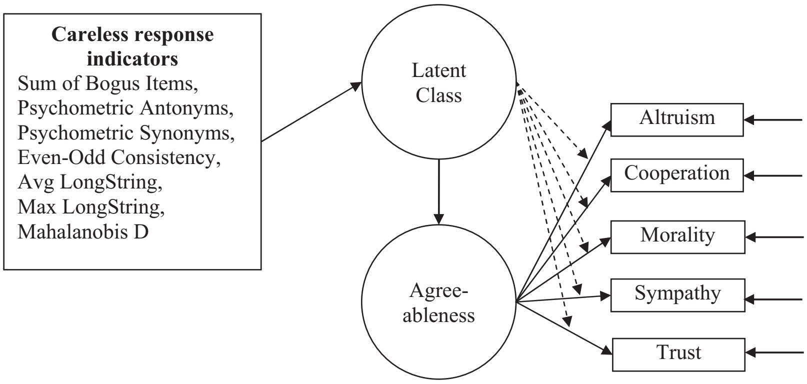

Figure 1. Factor mixture model.

We investigated a three latent class model as was found with our LPA. However, the three resultant classes did not match well to those found in the LPA; thus, we utilized a two-class solution, which resulted in very similar classes to those identified earlier in the LPA. Given that facet-level scores were utilized in these analyses, it is possible that the few respondents indicating the same response option for several consecutive items resulted in facet scores not altogether different from those of conscientious respondents. There were 45 respondents in Class 1 (careless responders) and 336 in Class 2; entropy was .984. As expected, factor loadings for the careless respondents in Class 1 were much lower than Class 2 (see Table 7). Class 1 also had a much lower factor mean value than Class 2 (d = 5.58), although differences in factor means should be judged with great caution when groups' factor loadings drastically differ.

As some of the careless responding indices were highly correlated, multicolinearity among the latent regression coefficients makes such comparisons problematic. We attempted to run separate mixture models, using a single careless responding index in each (as suggested by B. Muthén, personal communication, September 16, 2011). However, class membership changed slightly when we used different indices as covariates as these influence the determination of the latent class (see Lubke & Muthén, 2005). In the end, we exported maximum likelihood latent class membership based on probabilities (which were typically at or near 1.0 or 0, given the entropy of .984), merged the class data with our original data, and examined differences in the careless response indicators across the two latent classes using logistic regression, with class membership as the dichotomous dependent variable.

Table 7
Standardized Results of Factor Mixture Model

Variable	Class 1 (n = 45): Careless responders	Class 2 (N = 336)
Latent mean	−5.58	0.0
Factor loadings (uniqueness)		
Altruism	.55 (.69)	.83 (.32)
Cooperation	.12 (.99)	.65 (.58)
Morality	.29 (.91)	.54 (.71)
Sympathy	.18 (.97)	.60 (.64)
Trust	.11 (.99)	.61 (.63)

Regression of latent class membership on latent indicators

	B (p)	Odds ratio
Sum of Bogus Items	2.43 (.01)	11.34
Psychometric Antonyms	8.95 (.03)	7,666.30
Psychometric Synonyms	−29.23 (.01)	<0.01
Even-Odd Consistency	−22.99 (.01)	<0.01
Avg. LongString	0.31 (.31)	1.36
Max LongString	−0.25 (.05)	0.78
Mahalanobis D	−0.17 (.01)	0.85

Note. Factor variances constrained to 1.0 in each class. Indicator intercepts were similarly constrained and were as follows for the factors (5.45, 4.57, 5.09, 4.49, 4.52). Odds ratios < 0.01 were positive.

Given the correlations among the predictors, each index was utilized as a predictor in a separate model. The results of these analyses are given in Table 8. As can be seen, by several criteria, the Psychometric Synonyms index did the best job of predicting class membership, particularly in predicting membership in the careless response class (Class 1). Moreover, model fit was better for this index. The Even-Odd consistency indices and Sum of the Bogus items also worked well. Given that there were only two classes, it was not surprising that the LongString indices did not perform well. However, we believe that these methods will work well for screening the relatively few respondents utilizing the same response option on many consecutive items.

RQ5: Self-report. Several of the indices examined are difficult to compute or otherwise require special procedures. One convenient alternative is to simply ask respondents to indicate whether their data should be used. The SRSI UseMe variable provides a straightforward dichotomous assessment. On the whole, 90% of respondents indicated that their data should be used. Examining Table 5 shows that self-report indices are only moderately correlated with each other. Moreover, the SRSI UseMe correlated as highly with the number of bogus items missed as it did with any other self-report variable. Thus, it is possible that respondents may be forthcoming about not using their data based on their behaviors but may not be willing to admit to putting forth little effort when directly asked.

Table 8 indicates that the self-report indices were of mixed utility in predicting latent class membership, as derived from the factor mixture model. Of the self-report indices, SRSI UseMe and the SR Diligence scale appeared more useful than the SRSI Attention and Effort items. While the SRSI UseMe indicator had

Table 8
Results of Logistic Regression Models Predicting Factor Mixture Model Latent Class Membership

Variable	−2 LL	Cox & Snell	Nagelkerke	Overall % correct	Class 1 % correct (Sensitivity)	*B*	*SE*	Wald	Odds ratio
Sum of Bogus	130.55	.32	.62	94.8	64.4	0.96	0.12	63.41	2.62
Psy Antonyms	232.82	.11	.21	88.2	15.6	−2.39	0.37	41.47	0.09
Psy Synonyms	84.75	.40	.77	95.0	73.3	−13.11	1.77	54.65	<0.01
Even Odd Cons.	105.25	.36	.70	94.2	64.4	−10.11	1.39	52.87	<0.01
LongString Avg.	261.54	.04	.08	88.9	4.4	0.26	0.08	10.04	1.29
LongString Max	263.57	.03	.07	87.7	4.4	0.07	0.02	12.64	1.07
Mahalanobis D	203.81	.17	.34	90.6	31.3	0.06	0.01	53.03	1.06
Total Minutes	163.04	.13	.27	92.0	25.0	−0.09	0.02	31.35	0.91
SRSI UseMe	228.41	.11	.21	88.8	43.2	−2.64	0.39	45.24	0.07
SR Diligence	185.15	.21	.40	91.5	36.4	−1.63	0.22	55.44	0.20
SRSI Attention	203.32	.17	.32	89.4	15.9	−1.83	0.27	44.13	0.16
SRSI Effort	202.24	.17	.33	89.6	20.5	−1.81	0.27	45.57	0.16

Note. Psy = psychometric; Cons. = consistency; SR = self-report; SI = single item. Class 1 is careless responding class (N = 45). Class 2 N = 336. Wald = $(B/SE)^2$.

poorer model fit and lower pseudo-R^2 value (potentially because of its dichotomous nature), it did produce a higher Class 1 percentage correct (sensitivity). The SR Diligence scale, on the other hand, resulted in higher pseudo-R^2 values and overall percentage correct. One positive feature of the SRSI UseMe index, however, is that it has a natural cutoff for whether to include the response among the set of valid responses, whereas some cutoff would be required to make such a decision with the SR Diligence and other indices.

RQ6: Response time. RQ 6 asked whether response time can be used in a meaningful way. Total minutes did account for significant variance in class membership, despite not being included in the mixture model used to define latent classes. While this does portend some promising use of response time, pragmatically speaking, some type of cutoff value would need to be established in order to screen out persons responding too quickly. Unfortunately there was no clear break point in the distribution, as individual differences in response time were very large. In our opinion, clear outliers on the low end of a distribution of response time could likely be categorized as careless respondents, although there are better methods of making such characterizations.

Study 2

Method

RQ 7 pertained to the sensitivity of careless response indices. While our Study 1 provided highly useful information, more definitive conclusions can be drawn with simulated data wherein each case is known to follow an appropriate response model or to be "careless." Study 2 involved data simulated as such. There were a multitude of decisions required in order to be able to generate the data. We simulated data for 100 items, similar to what may be administered in common personality measures, such as the IPIP, where there are five factors with 20 items per factor. In order to increase external validity, we based our population data on our sample responses described earlier. We began by selecting 351 responses for which two or fewer bogus items were flagged in our earlier analysis. We then randomly selected 20 items from each of the Big 5 personality scales, subject to the restriction that they load at least .45 onto the common factor. Once we had 20 items per scale, we computed scale scores as the sum of these 20 items and determined the correlations among the scales (see Table 9). Next, for each of the Big 5 factors, we estimated item parameters under the item response theory graded response model (Samejima, 1969) using the Multilog 7.03 program (Thissen, 1991). However, as some items had very few persons utilize the extreme options for our 7-point scale, we collapsed these down to five response by merging response options one through three prior to estimation to avoid estimation errors and parameters with extremely large standard errors (cf. Stark, Chernyshenko, Drasgow, & Williams, 2006).[2] As a result, for each item, one *a* parameter and four *b* parameters (b_1 to b_4) were estimated. In order to be able to generate data with seven response options, we created two additional *b* parameters.[3] These parameters were hereafter treated as the population parameters of our simulated data.

To simulate purposeful response data, we started by simulating theta scores (M = 0, SD = 1) for each of our five personality scales. We used a Cholesky decomposition of the observed correlation among the five factors (see Table 9) to derive weights such that our five theta scores were correlated in the same way as the observed scale scores from Study 1. These thetas were then used to generate data under the graded response model using the population item parameters.

[2] The exception to this process was the neuroticism scale, in which the skew in the raw data was positive and collapsing was across Response Options 5–7.

[3] The seven-response-option b1 parameter was created by taking the estimated b1 parameter and subtracting 2.0 from it. We similarly subtracted 1.0 from the estimated b1 parameter in order to create the new b2 parameter for our seven response options. We added 1.0 and 2.0 for the largest estimated b parameter to create the largest two b parameters for the neuroticism items. This process led to utilization of the extreme response options in rough proportion to the observed data witnessed in our data sample, yet avoided estimation errors associated with insufficient data.

Table 9
Correlations Among Personality Scales

Variable	1	2	3	4	5
1. Agreeableness	1.00				
2. Conscientiousness	.37	1.00			
3. Extraversion	.42	.21	1.00		
4. Neuroticism	−.31	−.27	−.46	1.00	
5. Openness	.15	.04	.09	.03	1.00

In order to simulate the data, a program written for the purpose in Microsoft Excel's Visual Basic for Application was used. For each item, the relevant theta score, along with population item parameters, was used to compute the probability of response for each of the seven response options. Cumulative intervals were then formed across the range of 0 to 1.0, in which seven ranges corresponded to the seven response options. The interval corresponding to response option one included the range from zero to the value of the probability of response option one (p_1). The interval corresponding to response option two banded the range from p_1 to ($p_1 + p_2$), etc., with the range of response option seven banding ($1 - p_7$) to 1.0. A random number with a uniform distribution was then simulated and the response option with the corresponding range that contained the random number was then specified as the response for the item. This process was repeated for each of the 20 items per each of the five latent constructs.

Variables manipulated. We simulated 16 data conditions, each containing 100 sample replications and 1,000 simulated respondents. We varied three variables: the extent of carelessness of the data (full or partial), the type of carelessness (uniformly random, normally distributed random), and the percentage of careless respondents in the sample replication (5%, 10%, 15%, or 20%). Our design was 2 × 2 × 4, with a total of 1,600 replication samples. In order to simulate careless data, we used a random response model using either a uniform distribution or a normal distribution. For the portion of the sample for which responses were fully random, all 100 items were generated by simulating a random integer between 1 and 7 from the relevant (uniform or normal) distribution. While little guidance is available in the literature regarding the nature of random response, we sought to simulate truly random (uniform distribution) response as well as random response for persons that may stick primarily to the scale midpoints over concerns of being identified via outlier analysis.

We created partially random data as partially random patterns were present in our data as well as that of others (e.g., Baer et al., 1997; Berry et al., 1992). In order to create partially random data, for each of the relevant cases, we first simulated purposeful response data as described earlier, but then randomly selected approximately 25% of the 100 items for which we replaced the purposeful response with a random integer between 1 and 7. We did not simulate careless data by which respondents select the same response to many consecutive items; it is clear from Study 1 that the LongString indices are best able to detect such response patterns.

Indices examined. We examined the performance of several consistency indices. We did not examine the performance of the two LongString indices, the self-reported indices, or response time. Computation of the consistency indices can be rather intensive, particularly if different variables are used to form the indices across the 1,600 samples. Therefore, we did not examine the empirical correlations among the items in each sample in order to choose which items to use in the Psychometric Synonyms and Antonyms measures. Rather, we used all item pairs from among the 100 items with correlations above ± .60 in our sample data from which item parameters were derived. With replicated sample sizes of 1,000, there should be relatively little sampling error and thus not a high degree of variability across samples with respect to which items were highly correlated. There were six such pairs used in the Psychometric Antonyms and nine available for the Psychometric Synonyms. For the Even-Odd Consistency index, we created 10-item subscales for each of the five 20-item personality measures using the methods described previously. The Mahalanobis distance measure was computed for each of the five subscales and averaged into a single index.

Once the data were created, for each replication, we again examined the estimated logistic regression coefficients, model fit, Cox and Snell and Nagelkerke pseudo-R^2 coefficients, and correct classification percentages associated with regressing the dichotomous true nature of the data (0 = purposeful, 1 = careless) on each of the indices examined in separate model runs. In each regression analysis, the criterion variable was the dichotomous variable representing whether the simulated examinee represented a valid responder or a careless responder. The sample size for each regression was 1,000 (the number of simulated examinees per replication), and these analyses were performed in each of the 100 replications for each condition. The results of each of the 100 replication regression analyses were averaged and reported at the condition level.

Results

Uniformly random careless data. The results of our simulations involving uniformly random careless data can be found in Table 10. In interpreting the results, we relied primarily on the Cox and Snell as well as Nagelkerke pseudo-R^2 values as well as sensitivity and specificity values. While Wald statistics indicate whether the index could significantly predict simulated carelessness, the pseudo-R^2 and classification analyses more directly assess the relative efficacy of the indices, as they are less contingent upon power. Sensitivity values indicate the percentage of careless respondents correctly classified as careless (i.e., true positives). Specificity values indicate the percentage of simulated purposeful respondents correctly categorized as such (i.e., true negatives).

As expected, efficacy of the methods was higher when a totally, rather than partially, random response model was used to simulate data. A larger percentage of careless responders also resulted in somewhat better performance for each index. On the whole, the Mahalanobis D measure performed best with the data simulated with respect to variance explained and correct classification. Our logistic regression analyses encountered errors with Mahalanobis D, in which no maximum likelihood estimate of the slope parameter was available, as Mahalanobis D was able to nearly perfectly categorize each simulated respondent in each replication as careless or purposeful. We believe that this index performed well because of the skewed nature of the data simulated. When data are sufficiently skewed, careless responses from the infrequently used end of the observed response distribution are likely to result in a

Table 10

Logistic Regression Results for Simulated Totally and Partially Careless Data Under Uniform Random Distribution

% Random	Variable	−2 LL	Cox & Snell	Nagelkerke	*B*	*SE*	Wald	Sensitivity	Specificity
					Totally careless				
5	Psy Synonyms	279.86 (2.23)	.12 (.02)	.35 (.05)	−4.06 (0.46)	0.42 (0.04)	93.38 (6.85)	15.33 (7.12)	99.02 (0.31)
	Psy Antonyms	318.86 (15.42)	.08 (.01)	.24 (.04)	−2.99 (0.32)	0.35 (0.02)	74.49 (9.97)	6.98 (4.64)	99.58 (0.22)
	Even Odd Cons.	116.64 (25.9)	.25 (.02)	.75 (.06)	−13.15 (3.21)	1.75 (0.47)	57.59 (7.87)	73.58 (6.34)	99.66 (0.12)
	Mahalanobis D	[a]	.33 (.00)	1.00 (.00)	[a]	[a]	[a]	99.90 (0.44)	100.00 (0.00)
10	Psy Synonyms	426.00 (24.84)	.20 (.02)	.42 (.04)	−4.31 (0.36)	0.35 (0.03)	15.26 (5.99)	33.12 (5.87)	97.50 (0.29)
	Psy Antonyms	505.99 (23.63)	.14 (.02)	.29 (.04)	−2.99 (0.29)	0.27 (0.01)	12.32 (12.06)	19.26 (6.36)	98.38 (0.33)
	Even Odd Cons.	176.77 (37.23)	.38 (.02)	.79 (.05)	−14.59 (3.26)	1.60 (0.40)	85.02 (9.09)	78.87 (5.16)	99.36 (0.2)
	Mahalanobis D	[a]	.48 (.00)	1.00 (.00)	[a]	[a]	[a]	99.91 (0.29)	100.00 (0.00)
15	Psy Synonyms	536.79 (27.81)	.27 (.02)	.47 (.04)	−4.41 (0.32)	0.32 (0.02)	19.57 (5.82)	44.66 (5.54)	95.96 (0.41)
	Psy Antonyms	634.77 (25.54)	.19 (.02)	.34 (.04)	−3.11 (0.24)	0.25 (0.01)	159.77 (11.36)	31.89 (4.78)	96.74 (0.45)
	Even Odd Cons.	218.79 (36.42)	.47 (.02)	.82 (.03)	−15.57 (2.94)	1.57 (0.34)	99.33 (8.31)	82.20 (3.27)	99.21 (0.22)
	Mahalanobis D	[a]	.57 (.00)	1.00 (.00)	[a]	[a]	[a]	99.95 (0.18)	100.00 (0.00)
20	Psy Synonyms	617.92 (32.52)	.32 (.02)	.50 (.03)	−4.53 (0.34)	0.30 (0.02)	221.85 (5.5)	53.72 (4.08)	94.33 (0.47)
	Psy Antonyms	741.72 (26.18)	.23 (.02)	.36 (.03)	−3.14 (0.23)	0.23 (0.01)	186.85 (1.48)	40.26 (4.21)	94.82 (0.48)
	Even Odd Cons.	258.04 (35.32)	.53 (.02)	.83 (.03)	−15.72 (2.58)	1.48 (0.28)	114.49 (11.48)	83.81 (2.48)	98.90 (0.26)
	Mahalanobis D	[a]	.63 (.00)	1.00 (.00)	[a]	[a]	[a]	99.96 (0.14)	100.00 (0.01)
					Partially careless				
5	Psy Synonyms	375.97 (12.42)	.03 (.01)	.08 (.04)	−1.96 (0.42)	0.37 (0.02)	29.12 (12.51)	0.14 (0.58)	99.97 (0.07)
	Psy Antonyms	382.68 (11.76)	.02 (.01)	.06 (.03)	−1.48 (0.46)	0.34 (0.02)	21.36 (12.09)	0.00 (0.00)	99.99 (0.04)
	Even Odd Cons.	378.33 (15.64)	.02 (.02)	.07 (.05)	−6.60 (2.78)	1.41 (0.26)	22.76 (13.34)	2.56 (3.10)	99.78 (0.14)
	Mahalanobis D	97.40 (28.8)	.26 (.02)	.79 (.06)	0.65 (0.12)	0.09 (0.03)	58.99 (12.93)	75.29 (8.18)	99.49 (0.19)
10	Psy Synonyms	601.06 (19.3)	.05 (.02)	.11 (.04)	−2.10 (0.36)	0.29 (0.01)	53.76 (17.25)	1.55 (2.27)	99.57 (0.40)
	Psy Antonyms	617.80 (18.63)	.04 (.02)	.07 (.04)	−1.51 (0.40)	0.26 (0.01)	37.35 (18.03)	0.54 (1.16)	99.89 (0.19)
	Even Odd Cons.	608.27 (24.18)	.04 (.02)	.09 (.05)	−7.49 (2.51)	1.24 (0.17)	37.67 (17.82)	5.28 (4.33)	99.42 (0.26)
	Mahalanobis D	147.23 (37.65)	.40 (.02)	.83 (.05)	0.70 (0.11)	0.08 (0.02)	88.50 (16.49)	81.72 (5.67)	99.01 (0.29)
15	Psy Synonyms	772.20 (25.4)	.07 (.02)	.13 (.04)	−2.19 (0.38)	0.26 (0.01)	72.85 (21.08)	5.40 (4.02)	98.56 (0.67)
	Psy Antonyms	794.23 (24.27)	.05 (.02)	.09 (.04)	−1.58 (0.36)	0.22 (0.01)	53.20 (21.93)	3.32 (3.84)	99.37 (0.58)
	Even Odd Cons.	780.23 (29.86)	.07 (.03)	.12 (.05)	−8.60 (2.42)	1.21 (0.14)	50.61 (19.49)	8.87 (5.23)	98.94 (0.37)
	Mahalanobis D	178.82 (39.08)	.49 (.02)	.85 (.04)	0.73 (0.09)	0.07 (0.02)	107.59 (18.05)	85.41 (3.98)	98.59 (0.34)
20	Psy Synonyms	912.25 (22.08)	.09 (.02)	.14 (.03)	−2.18 (0.27)	0.24 (0.01)	84.57 (17.8)	10.20 (4.69)	96.92 (0.79)
	Psy Antonyms	940.55 (20.81)	.06 (.02)	.09 (.03)	−1.54 (0.25)	0.20 (0.01)	60.06 (18.38)	6.63 (4.53)	98.48 (0.82)
	Even Odd Cons.	922.75 (26.15)	.08 (.02)	.12 (.04)	−8.91 (2.0)	1.17 (.13)	58.33 (15.71)	11.45 (4.44)	98.26 (0.45)
	Mahalanobis D	213.53 (38.83)	.55 (.02)	.86 (.03)	0.76 (0.09)	0.07 (0.01)	127.85 (19.11)	87.02 (2.80)	98.03 (0.04)

Note. Psy = psychometric; Cons. = consistency. Numbers presented are means of model parameters across 100 replication conditions (*SD*s in parentheses). Wald = $(B/SE)^2$. *SD*s of the *B* coefficients for Even Odd Consistency are considerably larger than estimated mean *SE*; therefore, Wald tests for Even Odd Consistency should be treated with caution.

[a] Statistics not available due to "complete separation of data points" in which some value of Mahalanobis D could be found that accurately categorizes nearly all simulated respondent as either careless or purposeful.

greater impact on the multivariate outlier statistic than under conditions in which greater utilization of all response scales is observed. However, further simulation work is required to confirm this explanation. Among the three consistency indices, the Even-Odd Consistency index performed better than the Psychometric Synonyms and Antonyms measures. These findings contrast somewhat with those of our Study 1 results. There were several differences between these simulated data and those observed data. The largest difference was the number of items that were sufficiently correlated (>.60) to be used in the synonym approach. In Study 1, there were 27 such pairs, whereas there were only nine pairs in our simulation. However, the same is true of the Even-Odd Consistency measures, which was previously based on 32 subscale pairs and is here based on only five. In comparing the performance, however, we believe that reliability best explains these differences. The Psychometric Synonyms and Antonyms approaches used correlations across nine and six item pairs, respectively. Thus the building blocks of these indices were single items. Conversely, the building blocks of the Even-Odd Consistency measure were five 10-item subscale pairs, which are more reliable than single-item pairs. As with the observed data, we found poorer performance for our Psychometric Antonyms index than for any other index.

Normally distributed random careless data. Results for careless data simulated via a normal distribution varied substantially depending on whether all items were subject to careless response or whether only a portion were. When all items were subject to careless responding (total carelessness conditions), the Mahalanobis D measure was woefully inadequate for detecting random response under a normal distribution (see Table 11). Pseudo-R^2 values were very low, and sensitivity often approached zero. This was in stark contrast to our conditions of uniformly distributed careless response, in which the Mahalanobis D worked well. With the normally distributed, totally careless data simulated here, the Even-Odd Consistency measure outperformed the others by a good margin with respect to model fit; pseudo-R^2; and most importantly, sensitivity.

Table 11
Logistic Regression Results for Simulated Totally and Partially Careless Data Under Normal Random Distribution

% Random	Variable	−2 LL	Cox & Snell	Nagelkerke	*B*	*SE*	Wald	Sensitivity	Specificity
				Totally careless					
5	Psy Synonyms	282.45 (18.11)	.11 (.02)	.33 (.05)	−3.93 (0.43)	0.42 (0.03)	89.52 (6.56)	13.04 (5.94)	99.14 (0.29)
	Psy Antonyms	312.41 (19.57)	.08 (.02)	.24 (.05)	−2.97 (0.40)	0.35 (0.02)	71.81 (11.38)	6.01 (5.80)	99.63 (0.23)
	Even Odd Cons.	107.2 (25.92)	.25 (.02)	.77 (.06)	−14.35 (3.12)	1.94 (0.48)	55.89 (8.92)	75.42 (5.96)	99.68 (0.11)
	Mahalanobis D	343.67 (11.75)	.05 (.01)	.16 (.03)	0.25 (0.03)	0.04 (0.00)	47.07 (7.89)	0.38 (0.93)	99.86 (0.13)
10	Psy Synonyms	435.58 (20.53)	.19 (.02)	.40 (.03)	−4.14 (0.30)	0.34 (0.02)	145.94 (5.88)	3.24 (5.37)	97.59 (0.35)
	Psy Antonyms	498.81 (23.33)	.14 (.02)	.28 (.04)	−2.97 (0.27)	0.27 (0.01)	118.37 (11.83)	18.49 (5.77)	98.38 (0.42)
	Even Odd Cons.	177.14 (53.98)	.38 (.04)	.79 (.09)	−14.47 (3.20)	1.60 (0.38)	81.55 (12.03)	78.67 (8.55)	99.42 (0.20)
	Mahalanobis D	625.97 (7.55)	.02 (.01)	.05 (.02)	0.13 (0.02)	0.03 (0.00)	23.58 (7.03)	0.00 (0.00)	99.99 (0.04)
15	Psy Synonyms	545.33 (27.84)	.26 (.02)	.45 (.04)	−4.28 (0.33)	0.31 (0.02)	187.56 (6.29)	42.41 (4.79)	95.89 (0.43)
	Psy Antonyms	637.25 (25.9)	.18 (.02)	.32 (.04)	−2.99 (0.25)	0.24 (0.01)	153.88 (11.90)	29.68 (4.96)	96.78 (0.45)
	Even Odd Cons.	206.11 (36.55)	.47 (.02)	.83 (.03)	−16.48 (3.14)	1.67 (0.34)	97.95 (9.23)	83.22 (3.31)	99.25 (0.19)
	Mahalanobis D	841.44 (2.9)	.00 (0)	.01 (.01)	0.04 (0.02)	0.02 (0.00)	3.99 (2.90)	0.00 (0.00)	100.00 (0.00)
20	Psy Synonyms	623.59 (31.98)	.31 (.02)	.5 (.03)	−4.47 (0.32)	0.30 (0.02)	22.15 (5.75)	53.18 (4.51)	94.28 (0.49)
	Psy Antonyms	737.74 (32.97)	.22 (.03)	.36 (.04)	−3.09 (0.27)	0.23 (0.01)	183.33 (14.12)	39.09 (4.73)	95.01 (0.47)
	Even Odd Cons.	246.33 (34.49)	.53 (.02)	.84 (.03)	−17.14 (2.75)	1.63 (0.31)	112.03 (9.71)	84.53 (2.45)	98.96 (0.26)
	Mahalanobis D	997.85 (2.84)	.00 (0)	.00 (0)	−0.03 (0.02)	0.02 (0.00)	2.89 (2.75)	0.00 (0.00)	100.00 (0.00)
				Partially careless					
5	Psy Synonyms	380.69 (7.77)	.02 (.01)	.05 (.02)	−1.58 (0.37)	0.39 (0.02)	18.07 (8.48)	0.02 (0.20)	100.00 (0.01)
	Psy Antonyms	384.44 (7.26)	.01 (.01)	.04 (.02)	−1.19 (0.31)	0.35 (0.02)	12.90 (7.32)	0.00 (0.00)	100.00 (0.00)
	Even Odd Cons.	386.29 (7.78)	.01 (.01)	.03 (.02)	−4.92 (2.25)	1.51 (0.33)	11.44 (7.76)	0.64 (1.33)	99.87 (0.09)
	Mahalanobis D	257.34 (17.3)	.13 (.02)	.40 (.05)	0.39 (0.04)	0.04 (0.00)	85.89 (4.16)	22.90 (6.75)	99.3 (0.21)
10	Psy Synonyms	618.14 (10.49)	.03 (.01)	.07 (.02)	−1.68 (0.27)	0.29 (0.01)	33.71 (1.56)	0.16 (0.47)	99.92 (0.13)
	Psy Antonyms	627.19 (8.66)	.02 (.01)	.05 (.02)	−1.23 (0.23)	0.26 (0.01)	23.68 (8.99)	0.01 (0.10)	100.00 (0.02)
	Even Odd Cons.	628.94 (10.45)	.02 (.01)	.04 (.02)	−5.84 (2.00)	1.32 (0.19)	19.97 (9.47)	1.51 (1.46)	99.71 (0.16)
	Mahalanobis D	406.94 (22.69)	.22 (.02)	.45 (.04)	0.41 (0.03)	0.04 (0.00)	134.77 (4.32)	35.95 (5.49)	98.03 (0.35)
15	Psy Synonyms	800.84 (12.94)	.04 (.01)	.08 (.02)	−1.71 (0.26)	0.26 (0.01)	45.01 (12.28)	1.31 (1.37)	99.46 (0.41)
	Psy Antonyms	813.19 (12.37)	.03 (.01)	.05 (.02)	−1.24 (0.24)	0.22 (0.01)	32.50 (12.11)	0.52 (1.26)	99.88 (0.22)
	Even Odd Cons.	814.89 (15.04)	.03 (.01)	.05 (.03)	−6.32 (2.02)	1.25 (0.14)	26.49 (12.85)	2.85 (2.27)	99.41 (0.03)
	Mahalanobis D	527.85 (27.15)	.27 (.02)	.48 (.03)	0.42 (0.03)	0.03 (0.00)	169.40 (4.36)	44.46 (4.62)	96.62 (0.41)
20	Psy Synonyms	944.99 (17.17)	.05 (.02)	.09 (.03)	−1.74 (0.28)	0.24 (0.01)	54.39 (15.34)	3.80 (2.79)	98.31 (0.81)
	Psy Antonyms	962.14 (10.98)	.04 (.01)	.06 (.02)	−1.23 (0.18)	0.20 (0.01)	37.91 (1.53)	1.86 (1.88)	99.44 (0.49)
	Even Odd Cons.	965.41 (13.31)	.03 (.01)	.05 (.02)	−6.35 (1.58)	1.18 (0.11)	29.49 (1.36)	3.92 (2.32)	99.07 (0.35)
	Mahalanobis D	637.17 (28.71)	.30 (.02)	.48 (.03)	0.41 (0.03)	0.03 (0.00)	193.88 (5.30)	49.46 (3.58)	94.84 (0.46)

Note. Psy = psychometric; Cons. = consistency. Numbers presented are means of model parameters across 100 replication conditions (*SD*s in parentheses). Wald = $(B/SE)^2$. *SD*s of the *B* coefficients for Even Odd Consistency are considerably larger than estimated mean *SE*; therefore, Wald tests for Even Odd Consistency should be treated with caution.

Results for partially random normally distributed careless responses more closely resembled those of the uniform distribution, in which the Mahalanobis D measure again was the superior indicator of careless response. Interestingly, the Even-Odd Consistency index performed similarly for uniform and normal random distributions of careless data, although it was very strongly affected by the pervasiveness of the carelessness across items in each (being very ineffective when data were only partially careless). The Mahalanobis D measure worked well under most conditions, although it performed very poorly when a substantial number of simulated respondents were responding pervasively carelessly under a normal distribution.

Pseudo-R^2. We also conducted logistic regression analyses using the true population careless/purposeful dichotomous variable as the criterion and all of our consistency indices as well as their interactions as predictors. Our goal was to determine the maximum amount of variance that one could hypothetically account for using these indices. Table 12 indicates that virtually all careless responses can be detected under the conditions simulated with the uniformly distributed, totally random careless data, whereas very high percentages could be identified under uniformly distributed partially random and normally distributed totally random careless data. Results were more modest (Nagelkerke ~ .5) for the normally distributed partially random careless data condition.

Variance decomposition. Additionally, we wanted to determine the effect of each of our manipulated study variables on the efficacy of each index for detecting careless respondents. Earlier, we conducted logistic regression analyses for each of our 100 replication samples using only one of the four indices as a predictor. We harvested the sensitivity (true positive) values obtained from these analyses from all 100 replications for each study condition and used these as a dependent variable in a factorial ANOVA to determine the extent to which sensitivity of each index was influenced by our study variables. Our manipulated study variables were all treated as categorical and included type of generating distribution (uniform or normal distribution), extent of carelessness (totally careless or partial), and base rate of careless respondents (5%, 10%, 15%, or 20%).

Table 12
Average Pseudo-R^2 Values for Simulated Data

	Uniform distribution				Normal distribution			
	Totally random		Partially random		Totally random		Partially random	
% Random	Cox & Snell	Nagelkerke	Cox & Snell	Nagelkerke	Cox & Snell	Nagelkerke	Cox & Snell	Nagelkerke
5	.33	1.00	.27	.83	.28	.85	.15	.44
10	.48	1.00	.41	.85	.42	.87	.23	.48
15	.57	1.00	.50	.87	.52	.91	.29	.50
20	.63	1.00	.55	.87	.58	.92	.32	.51

Note. Averaged across 100 replications.

Computed η^2 values are reported in Table 13. As can be seen, the type of generating distribution had an enormous effect on the Mahalanobis D index, yet very little effect on the other indices. Conversely, Mahalanobis D was affected to a much lesser extent than the other indicators by the extent (totally vs. partially) of carelessness present as well as by the base rate of careless responding. However, there was a large interaction, as discussed earlier, such that Mahalanobis D performed poorly when data were partially careless and normally distributed.

Bogus items. Formal analyses for bogus or instructed response items are not necessary, as the underlying probabilities can be computed to illustrate the efficacy of the items. Presuming that there is a single correct answer to the item, careless responders using a uniform distribution of careless response will have a probability of correct response of $1/j$, where j is the number of response options. This would be best achieved via instructed response items (e.g., "please respond with "disagree" for this item). Thus, the probability of being flagged by the item is $(j - 1)/j$. If more than one such item is used, the probability of being flagged by any of the items is $1 - (1/j)^k$, where k is the number of instructed response items. For two instructed response items and five response options, the probability of being screened out when using random response is .96, while the probability is .98 for seven response options. With three such items, the probability is very close to 1.0 for five or more response options. As such, these types of items should be powerful indicators of respondents not bothering to read the item stem. If respondents were following a random normal distribution with their careless responses, instructed response items requiring an extreme response would be even more efficacious. When coupled with the careless response indicators we empirically examined, nearly all careless responses should be able to be detected when the two are used simultaneously.

General Discussion

This study provides a comprehensive investigation of the indicators of careless response tendencies present in survey data. As such, this study answers several questions not addressed by the extant literature. First, we found a significant but small effect of using instruction sets to influence data quality. There appeared to be small advantages in using identified surveys such that fewer bogus items were endorsed under such conditions and respondents self-reported paying more attention. Strong warnings about violations of the honor code approached significance with respect to decreasing respondent self-reported attitude toward the study and strong warnings provided no tangible benefit over using identified responses without the warning.

Second, we found that the different indicator variables flag different individuals, as is evident in the correlations among all quality indicators (see Tables 4 and 5). Logically this is understandable. For instance, a person randomly responding to items will not be identified by the LongString variable but may be flagged by other quality indicators. Conversely, a person responding with many 4s in a row may not provide suspect answers, as indicated by an internal consistency measure or outlier analysis, but may be identified by the LongString variable. Our factor analysis of the indicators found that the three consistency indices (Psychometric Antonyms, Psychometric Synonyms, and the Even-Odd Consistency measure) loaded onto the same factor as the Mahalanobis D outlier index and the sum of bogus items endorsed. While conceptually different, the Mahalanobis D and the bogus

Table 13
η^2 for Manipulated Variable Effects on Logistic Regression Sensitivity Indices

Manipulated variable	Psychometric synonyms	Psychometric antonyms	Even-odd consistency	Mahalanobis D
Generating Distribution (normal vs. uniform)	.00	.00	.00	.83
Partially vs. Totally Careless Model (Extent)	.67	.56	.98	.02
Base rate of Careless Respondents	.18	.22	.01	.01
Distribution × Extent	.00	.00	.00	.12
Distribution × Base Rate	.00	.00	.00	.00
Extent × Base Rate	.09	.14	.00	.01
Distribution × Extent × Base Rate	.00	.00	.00	.00

items appear to capture the same type of careless responses present in our data as do the consistency indices. Self-report measures and the LongString indices appear fundamentally different. The most effective data screening approach will utilize several data quality indicators simultaneously.

Our simulation results found that Mahalanobis D can be a powerful indicator of careless response, but its efficacy is very much dependent upon the nature of the entire sample. As with any outlier analysis, what is considered an outlier is dependent upon the distribution of responses in the sample. We found that when careless responses followed a uniform random distribution, Mahalanobis D performed well; it similarly performed well when the careless responses followed a normal distribution for only some of the careless data. However, it performed very poorly when careless respondents' careless data followed a normal distribution for all items. In contrast, the Even-Odd Consistency measure was relatively stable in its performance across conditions of uniform or normal distribution of totally random responses. However, the Even-Odd measure performed poorly only when occasional careless responses were provided. Given that the Even-Odd uses subscales, occasional careless responses have relatively little impact on these subscales.

Our LPA and factor mixture model analyses indicated that around 10%–12% of our undergraduate sample belonged to a latent class that can be considered careless in their responses, a number nearly identical to that found by Kurtz and Parish (2001). Examining self-report and response time measures indicated that their use to screen the data is better than doing nothing. Also, while a few persons completed the survey so quickly that careless responses were undoubtedly present, most persons identified as responding carelessly by other methods had response times such that their responses would not have been removed. Such persons may be distracted, engaged in conversation, etc. during the response process. On the whole, response time and self-report measures were not sufficient for a thorough data screening.

Recommendations

Our results afford a number of practical recommendations. First, we encourage the use of identified responses but not the harsh wording in the instruction set used here. While anonymity may afford more accurate reports of sensitive behaviors (e.g., cheating; Ong & Weiss, 2000), several studies have found differences between anonymous and confidential conditions to be minimal (Moore & Ames, 2002) and to have no effect on response rate (Campbell & Waters, 1990).

Second, we strongly endorse bogus items—or, preferably, instructed response items (e.g., "Respond with 'strongly agree' for this item")—for longer survey measures. Such items are easy to create, have the advantage of providing an obvious metric for scoring as correct or incorrect, and are not as vulnerable to figurative interpretation as the bogus items we created. We suggest incorporating approximately one such item in every 50–100 items up to a maximum of three. Respondents may become annoyed at such items if a large number appear.

Our further recommendations are dependent upon the stringency of data screening needed. We believe that every Internet-based survey research study would benefit from incorporating at least one careless response detection method. For instances in which robust correlations are of interest and highly rigorous data scrubbing is unnecessary, we suggest incorporating a simple self-report measure (i.e., "In your honest opinion, should we use your data?"), coupled with a cursory look at response time for outliers. Our results suggest that these methods will effectively screen out some careless responders, although further research is needed to determine the effects of remaining careless responders on data properties. When utilizing survey data collected online from undergraduate student samples, we believe this process should be employed as a minimum. If only post hoc methods are available, then inspection of response time and computation of the Even-Odd Consistency measure are suggested as minimums.

For cases in which more rigorous data cleaning is necessary, we suggest incorporating the instructed response items and computing three post hoc indices. Among the consistency indices, we recommend computing the Even-Odd Consistency index, as this index had slightly more favorable qualities than the Psychometric Synonyms or Antonyms indices. Care must be taken, however, that the scales used to calculate the indices are unidimensional, and several scales must be available in order to compute such an index. We also recommend the Maximum LongString index, which is useful for identifying persons responding to different items with a single response option, although we found a very small number of respondents employed this strategy. We found very little difference in the performance of the Maximum and Average LongString indices. Note that these indices work best when items are randomly ordered and, thus, similar responses across consecutive items are not expected. We also believe the Mahalanobis D measure to be very useful. Unfortunately, each of these three indices requires inspection of their frequency distributions in order to identify a suitable cutoff value for careless response, although probability values can be computed for the Mahalanobis distance measure. In practice, then, some judgment without much in the way of empirical basis will be needed for the implementation of these indices.

For instances in which few items are administered, computation of the consistency indices examined in this study may not be possible. In such cases, we recommend taking a highly internally consistent scale and examining the within-person variance in responses across the scale items. If the scale items correlate highly with one another for the sample as a whole, yet an individual shows excessive variance across his or her responses to those scale items, the person is likely to be responding inconsistently and can be eliminated from the sample.

Limitations

As with any study, this one is limited in scope. While we have every reason to believe that our respondent population is typical of that in most large research universities, we cannot be certain that this is the case. If careless response prevalence varies by location or time, then our results may not generalize to future research. Additionally, while we have included a far more comprehensive set of indices than any previous studies on this topic, there are an almost unlimited number of approaches that could be explored. We encourage future researchers to continue to compare other approaches with those examined here. For instance, alternative approaches, such as offering descriptive feedback to respondents, may provide incentive for more purposeful responses.

Another limitation is the lack of incorporation of explicit instructed response items. During data analysis, we found that one of our bogus items was apparently interpreted much more figuratively than we had intended. Also, we had some concern that otherwise careful respondents might endorse a bogus item because they find it humorous (e.g., “I am paid biweekly by leprechauns”). Moreover, we were forced to choose the level of dis/agreement at which such items were to be considered “correct.” Explicit instructed response items would not exhibit these weaknesses.

Another issue is that self-report measures are based on the assumption that respondents are attentively responding to them. In our study, our final survey page was formatted very differently from the earlier pages including personality items. Moreover, the instructions were presented with several aspects of highlighting to alert the respondent that the study was at a close and that candid responses were necessary. However, we cannot be sure that respondents heeded these features. Researchers hoping to use such measures should clearly delineate them from other study items, perhaps including them along with demographic questions at the end of the study.

Conclusion

There are many benefits of online survey methodology. Our own anecdotal experience is that Internet-based surveys are many orders of magnitude more popular among potential participants than similar studies requiring laboratory participation. For these reasons and others, Internet-based data collection will continue to be the dominant data collection paradigm in survey research. However, it is imperative that Internet survey data be properly screened for careless responses in order to safeguard the integrity of research conclusions.

References

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. doi:10.1207/s15327752jpa6801_11

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

Ben-Porath, Y. S., & Tellegen, A. (2008). *The Minnesota Multiphasic Personality Inventory–2 Restructured Form: Manual for administration, scoring, and interpretation.* Minneapolis, MN: University of Minnesota Press.

Berry, D. T., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–598. doi:10.1016/0272-7358(91)90005-F

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. doi:10.1037/1040-3590.4.3.340

Berry, D. T. R., Wetter, M. W., Baer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991). Detection of random responding on the MMPI-2: Utility of F, back F, and VRIN scales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 418–423. doi:10.1037/1040-3590.3.3.418

Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121–140). San Diego, CA: Academic Press.

Campbell, M. J., & Waters, W. E. (1990). Does anonymity increase response rate in postal questionnaire surveys about sensitive subjects? A randomised trial. *Journal of Epidemiology and Community Health, 44,* 75–76. doi:10.1136/jech.44.1.75

Carrier, L. M., Cheever, N. A., Rosen, L. D., Benitez, S., & Chang, J. (2009). Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans. *Computers in Human Behavior, 25,* 483–489. doi:10.1016/j.chb.2008.10.012

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (Eds.). (1993). *16PF Questionnaire* (5th ed.). Champaign, IL: Institute for Personality and Ability Testing.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment, 15,* 223–234. doi:10.1037/1040-3590.15.2.223

Clark, S. L., Muthén, B., Kaprio, J., D’Onofrio, B. M., Viken, R., Rose, R. J., & Smalley, S. L. (2009). *Models and strategies for factor mixture analysis: Two examples concerning the structure underlying psychological disorders.* Manuscript submitted for publication. Retrieved from http://www.statmodel.com/papers.shtml

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In D. H. Saklofske (Ed.), *The SAGE handbook of personality theory and assessment. Vol. 2: Personality measurement and testing* (pp. 179–198). Thousand Oaks, CA: Sage.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354. doi:10.1037/h0047358

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

Dillman, D. A., Smyth, J. D., Christian, L. M., & Dillman, D. A. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Douglas, K. M., & McGarty, C. (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology, 40,* 399–416. doi:10.1348/014466601164894

Ehlers, C., Greene-Shortridge, T. M., Weekley, J. A., & Zajack, M. D. (2009). *The exploration of statistical methods in detecting random responding.* Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA.

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* (pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Gordon, M. E., Slade, L. A., & Schmitt, N. (1986). The science of the sophomore revisited: From conjecture to empiricism. *Academy of Management Review, 11,* 191–207.

Gough, H. G., & Bradley, P. (1996). C*alifornia Psychological Inventory: Administrator’s guide* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics. Vol. 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1,* 104–121. doi:10.1177/109442819800100106

Huang, J. L., Curran, P. G., Keeney, J. Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27,* 99–114.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39,* 103–129. doi:10.1016/j.jrp.2004.09.009

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. doi:10.1207/S15327752JPA7602_12

Lee, H. (2006). Privacy, publicity, and accountability of self-presentation in an on-line discussion group. *Sociological Inquiry, 76,* 1–22. doi:10.1111/j.1475-682X.2006.00142.x

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

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21–39. doi:10.1037/1082-989X.10.1.21

Montgomery, K. C. (2007). *Generation digital: Politics, commerce, and childhood in the age of the Internet.* Cambridge, MA: MIT Press.

Moore, R. S., & Ames, G. M. (2002). Survey confidentiality vs. anonymity: Young men's self-reported substance use. *Journal of Alcohol and Drug Education, 47,* 32–41.

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. doi:10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569.

Ong, A. D., & Weiss, D. J. (2000). The impact of anonymity of responses to sensitive questions. *Journal of Applied Social Psychology, 30,* 1691–1708. doi:10.1111/j.1559-1816.2000.tb02462.x

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609. doi:10.1037/0022-3514.46.3.598

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, D. E. Wiley, H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54,* 890–902. doi:10.1037/0022-3514.54.5.890

Rogers, R., Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10,* 160–177. doi:10.1177/1073191103010002007

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 34*(Suppl.), 100–114.

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. doi:10.1207/s15327752jpa6801_10

Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin, 72,* 214–228. doi:10.1037/h0027880

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105. doi:10.1037/0003-066X.54.2.93

Spelke, E., Hirst, W., & Neisser, U. (1976). Skills of divided attention. *Cognition, 4,* 215–230. doi:10.1016/0010-0277(76)90018-4

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91,* 25–39. doi:10.1037/0021-9010.91.1.25

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson/Allyn & Bacon.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory.* Chicago, IL: Scientific Software International.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, England: Cambridge University Press.

Wallis, C. (2006). genM: The multitasking generation. *TIME Magazine.* Retrieved from http://www.time.com/time/magazine/article/0,9171,1174696,00.html

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28,* 186–191. doi:10.1007/s10862-005-9004-7

Appendix

Pilot Analyses

For this study, we created two scales of participant study engagement. In order to refine these scales, we removed any respondents flagged by any of the bogus items. The resulting 196 respondents were considered "clean" and suitable for scale development.

Seventeen items were written to assess participant self-reported diligence and engagement with the study. An exploratory factor analysis with principal factors extraction and promax rotation resulted in two clear factors (first three eigenvalues = 4.18, 2.42, 1.05), with an un-interpretable third factor composed primarily of cross-loading items (see Table A1). Investigating the factor loadings, using .45 as a retention criteria, suggested one factor consisting of nine items that we labeled *Diligence* (α = .83) and a second factor that contained six items more attitudinal in nature labeled *Interest* (α = .81). The correlation between the latent factors was –.19. Two items failed to load onto either factor at a level of .45 or higher.

Table A1
Exploratory Factor Analysis of Participant Engagement Items

Item	Diligence	Interest
1. I carefully read every survey item.	**.71**	.01
2. I could've paid closer attention to the items than I did.	**.71**	−.16
3. I probably should have been more careful during this survey.	**.71**	−.08
4. I worked to the best of my abilities in this study.	**.71**	.12
5. I put forth my best effort in responding to this survey.	**.65**	.05
6. I didn't give this survey the time it deserved.	**.59**	−.04
7. I was dishonest on some items.	**.55**	−.05
8. I was actively involved in this study.	**.48**	.19
9. I rushed through this survey.	**.45**	.15
10. I enjoyed participating in this study.	−.02	**.82**
11. This study was a good use of my time.	−.16	**.71**
12. I was bored during the study.	−.04	**.65**
13. This survey was too long.	−.06	**.59**
14. The work I did for this study is important to me.	.09	**.57**
15. I care about my performance in this study.	.29	**.50**
16. I would be interested in reading about the results of this study.	.18	.38
17. I'm in a hurry right now.	.20	.29

Note. Items 16 and 17 were not retained. Bold indicates loadings > .4.

Received June 3, 2011
Revision received February 11, 2012
Accepted February 23, 2012 ■