ABSTRACT

YENTES, RICHARD DEAN. In Search of Best Practices for the Identification and Removal of Careless Responders. (Under the direction of Dr. Adam W. Meade).

Online surveys are frequently used in research and practice to inform a wide variety of decisions. Consequently, the quality of these decisions is contingent upon the quality of survey responses. Therefore, the presence of careless responders in online surveys is of concern to both research and practitioners alike. Though many metrics of careless responding have been proposed, they are generally continuous in nature. To date there is little to no research providing a sound basis for choosing a cut score for these metrics, in order to make a final determination to remove a respondent from the data set or not. This study was designed to address this issue by simulating many datasets with careless and careful respondents, and then evaluating a variety of methods to provide a cut score for metrics individually, and as an ensemble. Specifically, I investigated three metrics, namely, longstring, even-odd, and outliers. The results demonstrate the general stability and usefulness of the longstring metric when using a cut score set at .4 standard deviations from the mean. Subsequently computing and cleaning data using the outlier metric with a cut score of .5 standard deviations is also recommended, but at the cost of approximately 30% of careful respondents. Additional best practices for survey design and future research are also discussed.

In Search of Best Practices for the Identification and Removal of Careless Responders

by
Richard Dean Yentes

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina
2020

APPROVED BY:

_____	_____
Dr. Adam W. Meade Committee Chair	Dr. Lori L. Foster
_____	_____
Dr. Samuel B. Pond, III	Dr. Mark A. Wilson

DEDICATION

For my family, faculty, and friends, who gave me the time and space to produce something I'm proud of.

BIOGRAPHY

Richard Yentes completed his undergraduate studies at Virginia Polytechnic Institute and State University (Virginia Tech) in Blacksburg, Va. In 2010 he graduated with a Bachelor of Science in Psychology with High Honors. He moved to Raleigh, North Carolina in order to begin graduate studies in the Ph.D program at North Carolina State University, with a focus on industrial and organizational psychology. Along the way Richard had his colon removed, and replaced with a j-pouch, which his friends immediately dubbed the “semi-colon.” He worked full time as a data scientist for five years before finally getting around to finishing his dissertation. He currently resides in Durham, North Carolina with his cat Salia.

ACKNOWLEDGMENTS

The faculty at N.C. State are amazing, and in particular Lori and Adam, whose patient guidance helped me find my way. Cristina and the rest of the staff of the psychology department are the best, without their help I could never have navigated the paperwork to progress through the program. Also Shiva, who took a chance on me, and offered me a job where I honed my data science skills, without which this dissertation would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

In Search of Best Practices for the Identification and Removal of Careless Responders

Surveys have long been popular among psychologists and other researchers as a method for collecting data from human research participants. The appeal of surveys is easily understood, as they allow researchers to access diverse, and sometimes difficult to reach populations, with efficiency of cost, speed, and ease of administration (Schmidt, 1997). Recognizing their widespread use, some researchers long ago expressed concerns that survey measures, as a method for data collection, may represent a source of measurement error (Evans & Dinning, 1983; Haertzen & Hill, 1963; Thompson, 1975). As the internet grew to permeate the world, researchers quickly adapted to the new technology and began to administer surveys online. Today online surveys and questionnaires are used in selection assessments, to collect employee and consumer opinions, for academic research, and many other purposes. While use of the medium has allowed researchers to further capitalize upon the strengths of survey methods, it has not assuaged concerns about the quality of resulting data. If anything, online surveys have introduced new concerns, such as questions about the representativeness of online samples (Thompson, Surface, Martin, & Sanders, 2003), and the increased likelihood of inattention due to environmental distractions or multitasking (Johnson, 2005).

Recognizing the important roles that data from online surveys play in the modern world, researchers have been working to develop a literature to address threats to data quality that result when respondents do not respond carefully to a survey (Bowling et al., 2016; Huang, Curran, Keeney, Poposki, & DeShon, 2012; Johnson, 2005; Meade & Craig, 2012; Ward & Pond, 2015). Research has confirmed that this phenomenon, which I refer to as careless responding (CR), does occur (Baer, Ballenger, Berry, & Wetter, 1997; Berry et al., 1992; Huang et al., 2012; Maniaci & Rogge, 2014). Further research demonstrates that CR, when present, poses a real threat to the

validity of statistical inferences (Maniaci & Rogge, 2014; Huang et al., 2015). Accepting the occurrence and implications of CR, survey methodologists have begun developing methods to effectively counter the threat it poses. Such efforts can generally be divided into two general approaches. The first approach focuses on the antecedents of CR, and how to prevent it. In contrast, the second approach generally seeks to detect CR, so that researchers can remove data that result from it before using those data for analyses or decision making purposes. Much of the research on CR, to date, has opted for the detection approach, rather than the prevention approach.

Several different methods for detecting careless responding have been proposed, resulting in the development of numerous different indices (Huang et al., 2012; Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012). Consequently, it can be difficult to ascertain which indices of CR should be used, and this problem is further compounded by issues surrounding ease of computation, selection of appropriate cut scores, and the appropriate combination of multiple indices. Researchers hoping to use methods for detecting CR to clean their data would benefit from the creation of clear best practices that are relatively easy to implement. Thus, with the goal of deriving actionable best practices, the purpose of this paper is to evaluate the performance of several indices of CR when used in different ways, and under varying circumstances, to classify respondents as careless or not.

Careless Responding

Researchers have long been concerned with the quality of data obtained from surveys and questionnaires (Berry et al., 1992; Evans & Dinning, 1983; Haertzen & Hill, 1963; Thompson, 1975). In survey contexts, there are multiple distinct behaviors that threaten to introduce measurement error, such as malingering, and responding in a socially desirable manner;

However, several recent studies have led to a particular interest in situations where respondents complete an online survey with little or no regard for item content (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012). This phenomenon has been studied under numerous different names, including random responding, protocol invalidity, content-independent responding, content non-responsivity, insufficient effort responding, and careless responding (Huang et al., 2012). Though none of the proposed names are perfect, I prefer the term careless responding, as it is one of the earliest to appear (Haertzen & Hill, 1963), relatively easy to say, and intuitive to understand. Huang and colleagues (2012) formally define careless responding as:

> a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses. (p. 100)

However, it is important to note that careless respondents may not engage in careless responding for the entire duration of a survey. Instead, survey respondents may carefully respond to survey items in the beginning of a survey, and begin to respond carelessly later, as their attention wanders. This pattern is consistent with reports from respondents reported by early research on careless responding (Berry et. al, 1992). One reason for this relates to recent findings that careless responding is closely related to attention (Yentes, Foster, Meade & Pond, 2017). Cognitive Resource theories of attention suggest that maintaining attentional control on a task becomes progressively harder as its duration increases, and that this is likely to be particularly true for tasks that are easy or dull (Kanfer & Ackerman, 1989; Randall, Oswald, & Beir, 2014).

The Importance of Careless Responding

Though concerns about careless responding have existed for quite some time, initially they were somewhat speculative. Thus, some research on careless responding has sought to

estimate its prevalence, and to establish whether or not it is, indeed, a threat to data quality. Reports on the prevalence of careless responding vary from as little as 1% (Huang, Liu, & Bowling, 2015) to as much as 73% (Baer, Ballenger, Berry, & Wetter, 1997). This discrepancy is likely attributable to differences in the criteria for classification as a careless responder. When responding carelessly to a single item is sufficient, reports of careless responding tend to be much higher (Baer et al., 1997; Berry et al., 1992; Meade & Pappalardo, 2013). In contrast, when more refined criteria are applied, much lower estimates of careless responding, generally between 3-12% are reported (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012; Meade & Pappalardo, 2013). While some evidence does suggest that careless responding may be slightly more prevalent in undergraduate samples (Meade & Craig, 2012), research has shown that it occurs even among respondents who self-selected into a survey (Johnson, 2005), and in relatively high-stakes situations, as when a survey is taken as part of the job application process (Baer et al., 1997; Ehlers, Green-Shortridge, Weekley, & Zajack, 2009).

Given that careless responding does occur, researchers have also worked to demonstrate its effects on data quality. Much of the early research on careless responding was conducted with various clinical diagnostic inventories, such as the MMPI, or ARCI (Evans & Dinning, 1983; Haertzen & Hill, 1963; Nichols, Greene, & Schmlock, 1989; Thompson, 1975). These studies seem to focus on careless responding in an effort to ensure correct diagnosis, and though validity and reliability are mentioned, problems resulting from careless responding are not specifically articulated. The merit of these concerns seems intuitive; however, scientific skepticism does leave room to question whether careless responding is prevalent enough for these problems to manifest in ways that are practically meaningful. More recently, researchers have articulated

specific concerns, with accompanying studies that provide empirical evidence to inform a definition of problematic levels of careless responding.

One concern is that careless responding may hinder the validation of psychometric tests and measures by altering the factor structure of the constructs under investigation. For example, Woods (2006) designed a simulation to test the effects of careless respondents on the factor structure of a unidimensional scale with reverse worded items. When the sample was comprised of 5% careless respondents, indices of fit for confirmatory factor analysis of a unidimensional model were good according to commonly accepted standards (i.e. Hu & Bentler, 1999). At 10% careless responders, indices of fit for a unidimensional model were only marginal (e.g. CFI > .91; TLI = .95; RMSEA = .07). When a sample was comprised of more that 10% careless responders, fit of a unidimensional model was unacceptable by any modern standard. In a different study, Johnson (2005) split respondents into quartiles according to two separate inconsistency indices. Responses to the IPIP-NEO, provided by the top and bottom quartiles, were then independently subjected to principle components analysis and compared. Though factor loadings were lower for the bottom quartile, the five-factor structure of the IPIP-NEO was still observed, and items generally loaded onto the appropriate factors. Taken together, these two studies suggest that careless responding is likely only problematic when it exceeds a certain threshold.

Other studies have attempted to provide empirical evidence that careless responding leads to an increased chance of Type II errors. Maniaci and Rogge (2014) found that statistical power was reduced when careless responders comprised as little as 5% of a sample. Additionally, they demonstrated that robust research findings, which replicated among careful respondents, failed to replicate among careless respondents; however, their test was rather lenient, as they utilized

homogenous groups of careless and careful respondents, which are unlikely to occur in practice. That said, further research has provided more empirical evidence that careless responding can bias parameter estimates. Specifically, when careless responders comprise as little as 5% of a sample, it can bias parameter estimate either upward or downward, depending on the position of the population mean relative to that of careless respondents' (Huang, Liu, & Bowling, 2014). Said differently, depending on the circumstances, careless responding can increase the risk of Type I or Type II errors. Thus the preponderance of evidence suggests that careless responding does have important consequences, and that it is prevalent enough for those consequences to manifest in ways that are practically harmful. Therefore, given these findings, and particularly in light of recent concerns about replicability in psychological research (Open Science Collaboration, 2015), it is important to find ways to mitigate the threat that careless responding poses for psychological science.

Managing the Threat of Careless Responding

Along with efforts to establish the threat posed by careless responding, researchers have also sought to identify methods for mitigating its effects. Studies in this area generally adhere to one of two approaches. The first approach to dealing with careless responding is to identify its motivational antecedents with the goal of designing research and instruments in such a way that respondents do not engage in careless responding. The second approach for addressing careless responding is to identify survey respondents who were careless, and to remove their data before conducting further analyses. Thus, techniques for mitigating the impact of careless responding can be categorized depending on whether their goal is the prevention of careless responding, or its detection.

Though these two strategies for assuaging concerns related to careless responding are not mutually exclusive, some researchers have argued that there is reason to prefer prevention to detection. One reason for this preference is that removing data from careless responders may skew the sample distribution in a way that biases estimates (Berinsky, Margolis, & Sances, 2014; Ward & Pond, 2015). More concretely, there is growing evidence indicating that careless responding manifests, in part, as a function of respondent personality (Bowling et al., 2016). When personality is relevant to the phenomena under investigation, then the removal of data from careless responders is likely to remove the very sub-populations that are the focus of researchers' interest. When this problem is considered along with other concerns relating to wasted time, effort, and money, it is apparent that it would be ideal to prevent careless responding before it occurs.

To date, a number of methods for preventing careless responding have been proposed, many in the form of suggestions for survey design. For example, one option is to add a warning to survey instructions, informing participants that their responses will be checked for quality, and that credit and rewards will not be awarded to those that provide low-quality responses. Empirical evidence suggests that such warnings can be effective, though survey respondents generally react negatively to them (Huang et al., 2012). Additionally, some evidence indicates that this effect can be augmented through the presence of a virtual human (Ward & Pond, 2015). Similarly, Meade and Craig (2012) altered their instructions to require that respondents sign their name on each page of the survey. Manipulation of respondent anonymity did reduce CR, even though the instructions also assured participants that their responses would be confidential. Drawing upon theory from the literature on motivation, Yentes, Foster, Meade, and Pond (2017) argued that long surveys place a larger burden on respondents, leading to careless responding. In

their study, careless responding was more prevalent among participants randomly assigned to take a long version of a survey, as compared to those assigned to take a shorter version of the same survey. Several additional methods have also been investigated, though researchers have generally been unable to demonstrate their effectiveness. For example, in two studies, the presence of a proctor, virtual or otherwise yielded no or small main effects for reducing CR (Francavilla, 2016; Ward & Pond, 2015). Similarly, neither the inclusion of a survey progress bar, nor the disclosure of an estimate of the time commitment required to complete a survey, resulted in significantly less careless responding (Yentes, Foster, Meade & Pond, 2017).

Despite the identification of several methods that have been demonstrated to reduce careless responding, there are several problems with relying solely upon prevention as a strategy for addressing it. First, even those methods for which statistically significant effects were observed, effect sizes are typically modest. Therefore, while these methods may prevent some occurrences of careless responding, they may not prevent enough to effectively mitigate the threat it represents for statistical inference. Second, even effective methods are not without drawbacks, and must be applied situationally. For example, it may not always be possible to design a short survey that is adequate for its intended purpose. Finally, of the three effective methods described previously, two of them place the burden of dealing with the problems posed by careless responding on the survey respondent, as evidenced by negative reactions to warnings embedded in survey instructions (Huang et. al, 2012), and discomfort reported by respondents when perceived anonymity is low, despite assurances of confidentiality (Saari & Scherbaum, 2011; Thompson & Surface, 2007). As a solution, this is undesirable both ethically, and because respondents' satisfaction with survey experiences is predictive of their willingness to complete future surveys (Thompson, Surface, Martin, & Sanders, 2003). Thus, for all of the reasons

discussed here, researchers cannot address the threat posed by careless responding by relying solely on techniques to prevent it from occurring.

Methods for Detecting Careless Responding

The performance of any given technique designed to prevent careless responding is typically evaluated with respect to how much reduction it causes in some measure(s) of careless responding. Consequently, as researchers have sought to establish techniques for preventing careless responding, they have also developed methods to more accurately distinguish diligent survey respondents from those engaging in careless responding. As mentioned previously, many researchers consider the detection of careless responders, and removal of their data from a sample prior to conducting analyses, to be a worthwhile strategy for countering the threats to scientific inference that careless responders introduce. As a result, myriad different indices of careless responding have been proposed and studied. Though an exhaustive review of these indices is beyond the scope of this paper, a brief overview of the most common types of CR indices follows. Readers requiring more detail may refer to Curran (2016) for a more exhaustive overview.

Infrequency. Indices of careless responding that use the infrequency approach rely upon the assumption that it is possible to write items to which careful respondents will almost always provide a similar response. For example, if a survey contained the item “I used to live on the planet Mars.”, it would be highly improbable for any careful respondent to report agreement with this item. Similarly, instructed response items are a special case of infrequency items, in which researchers explicitly tell survey participants to select a specific response option (Meade & Craig, 2012). Researchers generally include a number of such items distributed throughout a survey, and an index score is computed as the sum of total endorsement across all such items, or

as the total number of items to which a survey participant selected the incorrect response option. The infrequency subscale of the Attentive Responding Scale (ARS) is an example of a validated infrequency index of careless responding (Maniaci & Rogge, 2014).

Inconsistency. Inconsistency measures of careless responding rely upon the assumption that careful respondents should generally respond to similar questions in a similar manner. For example, one would not generally expect a respondent to strongly endorse the statement "I'm the life of the party", right after strongly endorsing a statement like "I hate being the center of attention". It is conceivable that such circumstance may arise as a result of measurement error, or some idiosyncrasy of the respondent; however, given the literature on cognitive dissonance and self-perception, it would be very unlikely that a careful respondent would exhibit a general pattern of inconsistent responses to a well-designed survey. Consequently, inconsistency measures of careless responding quantify a pattern of inconsistency with respect to similar or related items in a survey.

Though inconsistency measures share a common approach to detecting careless responding, there is considerable variation in the manner in which items are chosen and compared. For some measures, item pairs are specifically designed for this purpose, as is the case with the inconsistency subscale of the Attentive Responding Scale (Maniaci & Rogge, 2014). One disadvantage of such inconsistency measures is that they add to the burden of the survey respondent by adding items to the survey. For example, the ARS inconsistency subscale necessitates adding 12 items (for the ARS-18; 22 for the ARS-33) that are unrelated to research questions motivating the survey.

Some other measures, like the psychometric synonyms index, rely upon empirically derived item pairs (Johnson, 2005; Meade & Craig, 2012). These indices identify highly-

correlated items in the overall sample, and then compute the within-person correlation for the identified item-pairs. One drawback to this technique is that it is dependent upon sufficient correlated item pairs to manifest in the sample. No formal best practices have been established for minimum correlations for such item pairs, though some researchers have used and/or recommended .60 as a minimum requirement (Curran, 2016; Meade & Craig, 2012). In samples where no such pairs exist, it would not be advisable to compute or interpret the psychometric synonyms index.

The even-odd index is another variant of an inconsistency index, which is analogous to the application of split-half reliability to respondent consistency. For this technique, researchers split validated unidimensional scales that are included in their survey into two sub-scales, one for the even numbered items, and the other for the odd numbered items. Sub-scale scores are calculated as the average of items on each sub-scale. Then the even-odd index score is computed as the within-person correlation between the paired sub-scale scores, and then corrected for decreased length of the measure using the Spearman Brown prophecy formula (Meade & Craig, 2012).

Invariability. Longstring indices of careless responding were created based on the idea that careless respondents may select the same response for longer strings of contiguous items than careful respondents. They are generally computed by first examining survey participants' individual item responses to determine whether they are the same as their response to the preceding item. Researchers then create a longstring score for each respondent, typically either as the longest string of consecutive identical responses (e.g. Meade & Craig, 2012), or as the average of the lengths of the longest string on each page (Huang et al., 2012).

Outlier Analysis. The use of outlier analysis as an index of careless responding stems from the notion that disregard for item content is characteristic of careless responding. Consequently, for each item, a careless responder should have a consistently low probability of selecting a response that is close to the item mean. Conversely, unless a careful respondent were uniformly atypical, their responses should be more likely to trend toward item means. Mahalanobis' distance is a multivariate outlier statistic that, conceptually, is a measure of the distance between a survey participant's response vector, and the vector of item means in the overall sample. Thus, researchers have applied Mahalanobis' distance to the detection of careless responders (Ehlers et al., 2009; Meade & Craig, 2012).

Types of Careless responding

An open question pertaining to the nature of careless responding is whether there are different types of careless responding. One way to approach this question is to apply statistical classification methods to indices of careless responding. At least two recent studies applying such methods generally agree that a three-class solution describes the data very well (Maniaci & Rogge, 2014; Meade & Craig, 2012). These classes represent groups of respondents who are roughly typified by one of three patterns. The first group of respondents provide varied, yet consistent responses, as indicated by better scores across indices of careless responding. This response pattern is generally consistent with the expected behavior of a respondent diligently answering survey items. Consequently, respondents who demonstrate this pattern are referred to here as diligent respondents. A second pattern is characteristic of respondents who provide the same response to relatively large numbers of contiguous questions, and are distinguished by elevated scores on longstring indices. Accordingly, researchers may think of respondents with this pattern of behavior as longstring careless respondents. The third pattern is characterized by

relatively varied responses, with poorer scores on indices of careless responding based on consistency with self, or others (e.g., inconsistency and outlier indices). Consequently, respondents exhibiting this pattern are referred to as generally careless. Both longstring and generally careless respondents also tended to provide incorrect responses to infrequency items.

Different patterns of careless responding may be especially important in light of the nature of how indices of careless responding are computed and used for detection. The maximum long string index is determined by the pattern of responses provided by a survey participant, and is unaffected by data provided by other participants. In contrast, consistency and outlier indices of careless responding are computed with respect to sample means and correlations. Thus, unless identified and removed prior to computing consistency and outlier indices, the data provided by brazen careless respondents will distort the scores that are generated. Consequently, it is reasonable to suggest that the removal of longstring careless respondents, prior to computing consistency and outlier indices, should only improve their ability to accurately distinguish between diligent respondents and generally careless respondents. Thus, the following hypotheses are proposed.

Hypothesis 1: Cleaning data using the maximum long string index before computing the outlier index of careless responding will improve the outlier index's informedness.

Hypothesis 2: Cleaning data using the maximum long string index before computing the even/odd index of careless responding will improve the even/odd index's informedness.

Making Decisions About Careless Responders

In practical terms, the purpose of any index of careless responding is generally to inform decisions about whether or not a survey respondent should be considered careless. Such decisions may be used to inform a variety of goals, such as academic research on the nature of

careless responding, or for more applied purposes, like cleaning a data set. Regardless of the purpose for making such decisions, there is a need for evidence-based recommendations defining a process for bridging the gap between computing numerous indices of careless responding and reaching a final careless or not careless decision. The creation of such a process implies a stance on two key issues; namely, how cut scores can be used to inform these decisions, and whether indices of careless responding can be used together to arrive at better decisions.

The simplest process for making a decision about careless responding is to simply compute a single index of careless responding, choose a cut score, and then classify each respondent as careless or not accordingly. This method is unlikely to perform optimally because, as mentioned previously, several different types of careless responding are thought to exist. Further, categories of indices for detecting careless responding have different strengths and weakness with respect to what type of careless responding they can identify (Curran, 2016). Consequently, it is both intuitive and reasonable to conclude that using some combination of different indices of careless responding should be preferable to the simple use of one index.

To arrive at an actionable process, it is necessary to be more specific about how different indices should be combined to inform a final decision. Some empirical evidence suggests that a simple combination of indices may have detrimental effects (Maniaci & Rogge, 2014). Said differently, classifying a respondent as careless if they trigger a cut score on any one index of careless responding is unlikely to be optimal. Additionally, research on selection suggests that mechanical combination systems typically outperform human judgement (Wunder, Thomas, & Luo, 2010). And, though some rules of thumb exist (e.g. Curran, 2016), there is considerable uncertainty with respect to how arrive at a cut score that can be generalized from one survey to another, and computed without knowledge of who is careless and who is not. However, as

mentioned previously, in some cases it may make sense to cut on one index before computing other indices of careless responding, thus some researchers have recommended using a multiple-hurdle approach to identifying careless respondents (Curran, 2016). Such a practice has the benefit of trading off the loss of information resulting from the dichotomization of continuous indicators, prior to a final decision, with the possibility that the performance of some indices of careless responding may be distorted by presence of some relatively careless respondents that are fairly easy to detect. Regardless, it is clear that the matter of combining the information provided by different indices of careless responding is not as simple as it may initially seem, and more empirical information about what constitutes a good cut score for any given index of careless responding is needed. One technique that some researchers employ (e.g. Yentes, et al., 2017), is to use respondents' standard deviation from the sample mean on an index as the basis for a cut score. In order to determine the effectiveness of this practice, the following research question is proposed.

Research Question 1: How well do cut scores based on deviations from the norm perform?

Several researchers have suggested that the use of different indices of careless responding in combination should perform better than relying on any one single index (Curran, 2016; Huang et al., 2012; Meade & Craig, 2012). Despite this, researchers have found some methods of combining indices of careless responding to perform poorly (Maniaci & Rogge, 2014). However, there is a vast number of possible ways to combine different indices of careless responding that are, as yet, untested. For example, the generalization of hypotheses 1 and 2, that computing and cleaning based on the long string index prior to computing other indices of careless responding will improve accuracy, implies that a flag on the long string index should be sufficient for

exclusion from further analyses. However, an alternative method could require that at least two indices, computed on uncleaned data, must flag a respondent as careless, in order to exclude them for further analyses. While it would be impractical to examine all potential rules in any one study, given that few have actually been examined, it is too early to dismiss the notion that using CR indices in combination could improve upon the accuracy of any given individual index. Thus, the following research question is proposed.

Research Question 2: Can methods for combining the information provided by indices of careless responding improve upon the accuracy of the individual indices?

Latent Profile Analysis (LPA) is a statistical technique that attempts to identify homogenous groups in empirical data by examining the means and correlations among variables (Bauer & Curran, 2004). In contrast to the Common Factor Model that underlies factor analysis, LPA assumes that correlations between variables are indicative of the presence of discrete groups, and not a continuous latent variable. In LPA, distinct groups are characterized by their mean vectors with respect to the input variables. As mentioned previously, researchers applying latent profile analysis to indices of careless responding have generally found that a three-class solution fits the data well (Maniaci & Rogge, 2014; Meade & Craig, 2012). Given that these classes are generally interpreted as corresponding to careful (i.e., not careless), longstring careless, and generally careless responding, it would be reasonable to suggest that LPA could be capable of classifying careless respondents. Thus, the following research question is proposed.

Research Question 3: Does LPA function well as an index of careless responding?

Methods

Data Model

Datasets were simulated using a method similar to that of Meade and Craig (2012). A model of factor correlations and item parameters was obtained from a public dataset of responses to the IPIP HEXACO measure (Ashton, Lee, & Goldberg, 2007), which was downloaded from 'http://personality-testing.info/_rawdata/'. The dataset consists of 22,786 participants responding to 240 items spread across six factors, each with four facets. Four additional items were also present, indicating self-reported understanding of the instructions, accuracy, the country of the respondent, and the number of seconds between the survey being opened and its submission. Data from respondents who reported that they did not understand the instructions (N=626), or that they had not answered accurately (N=561), were dropped. Additionally, data from respondents who completed the measure in less than 17.5 minutes (N=960), or in more than 2 hours (N=11,070) were also dropped to help ensure that the dataset consisted of meaningful responses. All of these criteria were applied simultaneously, yielding a final dataset containing 10,216 observations.

A subset of 10 facets were chosen from across the factors to serve as a 100-item measure. Table 1 contains a list of the chosen facets, as well as the number of positively and negatively coded items that each facet contains. Correlations between each of the selected facets were computed. IRT item parameters for the graded response model were estimated for each facet using the MIRT package in R (Chalmers, 2012). Both the correlation matrix and estimated item parameters were used in subsequent data simulation.

Data Simulation

A total of 5,000 simulated datasets were created, each with 500 rows (i.e. simulated respondents). The R-script used to simulate these datasets can be found in Appendix A. The matrix of correlations obtained from the HEXACO dataset served as input into a Choleski decomposition, which was computed to obtain weights for use in creating facet-level scores for each simulated respondent. Then, the previously estimated item parameters were used to simulate item-level responses. The resulting datasets contained simulated responses ranging from 1 to 7, with each row representing a simulated careful survey respondent, and with a population correlation among factors identical to that of the observed data. Next, rows designed to simulate careless responders were generated and introduced into the dataset.

Careless responses were simulated according one of two models, namely, either longstring, or generally careless. The percentage of respondents following each model varied between datasets. The percentage of longstring careless respondents, in any one replication sample, was determined by sampling a normal distribution with a mean of .02 and a standard deviation of .01. Similarly, the proportion of generally careless respondents was sampled from a normal distribution with a mean of .06 and a standard deviation of .02. These values were chosen so that the total proportion of careless respondents for each sample would center around the midpoint of the 3-12% range observed in prior research, with substantial variation. The parameters for longstring careless responders were set lower in order to maintain consistency with empirical evidence that longstring careless responding is less prevalent than generally careless responding (Maniaci & Rogge, 2014; Meade & Craig, 2012). The dataset was then randomly sampled without replacement until a number of rows proportional to the total percentage of careless respondents were selected. These rows were designated careless

respondents, and this subset was then randomly sampled without replacement in order to proportionally assign careless respondents to either the longstring or the generally careless model.

Careless response models. For careless respondents that were simulated from a longstring model, one response option was randomly chosen, which was used for each careless response. For those that followed the generally careless model, careless responses were randomly selected from a normal distribution with a mean of 3.5 and a standard deviation of 1.25. A resilience score was generated for each careless respondent, from a normal distribution with a mean of 50, and a standard deviation of 10. This resilience score represented the point at which a survey participant would begin resorting to careless responding; therefore, if a simulated careless respondent had a resilience score of 50, then only responses to items 50-100 were replaced with careless responses. The resulting datasets were then complete, and ready for substantive analyses. More specific details of the implementation of these data generation procedures can be found in Appendix A.

Measures

Indices of careless responding. Three different indices of careless responding were examined, each representing a different category of careless responding indices. Namely, maximum longstring was chosen as a measure of invariability; the even-odd index, as described by Meade and Craig (2012), was used as a measure of individual consistency; and finally, Mahalanobis' distance served as a multivariate outlier measure. While there are numerous other indices of careless responding, ones that measure the same type of careless responding tend to be correlated, so I selected three of the higher performing indices based on previous research (e.g. Mead & Craig, 2012). These indices were computed using the careless R package (Yentes &

Wilhelm, 2018). Following recommendations from Curran (2016), maximum longstring was computed first, on the raw response vectors. Then reverse worded items were reverse coded, and both even-odd consistency and Mahalanobis' distance were computed on these reverse coded response vectors.

Classifier Performance. As mentioned previously, the focus of this paper is on developing an evidence-based process for arriving at a final determination of whether a survey respondent should be considered careless. The general form of this problem is known as binary classification. While there are a variety of named metrics for evaluating the performance of binary classifiers (e.g., sensitivity, specificity, accuracy, etc.), classifier performance was operationalized using Bookmaker Informedness which, conceptually, is a quantification of how informed a predictor is with respect to a given condition (Powers, 2011). Informedness is computed as:

$$B = 1 - fnr - fpr$$

Where *fnr* is the false negative rate, which is equal to the proportion of false negatives resulting from prediction, divided by the proportion of real negatives in the sample, and *fpr* is the false positive rate, which is equal to the proportion false positives resulting from prediction, divided by the proportion of real positives in the sample. Informedness was preferred over other metrics because it considers false positives and false negatives to be equally problematic, and confidence intervals for it are easily computed using a known formula for standard error (Powers, 2011).

$$\mathrm{SE} = \sqrt{[SSE_B \,/\, (N - 1)]}$$

In order to compute a standard error for informedness, it is first necessary to compute the sum of squared errors. In this context, the sum of squared errors is computed as the squared deviation from the optimal value of one.

$$SSE_B = (1 - B)^2$$

An additional reason for preferring Informedness to other measures is because of its relationship with ROC analyses, which are also sometimes used for model evaluation. A ROC curve is a plot of the sensitivity of a model on the y axis, against 1-specificity on the x axis. The diagonal of this plot generally represents chance performance, and a model performance is interpreted in terms of the total area between the curve and the line representing chance. Importantly, the line is plotted across all possible thresholds (cut scores) for the predictor, thus informedness can be computed for every point in the line, and by maximizing informedness we arrive at a basis for deciding upon a cut score.

Results

Research Question 1

Research question 1 was proposed to investigate the suitability of standard deviations from the mean as a cut-scoring criterion. For each of the 5,000 samples, the longstring, even-odd, and outlier indices of careless responding were computed for each simulated participant, and left in their raw form. For each sample, the informedness of each index was computed across the range of -3 standard deviations from the mean to 3 standard deviations from the mean. The maximum value of informedness (max_B for each index in each sample was recorded), which constitutes the recommended cut score for that sample. Across all 5,000 samples, for each index, the average value of max_B was recorded along with the proportion of samples in which that index performed better than chance.

Longstring. Figure 1 shows the distribution of the index of maximum informedness (i.e., max_B) for the longstring metric, across each of the 5,000 samples. As shown, max_B is fairly tightly distributed around the mean of .39 indicating relative consistency in suitable cut scores across the replication samples. Table 3 contains measures of central tendency for the standard deviation index of max_B across all 5,000 samples, for each of the three indices of careless responding. Figure 2 is a series of graphs plotting the informedness of the longstring metric when using standard deviations from the mean across the range [-3,3] as the cut scoring criterion. To accomplish this, four samples were randomly selected without replacement from the 5,000 simulated samples, and each graph depicts the informedness of the longstring index from one of those samples. Because careless respondents typically have higher longstring scores, very few are missed by a low threshold, consequently, the false negative rate is roughly equivalent to zero. At the same time, a very low threshold will flag many careful respondents, causing the false positive rate to approximate one. Consequently, at low thresholds informedness is close to zero, and rises from there, approximating a sigmoid function in each of the four graphs. Figure 3 is a plot of the estimated cumulative distribution function of max_B for the longstring metric. The curve represents the proportion of samples where longstring's informedness was maximized at a value less than or equal to the corresponding index on the x axis. Given the general sigmoid shape observed in figure 2, as long as informedness is maximized to the left of a proposed threshold in any given sample, one can reasonably assume good informedness. I proposed that the average index of maximum informedness ($\bar{x} = .4$) for the longstring metric would be used to evaluate this research question. Referencing that value, approximately 60% of samples have maximize informedness at or below .4.

An a-priori criterion was set such that, if the informedness of the longstring metric at the test value was better than chance in at least 75% of samples, then cut scoring on the basis of that value could be considered reasonable. The longstring metric far exceeded this criterion, being informant in 99.2% of samples ($\bar{B}$ = .26, SD_B = .12). Taken together with the information supplied in the previous paragraph, this suggests that the longstring metric may be relatively stable and performant when standard deviations are used as a method for cut scoring.

Outlier. Figure 4 shows the distribution of max_B for the outlier metric, across each of the 5,000 samples. This distribution is approximately normal around the mean of .5, indicating that max_B is somewhat less stable for the outlier metric than it was for the longstring metric. Figure 5 is a series of graphs plotting the informedness of the outlier metric using the same procedures employed to produce Figure 2 for the longstring metric. Each of the four graphs in figure 5 depict a rough approximation of an asymmetric bell shape, indicating that cut scores deviating from max_B in either direction can generally be expected to result in lower informedness. Consequently, given that max_B is generally unknown in real world situations, maximizing the probability that a proposed cut score is as close to max_B as possible is vital to proper application of the outlier metric.

I proposed that the average index of maximum informedness ($\bar{x} = .5$) for the outlier metric would be used to evaluate this research question. An a-priori criterion was set such that, if the informedness of the outlier metric at the test value was better than chance in at least 75% of samples, then cut scoring on the basis of that value could be considered reasonable. The outlier metric far exceeded this criterion, being informant in 97.3% of samples ($\bar{B}$ = .17, SD_B = .10). As shown in Table 3, the mean, median, and mode of the index at which max_B occurs across samples are largely consistent. Finally, though there is some erratic behavior to the informedness of the

outlier metric over the range of [-3,3] between samples, the metric is generally informant over the range of [0,1]. Consequently, the outlier metric may be relatively stable and performant when standard deviations are used as a method for choosing a cut score.

Even-odd. Figure 6 shows the distribution of the index of maximum informedness (i.e., max_B) for the even-odd metric, across each of the 5,000 samples. As shown, max_B is not particularly tightly distributed for the even-odd metric. The graph is right skewed with a pronounced tail, indicating a relative lack of stability for max_B compared to the outlier and longstring metrics. Figure 7 is a series of graphs plotting the informedness of the even-odd metric generated according to the same procedures used to produce Figures 2 and 5.The four graphs of informedness in Figure 7 do not appear to vary around a consistent shape between samples, and the behavior across the range for each sample is best characterized as erratic. As a result, these findings do not elicit any clear inferences to guide the selection of a cut score for the even-odd metric.

I proposed that the average index of maximum informedness ($\bar{x} = .2$) for the even-odd metric would be used to evaluate this research question. An a-priori criterion was set such that, if the informedness of the even-odd metric at the test value was better than chance in at least 75% of samples, then cut scoring on the basis of that value could be considered reasonable. The even-odd metric met this criterion, being informant in 84.1% of samples ($\bar{B}$ = .08, SD_B = .08). That said, the mean, median, and mode of the index of max_B shown in Table 3 are all markedly different. Together with the erratic behavior of informedness shown in Figure 7, there is reason to proceed with caution. Consequently, it may be reasonable to use a standard deviation-based cut score for the even-odd metric; however, I do not have great confidence in this finding.

Hypothesis 1

The purpose of Hypothesis 1 was to determine whether it was better to flag and remove respondents using the longstring metric before computing the outlier metric, or not. In order to test this hypothesis, the informedness of the outlier metric was computed after it was used to clean each simulated dataset, and then again on each simulated dataset that was first cleaned using the longstring metric. In all cases, the cut scores derived in research question 1 were used as the basis for cleaning the datasets.

Using the formula for Standard Error mentioned previously, a 95% confidence interval was calculated for the informedness that was computed on each simulated sample. In order to assess whether a given sample demonstrated significant improvement in informedness of the outlier metric as a result of computing it after cleaning with the longstring metric, the 95% confidence intervals of the two informedness values were compared. When the confidence levels did not overlap, and informedness was higher when computing the outlier metric after the longstring metric, then that method was considered significantly better for that sample than computing the two metrics simultaneously. An a priori criterion was set such that if the proposed method was significantly better in 75% of samples, then that would be a data point in favor of Hypothesis 1. The proposed method was only significantly better in 1,741 (34.8%) samples.

Additionally, an omnibus test of improvement was computed on the mean of informedness across samples. A one-tailed Wilcox rank sum test (W=19,501,286, $p < .001$) with continuity correction was conducted on the means of the experimental ($\bar{B} = .29$) and control conditions ($\bar{B} = .18$). There was a significant improvement in informedness from the omnibus test, with the proposed method being roughly .11 more informant than simultaneous computation. Given that the proposed method failed to meet the a-priori criterion of being

significantly more informant in 75% of samples, but did pass the omnibus test of improved informedness, Hypothesis 1 was partially supported.

Hypothesis 2

The purpose of Hypothesis 2 was to determine whether it was better to flag and remove respondents using the longstring metric before computing the even-odd metric, or not. It was tested using the same procedure used to test Hypothesis 1. An a-priori criterion was set such that if the proposed method was significantly better in 75% of samples, then that would be a data point in favor of Hypothesis 2. The proposed method was only significantly better in 294 (6%) samples.

Additionally, an omnibus test of improvement was computed on the mean of informedness across samples. A one tailed Wilcox rank sum test (W=16,772,833, p <. 001) with continuity correction was conducted on the means of the experimental ($\bar{B} = .14$) and control conditions ($\bar{B} = .08$). There was significant improvement in informedness from the omnibus test, with the proposed method being roughly .06 more informant than simultaneous computation. Given that the proposed method failed to meet the a-priori criterion of being significantly more informant in 75% of samples, and the statistically significant omnibus test of improved informedness, Hypothesis 2 was partially supported.

Research Question 2

Research Question 2 was proposed to determine whether a method for combining the information provided by different indices of careless responding could improve upon the informedness of each metric by itself. Each metric was computed and scored using the appropriate cut-score from Research Question 1. These votes were then used in a scoring process that evaluated each of the different methods described in Table 2. Overall descriptive statistics

for the performance of each method over the 5,000 methods can be found in Table 4. Figure 8 is a frequency plot of the method that had the highest informedness for each sample.

According to the a priori conditions set in my proposal, in order for a combination method to be considered an improvement upon the singular indices, it needed to have a greater average informedness across the 5,000 samples. As shown in Table 4, the "Longstring or Outlier" method, with longstring computed first, had a markedly higher informedness than any other method. Likewise, its sensitivity and specificity were both approximately .7, and the average proportion of careless responders remaining after cleaning with this method was .039, which is below the .05 threshold that has been demonstrated to be problematic (Hu & Bentler, 1999; Huang, Liu, & Bowling, 2014; Maniaci & Rogge, 2014). Additionally, it was the most informant method in over 80% of samples. Consequently, it is clear that indices of careless responding can be combined in ways that augment the performance of any one metric. Thus, Research Question 2 was answered in the affirmative.

Research Question 3

Research Question 3 was proposed to determine whether latent profile analysis (LPA) could be usefully applied as a method for detecting careless respondents. In order to address this research question, respondents raw scores on each of the three indices of careless responding were input into an LPA using the tidyLPA R package (Rosenberg, Beymer, & Schmidt, 2018). Given the nature of careless responding, I assumed that the variances and covariances of careless responders would not necessarily be similar to those of a careful population. Consequently, I specified parameters that allowed them to vary independently between profiles. Out of 5,000 samples, the function was not able to estimate a 3-profile solution 28 times.

Each simulated respondent was assigned to a class based on the profile with which it had the highest conditional probability of membership (Scrucca, Fop, Murphy, & Raftery, 2016). Then a series of rules was applied on the means of each class to programmatically determine the careful group. The primary rule for identifying the careful class was to find the class to which the highest number of simulated respondents were assigned. A second rule added to the strength of that decision if that class's mean vector was closest to the vector (-.2, .4, 0), as an examination of a few datasets indicated that the careful class generally had a slightly elevated even-odd index. If both rules indicated the same class then the match was considered strong, otherwise, it was considered weak. In 99.2% of the 4,972 simulated samples for which the LPA ran, the match was considered strong by these criteria. The mean informedness of the weak matches (M =.129) was markedly lower than that of the strong matches (M = .294), but did not meaningfully change the overall mean of the method (M = .295)

Full metrics for performance of LPA as a classifier are included with the other methods in Table 4. When used as a classification method, LPA was among the methods with the highest informedness ($\bar{B} = .29$). Consequently, it seems that LPA can be applied to the detection of careless respondents in most samples with some simple heuristics for assignment to a prediction of 'careful'. Consequently, Research Question 3 was answered in the affirmative.

Discussion

The purpose of this paper was to gain deeper insight into the function of different metrics and methods for detecting careless responding. I began by examining three metrics of careless responding, and in particular, whether standard deviations from the mean of each metric could provide a reasonable basis for choosing a cut score for those metrics. In particular, the longstring metric is very well behaved when used in this manner. Given that the informedness of the metric

takes on the shape of a sigmoid function over the range of [-3,3] standard deviations, one can be fairly confident that the metric is always worth applying and will generally cut only those respondents who are choosing the same response repeatedly. As the cut score is raised toward three standard deviations, one can generally assume that few false positive classifications are being made, at the cost of a minimal number of true positives. Importantly, this is likely to hold true only when appropriate choices have been made in the survey design phase, such as interspersing negatively worded items throughout the survey. From a design perspective, this seems strictly better than adding to respondent burden by including instructed response items, which serve essentially the same purpose.

It is also worth noting that in this study I chose to use the mean of the distribution of maximum informedness to select a cut score for each of the metrics. Once again, given the sigmoid shape of the informedness of the longstring metric over the range of standard deviation, so long as informedness is maximized to the left of the chosen cut score, one can be confident that the metric will be quite performant. Consequently, there is also an argument to be made from looking at the estimated cumulative distribution function (ECDF), that a standard deviation cut-score value set at one would ensure that informedness is maximized to the left of the cut-score in almost all samples. Though that would require eschewing the theoretical ideal cut-score in most samples, the loss of informedness for cutting to the right of the max is minimal, and across samples, the science of psychology is likely to benefit from such a conservative application of the longstring metric.

In contrast, the even-odd metric, and to a lesser extent the outlier metric, were more erratic over the range of standard deviations. Despite this, the outlier metric is generally informant when used with a cut score of .5 standard deviations from the mean. However, the

shape of the distribution of informedness over the range of standard deviations in Figure 5 generally forms a peak around the maximum value of informedness. Consequently, the logic I applied to the sigmoid shape of the longstring metric does not generalize to the outlier metric.

I hypothesized that computing the even-odd, and outlier metrics after cleaning a dataset using the longstring technique would improve their informedness. The results for both metrics were metric were mixed, significant at the omnibus level, but failing to achieve statistical significance for most individual samples. Despite mixed support application of this technique generally did result in improvements of informedness. Consequently, it seems advisable for researchers to clean their datasets using the longstring metric, before computing and applying the outlier and even-odd metrics.

Finally, I compared the performance of a variety of different methods for combining metrics of careless responding into a binary classifier. As expected from previous research (Maniaci & Rogge, 2014), using all the metrics of careless responding in such a way that triggering the threshold of any of them is sufficient for removal is a poor method for using them in combination. In contrast, however, we did find several methods for combining the metrics that warrant further discussion. The first method is not really a combination, but just an application of the longstring metric as described in Research Question 1. This method is extremely specific to identifying careless respondents demonstrating a longstring pattern of careless respondents, and almost no one else. That said, on average, while it did substantially reduce the number of careless respondents in a sample, it did not reduce them below the .05 threshold that previous research has shown to be problematic. Consequently, cleaning using the longstring metric alone should be viewed as a highly conservative approach, to be applied when sample sizes are very small, or when there is reason to suspect that careless responding is correlated with the outcome

of interest. Nevertheless, it seems warranted to always compute this metric and clean data according to its result, though it seems reasonable to suspect that including negatively worded items in the design of the survey is likely helpful to maximizing the benefit of this metric.

Computing and cutting using the longstring metric, and then the outlier method in sequence was, on average, the most informant method for detecting careless respondents. It was one of only two methods to reduce the average proportion of careless responders remaining in a sample to less than .4, and alone of all the methods, never had degenerate (i.e., negative) informedness. As such, this method is the clear winner in terms of the best compromise of retaining careful respondents and removing careless ones. That said, there is a cost to employing this method; concretely, about 30% of careful respondents may be cut along with careless respondents when using this technique, which may mean it is infeasible when sample sizes are small.

A third method I refer to as “longstring or agree”, cleaned data using the longstring method before computing and applying the outlier and even-odd metrics. With this method, respondents are considered careless if they are flagged by the longstring metric, or if they are flagged by both the outlier and even-odd metrics. This method was among the most informant methods of those I examined, and it came very close to reducing the percentage of careless respondents to below the 5% threshold identified by previous research, and at the cost of only about 11% of careful respondents. However, in this paper the even-odd metric was computed based on 10 facets, each with 10 items used to measure it. In practice, such lengthy scales may rarely be used, and it’s possible that the metric may grow more unstable with shorter, less established scales. Consequently, though this method may be appealing to the pragmatic researcher, caution is warranted in its application.

A final method that warrants discussion is the application of latent profile analysis (LPA) on the metrics of careless responding in order to clean data. For samples in which the model converged, this method actually performed reasonably well, but was substantially less informant than other methods. Further, on average the proportion of careless respondents in the remaining sample was .05. Consequently, while application of LPA does show some promise for identifying careless respondents for samples in which a model can be estimated, in almost every sample it was generally outperformed by other methods.

Limitations and Future Research

As with all research, the scope of this investigation was necessarily limited. There is a large constellation of effects and sample manipulations that remain uninvestigated, and as such the results of the current paper represent the best of our current thinking related to detecting careless respondents. One issue is that careful respondents cut using methods for detecting careless respondents could be different in some way that is meaningful to the purpose of the study (Ward & Pond, 2015). Consequently, one avenue of future research could seek to examine how manipulations of sample characteristics and research designs might interact with the metrics of careless responding.

Another issue is that 30% of careful respondents is a non-negligible portion of people to cut from a sample and collecting large amounts of data is often difficult or expensive. As a matter of research design, we eschewed the use of machine learning techniques (e.g. logistic regression), because they would almost certainly result in a model that over-fit the data. That said, while the current paper did investigate a number of different methods for combining metrics of careless responding, there are still numerous other metrics and methods that could be conceived of. For example, examining polynomial transformations of the current metrics, adding

new metrics like IRV, or survey completion time, possibly in combination with latent profile analysis, are all potential avenues for improving the specificity of classification in the real world.

Finally, the current paper made large strides toward representing real samples with varying amounts of careless respondents, etc. Despite this, all samples were derived from the same survey model. More concretely, the samples were all simulated from the factor correlations and item parameters estimated on the 10 facets of the IPIP we chose from a publicly available dataset. In the interest of becoming even more confident in the generalizability of our best practices, it seems worth attempting to investigate whether our findings are stable in data sampled from a variety of different survey models. I believe the methods used in the current paper could be adapted to address such inquiries.

In the meantime, I remain firm in the belief that prevention should be pursued as the primary defensive measure for dealing with careless responding, with detection serving as a necessary backup precaution. I were able to identify a method that succeeds in detecting careless respondents in such a way that it reduces the proportion of careless respondents to be lower than known problematic thresholds, without adding to the burden of survey participants. Consequently, there seems to be practical and ethical reason to eschew the use of lengthy careless responding scales or instructed response items, in favor of metrics of careless responding that can be computed and applied according to the methods described previously. For convenience, Table 5 contains a list of recommended best practices for cleaning data.

References

Ashton, M. C., Lee, K., & Goldberg, L. R. (2007). The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, *42*(8), 1515-1526. doi:10.1016/j.paid.2006.10.027

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of Random Responding on the MMPI--A. *Journal of personality assessment*, *68*(1), 139-151. doi:10.1207/s15327752jpa6801_11

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods*, *9*(1), 3-29. doi: 10.1037/1082-989X.9.1.3

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

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739-753. doi: 10.1111/ajps.12081

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who Cares and Who Is Careless? Insufficient Effort Responding as a Reflection of Respondent Personality. doi: 10.1037/pspp0000085

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software,* 48(6), 1-29. doi:10.18637/jss.v048.i06

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66,* 4-19. doi:10.1016/j.jesp.2015.07.006

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.

Evans, R. G., & Dinning, W. D. (1983). Response consistency among high F scale scorers on the MMPI. *Journal of Clinical Psychology*. doi:10.1002/1097-4679(198303)39:2<246::AID-JCLP2270390217>3.0.CO;2-9

Francavilla, N. M. (2016). *Examining the Influence of Virtual and In-Person Proctors on Careless Response in Survey Data*. Unpublished Manuscript. North Carolina State University.

Haertzen, C. A., Hill, H. E., & Belleville, R. E. (1963). Development of the Addiction Research Center Inventory (ARCI): selection of items that are sensitive to the effects of various drugs. *Psychopharmacologia, 4*(3), 155-166. doi:10.1007/BF02584088

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55. doi:10.1080/10705519909540118

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*(1), 99-114. doi:10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828. doi:10.1037/a0038510

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, *39*(1), 103-129. doi:10.1016/j.jrp.2004.09.009

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of applied psychology*, *74*(4), 657. doi:10.1037/0021-9010.74.4.657

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. doi:10.1016/j.jrp.2013.09.008

Meade, A. W., & Craig, S. B. (2012) Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. doi:10.1037/a0028085

Meade, A. W., & Pappalardo, G. (2013). Predicting careless responses and attrition in survey data with personality. In *28th Annual Meeting of the Society for Industrial and Organizational Psychology, Houston, TX*.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2 (1),* 37-63. dc.identifier.uri: https://hdl.handle.net/2328/27165

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological bulletin*, *140*(6), 1411. doi:10.1037/a0037428

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., & Schmidt, J. A. (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. *Journal of Open Source Software, 3*(30), 978, https://doi.org/10.21105/joss.00978

Saari, L. M., & Scherbaum, C. A. (2011). Identified employee surveys: Potential promise, perils, and professional practice guidelines. *Industrial and Organizational psychology*, *4*(4), 435-448. doi: 10.1111/j.1754-9434.2011.01369.x

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods*, *29*(2), 274-279. doi:10.3758/BF03204826

Scrucca L., Fop M., Murphy T.B., & Raftery A.E. (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." *The R Journal*, *8*(1), 205–233. https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf.

Thompson, A. H. (1975). Random responding and the questionnaire measurement of psychoticism. *Social Behavior and Personality: an international journal*, *3*(2), 111-115. doi:10.2224/sbp.1975.3.2.111

Thompson, L. F., & Surface, E. A. (2007). Employee surveys administered online attitudes toward the medium, nonresponse, and data representativeness. *Organizational Research Methods, 10*(2), 241-261. doi: 10.1177/1094428106/294696

Thompson, L. F., Surface, E. A., Martin, D. L., & Sanders, M. G. (2003). From paper to pixels: Moving personnel surveys to the Web. *Personnel Psychology*, *56*(1), 197-227. doi:10.1111/j.1744-6570.2003.tb00149.x

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554-568. doi:10.1016/j.chb.2015.01.070

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186. doi: 10.1007/s10862-005-9004-7

Wunder, R. S., Thomas, L. L., & Luo, Z. (2010). Administering assessments and decision-making In Farr, J.L., & Tippins, N. T. (Eds.), *Handbook of employee selection*, 377-398. New York, New York: Routledge.

Yentes R.D., & Wilhelm, F. (2018). careless: Procedures for computing indices of careless responding. R package version 1.1.3. url: https://cran.r-project.org/web/packages/careless/index.html

Yentes, R. D., Foster L. L., Meade, A. W., & Pond, S. B. (2017). *Attention and Data Quality in Online Surveys.* Paper presented at the 32nd annual meeting of the Society for Industrial and Organizational Psychology, Orlando, Florida

Table 1

List of IPIP HEXACO facets used in simulating data

Factor	Facet	+ items	- items
Honesty/Humility	Sincerity	1	9
Honesty/Humility	Fairness	5	5
Emotionality	Anxiety	5	5
Emotionality	Dependence	10	0
Extraversion	Liveliness	8	2
Agreeableness	Forgiveness	4	6
Agreeableness	Patience	5	5
Conscientiousness	Perfectionism	8	2
Openness to Experience	Inquisitiveness	6	4
Openness to Experience	Unconventionality	5	5

Note: +/- Refer to the wording of the item with respect to the construct

Table 2

List of combination methods to attempt for Research Question 2

Method	Description
Any Flag (Simul)	All indices computed and applied at the same time. Any flag is sufficient for removal
Any Flag (LSF)	Longstring computed and applied first, then Mahalanobis distance and even/odd are computed and applied. Any flag is sufficient for removal
All Flags (LSF)	All indices computed and applied at the same time. Respondents are only flagged as careless if they are flagged by all three metrics
Longstring or Agree (LSF)	Longstring computed and applied first, then Mahalanobis distance and even/odd are computed and applied. Respondents are flagged as careless if longstring flags, or if Mahalanobis distance and even/odd both flag them as careless
Longstring or Agree (Simul)	All indices computed and applied at the same time. Responds are flagged as careless if they are flagged by longstring, or if they are flagged by both the outlier and even-odd.
Longstring or Outlier (LSF)	Longstring will be computed and flagged respondents will be removed. Then outlier will be computed and flagged respondents will be removed
Longstring or Even-odd (LSF)	Longstring will be computed and flagged respondents will be removed. Then even-odd will be computed and flagged respondents will be removed

Table 3

Measures of central tendency for the index of max_B

Metric	Mean	SD	Median	Mode
Longstring	.39	.36	.40	.40
Even-odd	.22	.74	.10	-.30
Outlier	.54	.50	.50	.30

Table 4

Summary statistics for each of the methods of cleaning careless respondents

Method	M_B	SD_B	Min_B	Max_B	M_{sens}	M_{spec}	$M_R^{Careless}$	M_R
Longstring	.261	.124	-.106	.962	.274	.988	.061	.967
Even-odd	.080	.090	-.291	.410	.346	.738	.072	.731
Outlier	.175	.096	-.126	.616	.460	.715	.063	.701
Any Flag (simul)	.302	.083	.032	.589	.757	.546	.039	.522
Any Flag (LSF)	.331	.087	.013	.593	.783	.548	.035	.522
All (simul)	.015	.021	-.015	.152	.016	.999	.08	.999
Longstring or Agree (LSF)	.357	.117	-.081	.888	.469	.888	.051	.860
Longstring or Agree (simul)	.340	.113	-.081	.884	.453	.887	.052	.860
Longstring or Outlier (LSF)	.392	.108	.016	.750	.686	.706	.039	.675
Longstring or Even-odd (LSF)	.296	.103	-.133	.742	.566	.730	.050	.730
Latent Profile Analysis	.295	.103	-.092	.799	.564	.731	.050	.707

Table 5

List of recommended practices for cleaning data of careless respondents.

Stage	Recommendation
Survey Design	1. Limit survey length to minimize the occurrence of careless responding 2. Don't add items for the purpose of detecting careless respondents 3. Use scales that include a few negatively worded items to maximize the effectiveness of the longstring metric
Post-hoc Detection	4. Compute and clean data using the longstring metric at .4 standard deviations 5a. Compute and clean data using the outlier metric at .5 standard deviations 5b. (Only with caution) Compute and clean data using the outlier and even-odd metrics in combination, at .5, and .2 standard deviation thresholds. Removing only those respondents that both metrics flag as careless.

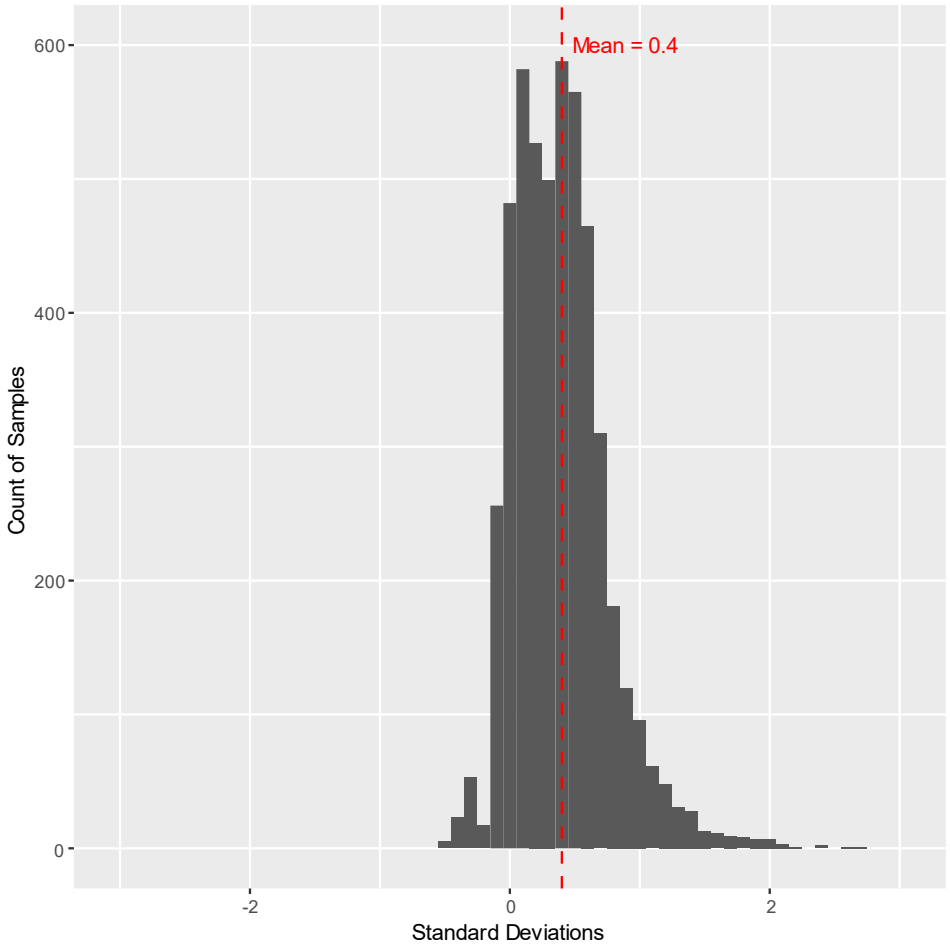

Figure 1. Distribution of max_B for the longstring metric across 5,000 samples.

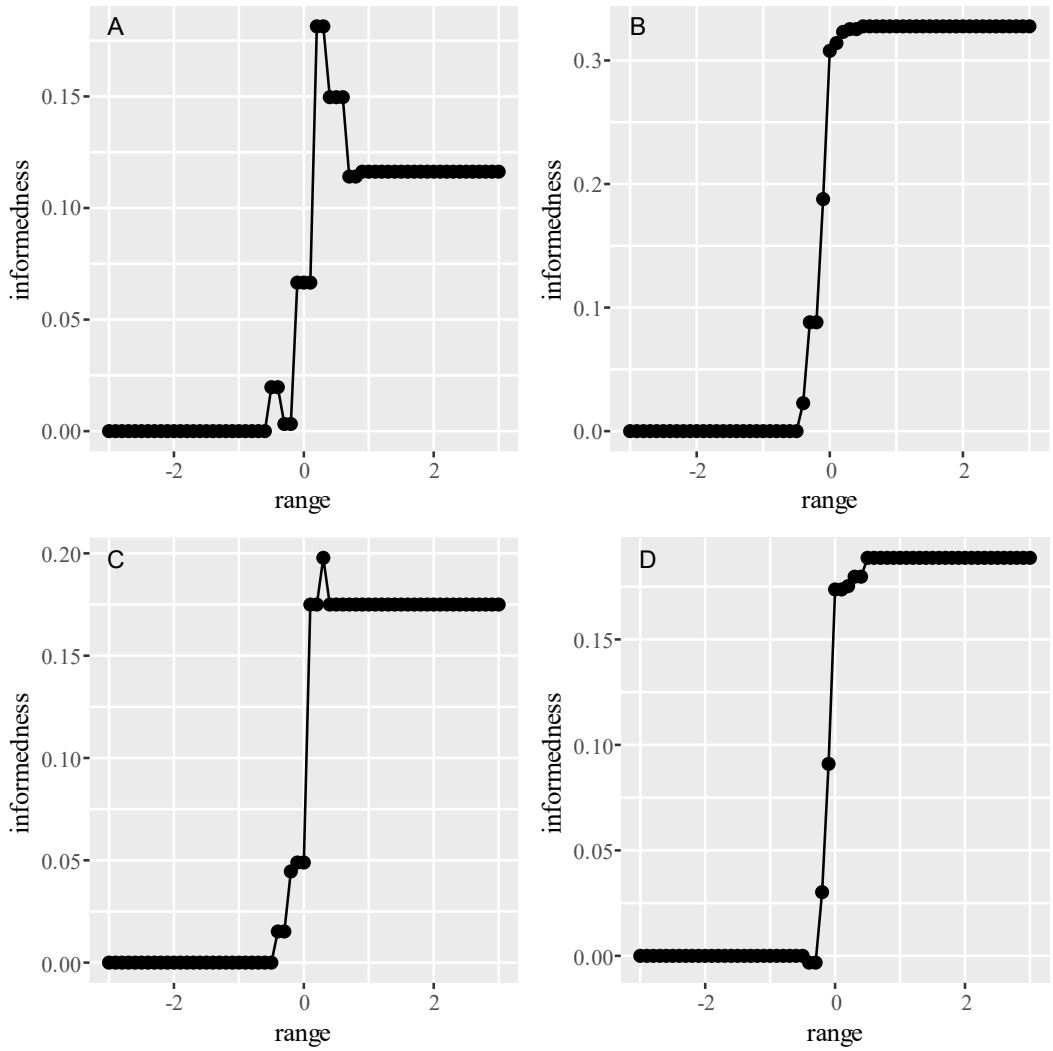

Figure 2. Plots of informedness of the longstring metric drawn from 4 random samples.

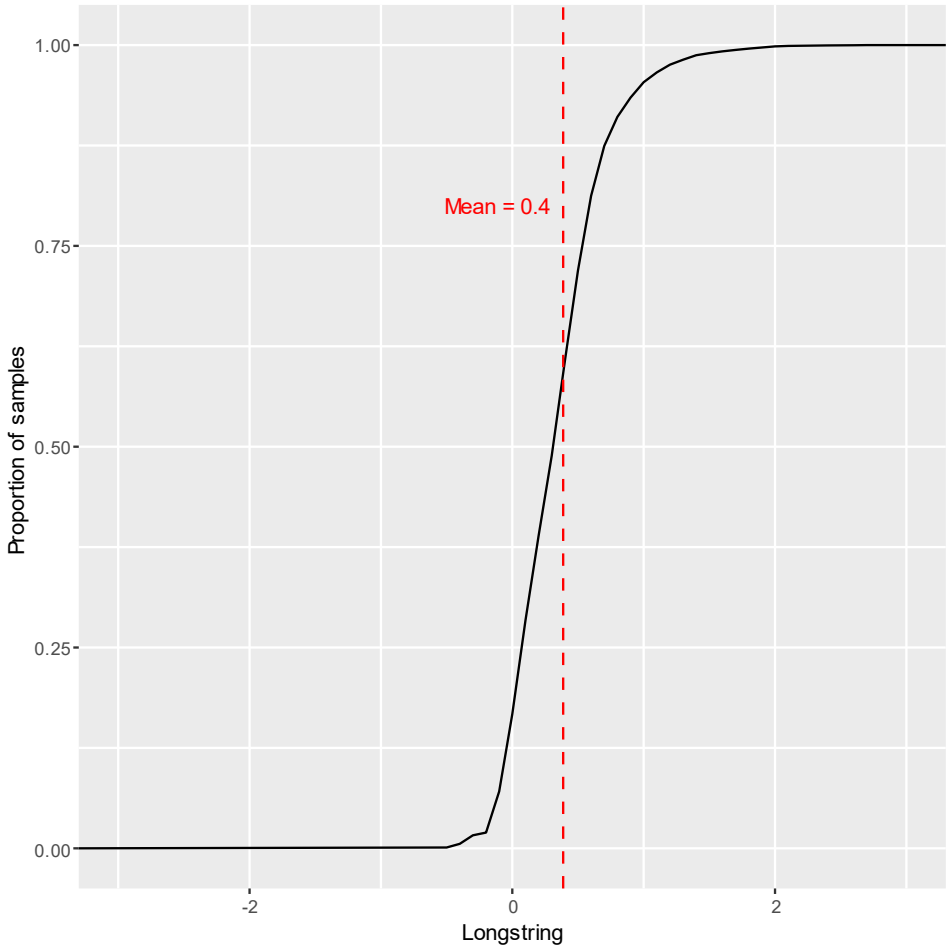

Figure 3. Estimated cumulative distribution function of longstring max_B.

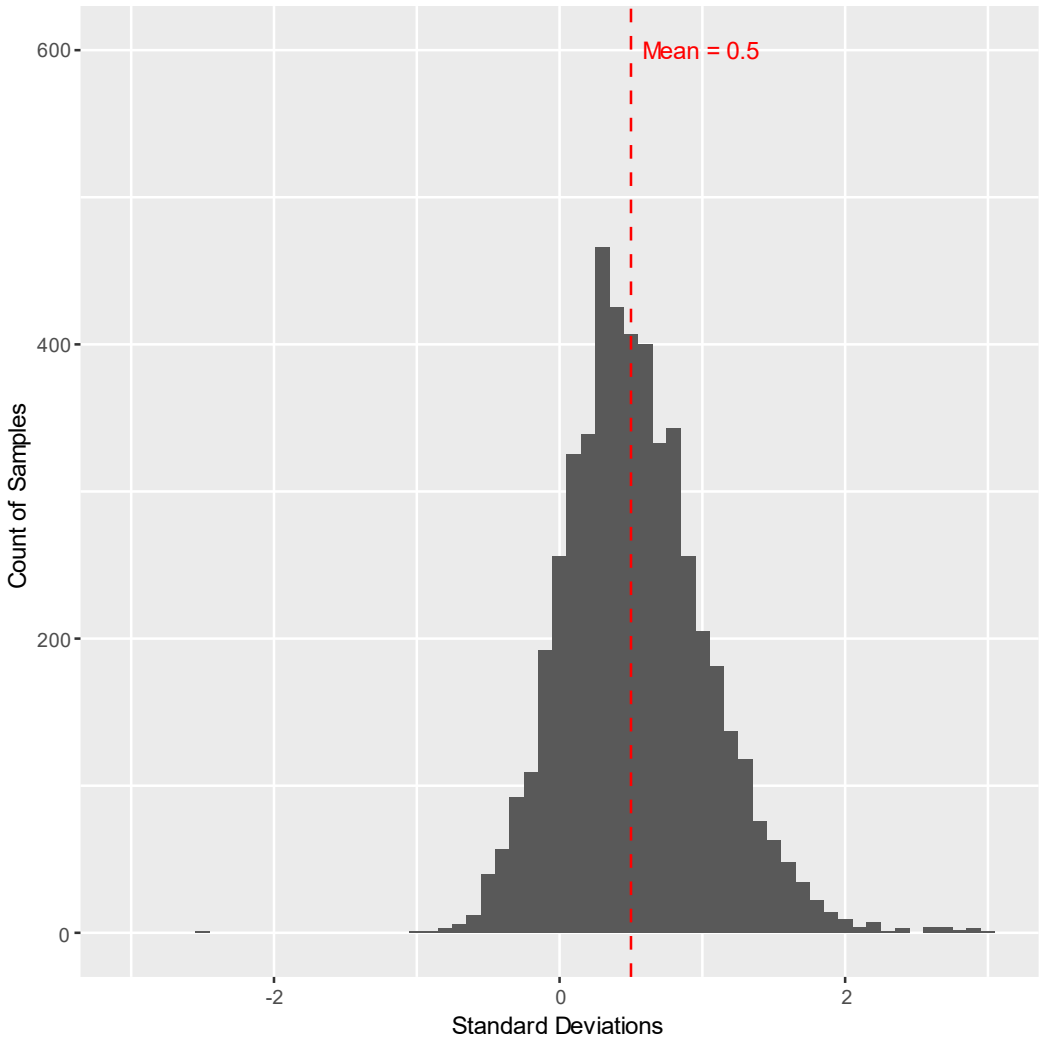

Figure 4. Distribution of max_B for the outlier metric over 5,000 samples.

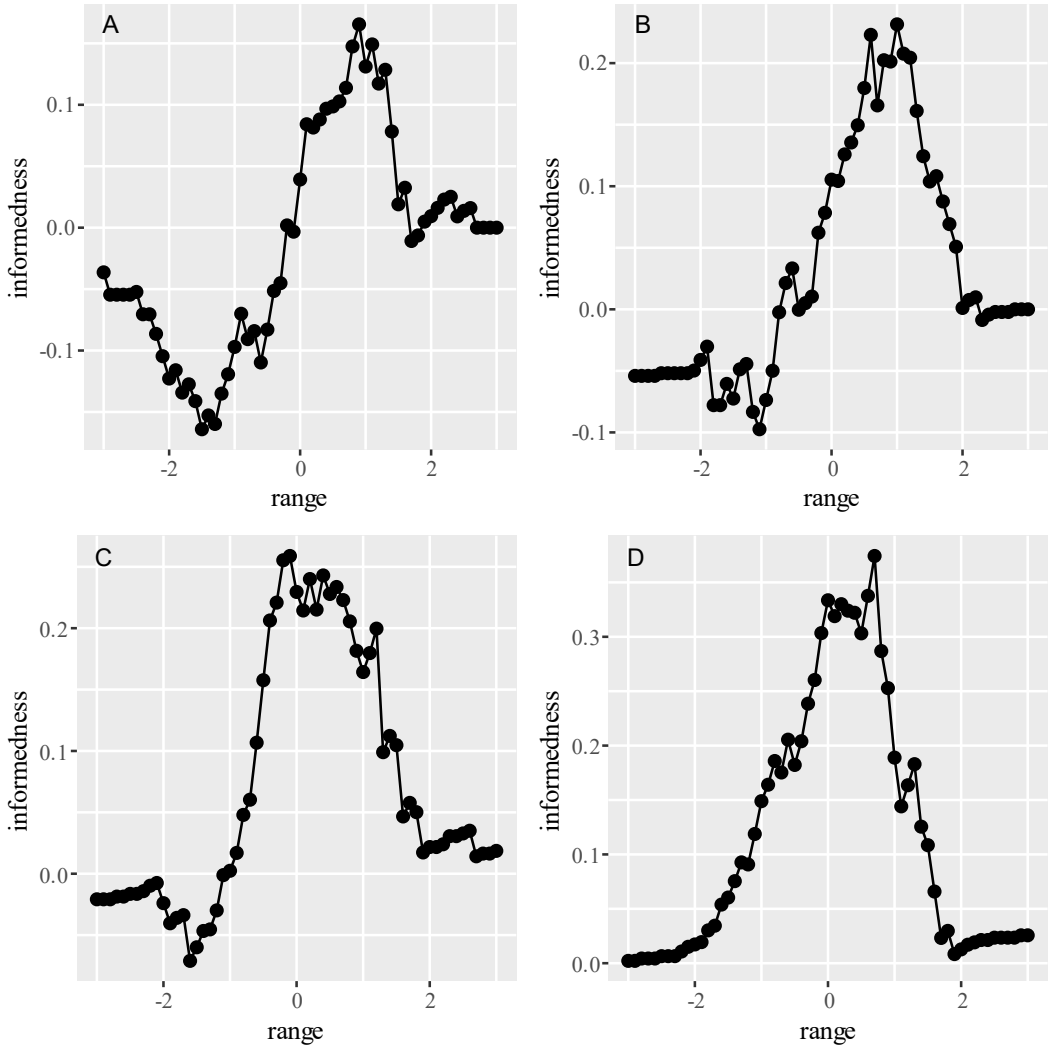

Figure 5. Plots of informedness of the outlier metric drawn from 4 random samples.

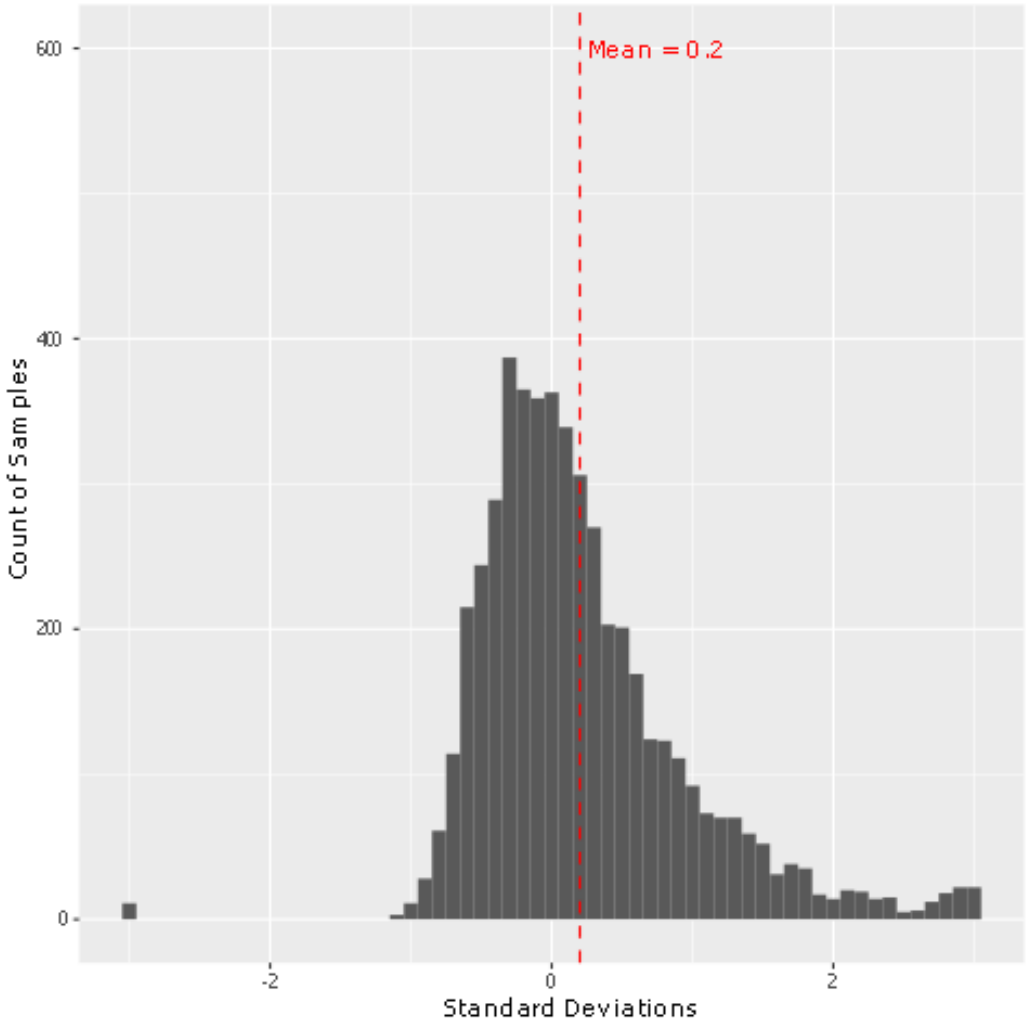

Figure 6. Distribution of max_B for the even-odd metric across 5,000 samples.

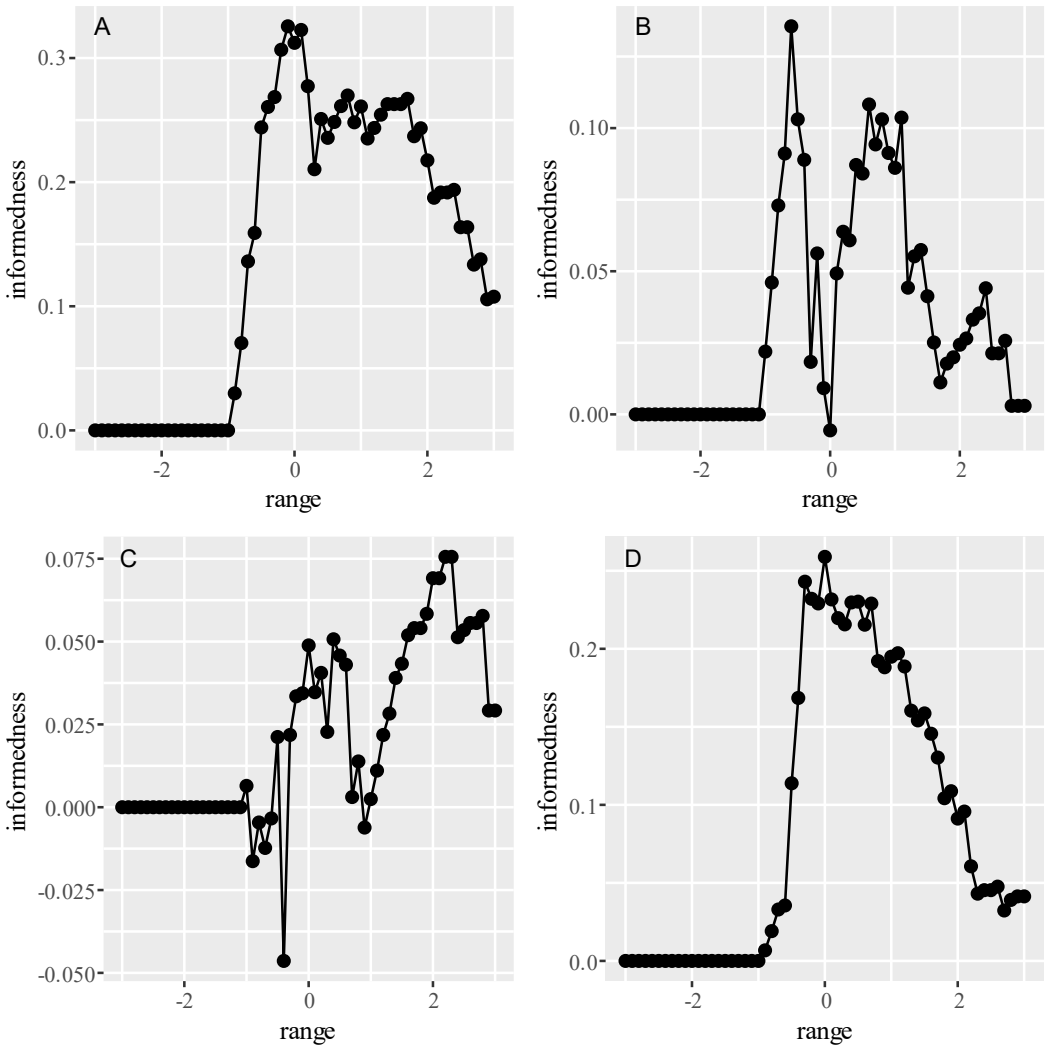

Figure 7. Plots of informedness of the even-odd metric drawn from 4 random samples.

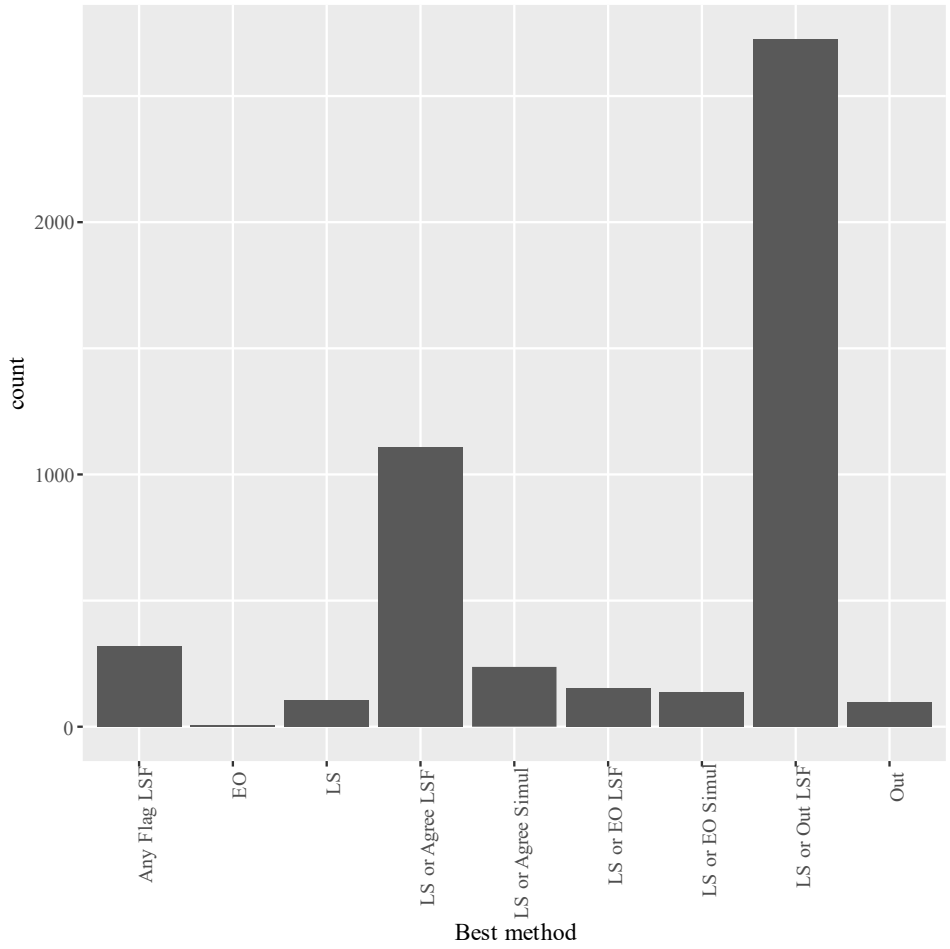

Figure 8. Frequency distribution of the method with highest informedness for each sample.

APPENDICES

Appendix A

Accessing the Code Repositories for this Project

Code for this project can be found in two repositories:

https://github.com/ryentes/dissertation contains the R code used to simulate carry out the analyses discussed int his dissertation.

https://github.com/ryentes/rdy-dissertation-tools is a lower level repository that contains many helper functions called in the main dissertation repository.

Appendix B

Original Proposal Document

In Search of Best Practices for the Identification and Removal of Careless Responders

by
Richard Dean Yentes

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2018

APPROVED BY:

_____	_____
Dr. Lori L. Foster Chair of Advisory Committee	Dr. Adam W. Meade
_____	_____
Dr. Samuel B. Pond, III	Dr. Mark A. Wilson

In Search of Best Practices for the Identification and Removal of Careless Responders

Surveys have long been popular among psychologists and other researchers as a method for collecting data from human research participants. The appeal of surveys is easily understood, as they allow researchers to access diverse, and sometimes difficult to reach populations, with efficiency of cost, speed, and ease of administration (Schmidt, 1997). Recognizing their widespread use, some researchers long ago expressed concerns that survey measures, as a method for data collection, may represent a source of measurement error (Evans & Dinning, 1983; Haertzen & Hill, 1963; Thompson, 1975). As the internet grew to permeate the world, researchers quickly adapted to the new technology and began to administer surveys online. Today online surveys and questionnaires are used in selection assessments, to collect employee and consumer opinions, for academic research, and many other purposes. While use of the new medium has allowed researchers to further capitalize upon the strengths of survey methods, it has not assuaged concerns about the quality of resulting data. If anything, online surveys have introduced new concerns, such as questions about the representativeness of online samples (Thompson, Surface, Martin, & Sanders, 2003), and the increased likelihood of inattention due to environmental distractions or multitasking (Johnson, 2005).

Recognizing the important roles that data from online surveys play in the modern world, researchers have been working to develop a literature to address threats to data quality that result when respondents do not respond carefully to a survey (Bowling et al., 2016; Huang, Curran, Keeney, Poposki, & DeShon, 2012; Johnson, 2005; Meade & Craig, 2012; Ward & Pond, 2015;). Research has confirmed that this phenomenon, which I refer to as careless responding (CR), does occur (Baer, Ballenger, Berry, & Wetter 1997; Berry et al., 1992; Maniaci & Rogge, 2014; Huang et al., 2012). Further research demonstrates that CR, when present, poses a real

threat to the validity of statistical inferences (Maniaci & Rogge, 2014; Huang et al., 2015). Accepting the occurrence and implications of CR, survey methodologists have begun developing methods to effectively counter the threat it poses. Such efforts can generally be divided into two general approaches. The first approach focuses on the antecedents of CR, and how to prevent it. In contrast, the second approach generally seeks to detect CR, so that researchers can remove data that result from it before using those data for analyses or decision making purposes. Much of the research on CR, to date, has opted for the detection approach, rather than the prevention approach.

Several different methods for detecting careless responding have been proposed, resulting in the development of numerous different indices (Meade & Craig, 2012; Johnson, 2005; Huang, 2012; Maniaci & Rogge, 2014). Consequently, it can be difficult to ascertain which indices of CR should be used, and this problem is further compounded by issues surrounding ease of computation, selection of appropriate cut scores, and the appropriate combination of multiple indices. Researchers, hoping to use methods for detecting CR to clean their data, would benefit from the creation of clear best practices that are relatively easy to implement. Thus, with the goal of deriving actionable best practices, the purpose of this paper is to evaluate the performance of several indices of CR when used in different ways, and under varying circumstances, to classify respondents as careless or not.

Careless Responding

Researchers have long been concerned with the quality of data obtained from surveys and questionnaires (Berry et al., 1992; Evans & Dinning, 1983; Haertzen & Hill, 1963; Thompson, 1975). In survey contexts, there are multiple distinct behaviors that threaten to introduce measurement error, such as malingering, and responding in a socially desirable manner;

However, several recent studies have led to a particular interest in situations where respondents complete an online survey with little or no regard for item content (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012). This phenomenon has been studied under numerous different names, including random responding, protocol invalidity, content-independent responding, content non-responsivity, insufficient effort responding, and careless responding (Huang et al., 2012). Though none of the proposed names are perfect, I prefer the term careless responding, as it is one of the earliest to appear (Haertzen & Hill, 1963), relatively easy to say, and intuitive to understand. Regardless, Huang and colleagues (2012) formally define careless responding as:

> "a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses."

However, it is important to note that careless respondents may not engage in careless responding for the entire duration of a survey. Instead, survey respondents may carefully respond to survey items in the beginning of a survey, and begin to respond carelessly later, as their attention wanders. This pattern is consistent with reports from respondents reported by early research on careless responding (Berry et. al, 1992). One reason for this relates to recent findings that careless responding is closely related to attention (Yentes, Foster, Meade & Pond, 2017). Cognitive Resource theories of attention suggest that maintaining attentional control on a task becomes progressively harder as its duration increases, and that this is likely to be particularly true for tasks that are easy or dull (Kanfer & Ackerman, 1989; Randall, Oswald, & Beir, 2014).

The Importance of Careless Responding

Though concerns about careless responding have existed for quite some time, initially they were somewhat speculative. Thus, some research on careless responding has sought to

estimate its prevalence, and to establish whether or not it is, indeed, a threat to data quality. Reports on the prevalence of carless responding vary from as little as 1% (Huang, Liu, and Bowling, 2014) to as much as 73% (Baer, Ballenger, Berry, & Wetter, 1997). This discrepancy is likely attributable to differences in the criteria for classification as a careless responder. When responding carelessly to a single item is sufficient, reports of careless responding tend to be much higher (Baer et al., 1997; Berry et al., 1992; Meade & Pappalardo, 2013). In contrast, when more refined criteria are applied, much lower estimates of careless responding, generally between 3-12% are reported (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012; Meade & Pappalardo, 2013). While some evidence does suggest that careless responding may be slightly more prevalent in undergraduate samples (Meade & Craig, 2012), research has shown that it occurs even among respondents who self-selected into a survey (Johnson, 2005), and in relatively high-stakes situations, as when a survey is taken as part of the job application process (Baer et al., 1997; Ehlers, Green-Shortridge, Weekley, & Zajack, 2009).

Given that careless responding does occur, researchers have also worked to demonstrate its effects on data quality. Much of the early research on careless responding was conducted with various clinical diagnostic inventories, such as the MMPI, or ARCI (Evans & Dinning, 1983; Haertzen & Hill, 1963; Nichols, Greene, & Schmlock, 1989; Thompson, 1975). These studies seem to focus on careless responding in an effort to ensure correct diagnosis, and though validity and reliability are mentioned, problems resulting from careless responding are not specifically articulated. The merit of these concerns seems intuitive; however, scientific skepticism does leave room to question whether careless responding is prevalent enough for these problems to manifest in ways that are practically meaningful. More recently, researchers have articulated

specific concerns, with accompanying studies that provide empirical evidence to inform a definition of problematic levels of careless responding.

One concern is that careless responding may hinder the validation of psychometric tests and measures by altering the factor structure of the constructs under investigation. For example, Woods (2006) designed a simulation to test the effects of careless respondents on the factor structure of a unidimensional scale with reverse worded items. When the sample was comprised of 5% careless respondents, indices of fit for confirmatory factor analysis of a unidimensional model were good according to commonly accepted standards (i.e. Hu & Bentler, 1999). At 10% careless responders, indices of fit for a unidimensional model were only marginal (e.g. CFI > .91; TLI = .95; RMSEA = .07). When a sample was comprised of more that 10% careless responders, fit of a unidimensional model was unacceptable by any modern standard. In a different study, Johnson (2005) split respondents into quartiles according to two separate inconsistency indices. Responses to the IPIP-NEO, provided by the top and bottom quartiles, were then independently subjected to principle components analysis and compared. Though factor loadings were lower for the bottom quartile, the five-factor structure of the IPIP-NEO was still observed, and items generally loaded onto the appropriate factors. Taken together, these two studies suggest that careless responding is likely only problematic when it exceeds a certain threshold.

Other studies have attempted to provide empirical evidence that careless responding leads to an increased chance of Type II errors. Maniaci and Rogge (2014) found that statistical power was reduced when careless responders comprised as little as 5% of a sample. Additionally, they demonstrated that robust research findings, which replicated among careful respondents, failed to replicate among careless respondents; however, their test was rather lenient, as they utilized

homogenous groups of careless and careful respondents, which are unlikely to occur in practice. That said, further research has provided more empirical evidence that careless responding can bias parameter estimates. Specifically, when careless responders comprise as little as 5% of a sample, it can bias parameter estimate either upward or downward, depending on the position of the population mean relative to that of careless respondents' (Huang, Liu, & Bowling, 2014). Said differently, depending on the circumstances, careless responding can increase the risk of Type I or Type II errors. Thus the preponderance of evidence suggests that careless responding does have important consequences, and that it is prevalent enough for those consequences to manifest in ways that are practically harmful. Therefore, given these findings, and particularly in light of recent concerns about replicability in psychological research (Open Science Collaboration, 2015), it is important to find ways to mitigate the threat that careless responding poses for psychological science.

Managing the Threat of Careless Responding

Along with efforts to establish the threat posed by careless responding, researchers have also sought to identify methods for mitigating its effects. Studies in this area generally adhere to one of two approaches. The first approach to dealing with careless responding is to identify its motivational antecedents with the goal of designing research and instruments in such a way that respondents do not engage in careless responding. The second approach for addressing careless responding is to identify survey respondents who were careless, and to remove their data before conducting further analyses. Thus, techniques for mitigating the impact of careless responding can be categorized depending on whether their goal is the prevention of careless responding, or its detection.

Though these two strategies for assuaging concerns related to careless responding are not mutually exclusive, some researchers have argued that there is reason to prefer prevention to detection. One reason for this preference is that removing data from careless responders may skew the sample distribution in a way that biases estimates (Berinsky, Margolis, & Sances, 2014; Ward & Pond, 2015). More concretely, there is growing evidence indicating that careless responding manifests, in part, as a function of respondent personality (Bowling et al., 2016). When personality is relevant to the phenomena under investigation, then the removal of data from careless responders is likely to remove the very sub-populations that are the focus of researchers' interest. When this problem is considered along with other concerns relating to wasted time, effort, and money, it is apparent that it would be ideal to prevent careless responding before it occurs.

To date, a number of methods for preventing careless responding have been proposed, many in the form of suggestions for survey design. For example, one option is to add a warning to survey instructions, informing participants that their responses will be checked for quality, and that credit and rewards will not be awarded to those that provide low-quality responses. Empirical evidence suggests that such warnings can be effective, though survey respondents generally react negatively to them (Huang et al., 2012). Additionally, some evidence indicates that this effect can be augmented through the presence of a virtual human (Ward & Pond, 2015). Similarly, Meade and Craig (2012) altered their instructions to require that respondents sign their name on each page of the survey. Manipulation of respondent anonymity did reduce CR, even though the instructions also assured participants that their responses would be confidential. Drawing upon theory from the literature on motivation, Yentes, Foster, Meade, and Pond (2017) argued that long surveys place a larger burden on respondents, leading to careless responding. In

their study, careless responding was more prevalent among participants randomly assigned to take a long version of a survey, as compared to those assigned to take a shorter version of the same survey. Several additional methods have also been investigated, though researchers have generally been unable to demonstrate their effectiveness. For example, in two studies, the presence of a proctor, virtual or otherwise yielded no or small main effects for reducing CR (Francavilla, 2016; Ward & Pond, 2015). Similarly, neither the inclusion of a survey progress bar, nor the disclosure of an estimate of the time commitment required to complete a survey, resulted in less careless responding (Yentes, Foster, Meade & Pond, 2017).

Despite the identification of several methods that have been demonstrated to reduce careless responding, there are several problems with relying solely upon prevention as a strategy for addressing it. First, even those methods for which statistically significant effects were observed, effect sizes are typically modest. Therefore, while these methods may prevent some occurrences of careless responding, they may not prevent enough to effectively mitigate the threat it represents for statistical inference. Second, even effective methods are not without drawbacks, and must be applied situationally. For example, it may not always be possible to design a short survey that is adequate for its intended purpose. Finally, of the three effective methods described previously, two of them place the burden of dealing with the problems posed by careless responding on the survey respondent, as evidenced by negative reactions to warnings embedded in survey instructions (Huang et. al, 2012), and discomfort reported by respondents when perceived anonymity is low, despite assurances of confidentiality (Saari & Scherbaum, 2011; Thompson & Surface, 2007). As a solution, this is undesirable. Both ethically, and because respondents' satisfaction with survey experiences is predictive of their willingness to complete future surveys (Thompson, Surface, Martin, & Sanders, 2003). Thus, for all of the reasons

discussed here, researchers cannot address the threat posed by careless responding by relying solely on techniques to prevent it occurring.

Methods for Detecting Careless Responding

The performance of any given technique designed to prevent careless responding is typically evaluated with respect to how much reduction it causes in some measure(s) of careless responding. Consequently, as researchers have sought to establish techniques for preventing careless responding, they have also developed methods to more accurately distinguish diligent survey respondents from those engaging in careless responding. As mentioned previously, many researchers consider the detection of careless responders, and removal of their data from a sample prior to conducting analyses, to be a worthwhile strategy for countering the threats to scientific inference that careless responders introduce. As a result, myriad different indices of careless responding have been proposed and studied. Though an exhaustive review of these indices is beyond the scope of this paper, a brief overview of the most common types of CR indices follows. Readers requiring more detail may refer to Curran (2016) for a more exhaustive overview.

Infrequency. Indices of careless responding that use the infrequency approach rely upon the assumption that it is possible to write items to which careful respondents will almost always provide a similar response. For example, if a survey contained the item “I used to live on the planet Mars.”, it would be highly improbable for any careful respondent to report agreement with this item. Similarly, instructed response items are a special case of infrequency items, in which researchers explicitly tell survey participants to select a specific response option (Meade & Craig, 2012). Researchers generally include a number of such items distributed throughout a survey, and an index score is computed as the sum of total endorsement across all such items, or

as the total number of items to which a survey participant selected the incorrect response option. The infrequency subscale of the Attentive Responding Scale (ARS) is an example of a validated infrequency index of careless responding (Maniaci & Rogge, 2014).

Inconsistency. Inconsistency measures of careless responding rely upon the assumption that careful respondents should generally respond to similar questions in a similar manner. For example, one would not generally expect a respondent to strongly endorse the statement "I'm the life of the party", right after strongly endorsing a statement like "I hate being the center of attention". It is conceivable that such circumstance may arise as a result of measurement error, or some idiosyncrasy of the respondent; however, given the literature on cognitive dissonance and self-perception, it would be very unlikely that a careful respondent would exhibit a general pattern of inconsistent responses to a well-designed survey. Consequently, Inconsistency measures of careless responding quantify a pattern of inconsistency with respect to similar or related items in a survey.

Though inconsistency measures share a common approach to detecting careless responding, there is considerable variation in the manner in which items are chosen and compared. For some measures, item pairs are specifically designed for this purpose, as is the case with the inconsistency subscale of the Attentive Responding Scale (Maniaci & Rogge, 2014). One disadvantage of such inconsistency measures is that they add to the burden of the survey respondent by adding items to the survey. For example, the ARS inconsistency subscale necessitates adding 12 items (for the ARS-18; 22 for the ARS-33) that are unrelated to research questions motivating the survey.

Some other measures, like the psychometric synonyms index, rely upon empirically derived item pairs (Johnson, 2005; Meade & Craig, 2012). These indices identify highly-

correlated items in the overall sample, and then compute the within-person correlation for the identified item-pairs. One drawback to this technique is that it is dependent upon sufficient correlated item pairs to manifest in the sample. No formal best practices have been established for minimum correlations for such item pairs, though some researchers have used and/or recommended .60 as a minimum requirement (Curran, 2016; Meade & Craig, 2012). In samples where no such pairs exist, it would not be advisable to compute or interpret the psychometric synonyms index.

The even-odd index is another variant of an inconsistency index, which is analogous to the application of split-half reliability to respondent consistency. For this technique, researchers split validated unidimensional scales that are included in their survey into two sub-scales, one for the even numbered items, and the other for the odd numbered items. Sub-scale scores are calculated as the average of items on each sub-scale. Then the even-odd index score is computed as the within-person correlation between the paired sub-scale scores, and then corrected for decreased length of the measure using the Spearman Brown prophecy formula (Meade & Craig, 2012).

Invariability. Longstring indices of careless responding were created based on the idea that careless respondents may select the same response for longer strings of contiguous items than careful respondents. They are generally computed by first examining survey participants' individual item responses to determine whether they are the same as their response to the preceding item. Researchers then create a longstring score for each respondent, typically either as the longest string of consecutive identical responses (e.g. Meade & Craig, 2012), or as the average of the lengths of the longest string on each page (Huang et al., 2012).

Outlier Analysis. The use of outlier analysis as an index of careless responding stems from the notion that disregard for item content is characteristic of careless responding. Consequently, for each item, a careless responder should have a consistently low probability of selecting a response that is close to the item mean. Conversely, unless a careful respondent were uniformly atypical, their responses should be more likely to trend toward item means. Mahalanobis' distance is a multivariate outlier statistic that, conceptually, is a measure of the distance between a survey participant's response vector, and the vector of item means in the overall sample. Thus, researchers have applied Mahalanobis' distance to the detection of careless responders (Ehler's et al., 2009; Meade & Craig, 2012).

Types of Careless responding

An open question pertaining to the nature of careless responding is whether there are different types of careless responding. One way to approach this question is to apply statistical classification methods to indices of careless responding. At least two recent studies applying such methods generally agree that a three class solution describes the data very well (Maniaci & Rogge, 2014; Meade & Craig, 2012). These classes represent groups of respondents who are roughly typified by one of three patterns. The first group of respondents provide varied, yet consistent responses, as indicated by better scores across indices of careless responding. This response pattern is generally consistent with the expected behavior of a respondent diligently answering survey items. Consequently, respondents who demonstrate this pattern are referred to here as diligent respondents. A second pattern is characteristic of respondents who provide the same response to relatively large numbers of contiguous questions, and are distinguished by elevated scores on longstring indices. Accordingly, researchers may think of respondents with this pattern of behavior as longstring careless respondents. The third pattern is characterized by

relatively varied responses, with poorer scores on indices of careless responding based on consistency with self, or others (e.g., inconsistency and outlier indices). Consequently, respondents exhibiting this pattern are referred to as generally careless. Both longstring and generally careless respondents also tended to provide incorrect responses to infrequency items.

Different patterns of careless responding may be especially important in light of the nature of how indices of careless responding are computed and used for detection. The maximum long string index is determined by the pattern of responses provided by a survey participant, and is unaffected by data provided by other participants. In contrast, consistency and outlier indices of careless responding are computed with respect to sample means and correlations. Thus, unless identified and removed prior to computing consistency and outlier indices, the data provided by brazen careless respondents will distort the scores that are generated. Consequently, it is reasonable to suggest that the removal of longstring careless respondents, prior to computing consistency and outlier indices, should only improve their ability to accurately distinguish between diligent respondents and generally careless respondents. Thus, the following hypotheses are proposed.

Hypothesis 1: Cleaning data using the maximum long string index before computing the outlier index of careless responding will improve the outlier index's informedness

Hypothesis 2: Cleaning data using the maximum long string index before computing the even/odd index of careless responding will improve the even/odd index's informedness

Making Decisions About Careless Responders

In practical terms, the purpose of any index of careless responding is generally to inform decisions about whether or not a survey respondent should be considered careless. Such decisions may be used to inform a variety of goals, such as academic research on the nature of

careless responding, or for more applied purposes, like cleaning a data set. Regardless of the purpose for making such decisions, there is a need for evidence-based recommendations defining a process for bridging the gap between computing numerous indices of careless responding and reaching a final careless or not careless decision. The creation of such a process implies a stance on two key issues; namely, how cut scores can be used to inform these decisions, and whether indices of careless responding can be used together to arrive at better decisions.

The simplest process for making a decision about careless responding is to simply compute a single index of careless responding, choose a cut score, and then classify each respondent as careless or not accordingly. This method is unlikely to perform optimally because, as mentioned previously, several different types of careless responding are thought to exist. Further, categories of indices for detecting careless responding have different strengths and weakness with respect to what type of careless responding they can identify (Curran, 2016). Consequently, it is both intuitive and reasonable to conclude that using some combination of different indices of careless responding should be preferable to the simple use of one index.

To arrive at an actionable process, it is necessary to be more specific about how different indices should be combined to inform a final decision. Some empirical evidence suggests that a simple combination of indices may have detrimental effects (Maniaci & Rogge, 2014). Said differently, classifying a respondent as careless if they trigger a cut score on any one index of careless responding is unlikely to be optimal. Additionally, research on selection suggests that mechanical combination systems typically outperform human judgement (Wunder, Thomas, & Luo, 2010). And, though some rules of thumb exist (e.g. Curran, 2016), there is considerable uncertainty with respect to how arrive at a cut score that can be generalized from one survey to another, and computed without knowledge of who is careless and who is not. However, as

mentioned previously, in some cases it may make sense to cut on one index before computing other indices of careless responding, thus some researchers have recommended using a multiple-hurdle approach to identifying careless respondents (Curran, 2016). Such a practice has the benefit of trading off the loss of information resulting from the dichotomization of continuous indicators, prior to a final decision, with the possibility that the performance of some indices of careless responding may be distorted by presence of some relatively careless respondents that are relatively easy to detect. Regardless, it is clear that the matter of combining the information provided by different indices of careless responding is not as simple as it may initially seem, and more empirical information about what constitutes a good cut score for any given index of careless responding is needed. One technique that some researchers employ (e.g. Yentes, Foster, Meade & Pond, 2017), is to use respondents' standard deviation from the sample mean on an index as the basis for a cut score. In order to determine the effectiveness of this practice, the following research question is proposed.

Research Question 1: How well do cut scores based on deviations from the norm perform?

Several researchers have suggested that the use of different indices of careless responding in combination should perform better than relying on any one single index (Curran, 2016; Huang et Al. 2012; Meade & Craig, 2012). Despite this, researchers have found some methods of combining indices of careless responding to perform poorly (Maniaci & Rogge, 2014). However, there is a vast number of possible ways to combine different indices of careless responding that are, as yet, untested. For example, the generalization of hypotheses 1 and 2, that computing and cleaning based on the long string index prior to computing other indices of careless responding will improve accuracy, implies that a flag on the long string index should be sufficient for exclusion from further analyses. However, an alternative method could require that at least two

indices, computed on uncleaned data, must flag a respondent as careless, in order to exclude them for further analyses. While that it would be impractical to examine all potential rules in any one study, given that few have actually been examined, it is too early to dismiss the notion that using CR indices in combination could improve upon the accuracy of any given individual index. Thus, the following research question is proposed.

Research Question 2: Can methods for combining the information provided by indices of careless responding improve upon the accuracy of the individual indices?

Latent Profile Analysis (LPA) is a statistical technique that attempts to identify homogenous groups in empirical data by examining the means and correlations among variables (Bauer & Curran, 2004). In contrast to the Common Factor Model that underlies factor analysis, LPA assumes that correlations between variables are indicative of the presence of discrete groups, and not a continuous latent variable. In LPA, distinct groups are characterized by their mean vectors with respect to the input variables. As mentioned previously, researchers applying latent profile analysis to indices of careless responding have generally found that a three-class solution fits the data well (Maniaci & Rogge, 2014; Meade & Craig, 2012). Given that these classes are generally interpreted as corresponding to careful (i.e. not careless), longstring careless, and generally careless responding, it would be reasonable to suggest that LPA could be capable of classifying careless respondents. Thus, the following research question is proposed.

Research Question 3: Does LPA function well as an index of careless responding?

Methods

Data Model

Datasets were simulated using a method similar to that of Meade and Craig (2012). A model of factor correlations and item parameters was obtained from a public dataset of responses to the

IPIP HEXACO measure, which was downloaded from ‘http://personality-testing.info/_rawdata/’. The dataset consists of 240 items spread across six factors, each with four facets. Four additional items were also present, indicating self-reported understanding of the instructions, accuracy, the country of the respondent, and the number of seconds between the survey being opened and its submission. Data from respondents who reported that they did not understand the instructions, or that they had not answered accurately, were dropped. Additionally, data from respondents who completed the measure in less than 17.5 minutes, or in more than 2 hours were also dropped to help ensure that the dataset consisted of meaningful responses. The final dataset used contained 10,216 observations.

A subset of 10 facets were chosen from across the factors to serve as a 100-item measure. Table 1 contains a list of the chosen facets, as well as the number of positively and negatively coded items that each facet contains. Correlations between each of the selected facets were computed. IRT item parameters for the graded response model were estimated for each facet using the MIRT package in R (Chalmers, 2012). Both the correlation matrix and estimated item parameters were used in subsequent data simulation.

Data Simulation

A total of 5,000 simulated datasets were created, each with 500 rows (i.e. simulated respondents). The R-script used to simulate these datasets can be found in Appendix X. The matrix of correlations obtained from the HEXACO dataset served as input into a Choleski decomposition, which was computed to obtain weights for use in creating facet-level scores for each simulated respondent. Then, the previously estimated item parameters were used to simulate item-level responses. The resulting datasets contained simulated responses ranging from 1 to 7, with each row representing a simulated careful survey respondent, and with a population

correlation among factors identical to that of the observed data. Next, rows designed to simulate careless responders were generated and introduced into the dataset.

Careless responses were simulated according one of two models, namely, either longstring, or generally careless. The percentage of respondents following each model varied between datasets. The percentage of longstring careless respondents, in any one replication sample, was determined by sampling a normal distribution with a mean of .02 and a standard deviation of .01. Similarly, the proportion of generally careless respondents was sampled from a normal distribution with a mean of .06 and a standard deviation of .02. These values were chosen so that the total proportion of careless respondents for each sample would center around the midpoint of the 3-12% range observed in prior research, with substantial variation. The parameters for longstring careless responders were set lower in order to maintain consistency with empirical evidence that longstring careless responding is less prevalent than generally careless responding (Maniaci & Rogge, 2014; Meade & Craig, 2012). The dataset was then randomly sampled without replacement until a number of rows proportional to the total percentage of careless respondents were selected. These rows were designated careless respondents, and this subset was then randomly sampled without replacement in order to proportionally assign careless respondents to either the longstring or the general careless model.

Careless response models. For careless respondents that were simulated from a longstring model, one response option was randomly chosen, which was used for each careless response. For those that followed the generally careless model, careless responses were randomly selected from a normal distribution with a mean of 3.5 and a standard deviation of 1.25. A resilience score was generated for each careless respondent, from a normal distribution with a mean of 50, and a standard deviation of 10. This resilience score represented the point at which a survey

participant would begin resorting to careless responding; therefore, if a simulated careless respondent had a resilience score of 50, then only responses to items 50-100 were replaced with careless responses. The resulting datasets were then complete, and ready for substantive analyses. More specific details of the implementation of these data generation procedures can be found in appendix A.

Measures

Indices of careless responding. Three different indices of careless responding were examined, each representing a different category of careless responding indices. Namely, maximum longstring was chosen as a measure of invariability; the even-odd index, as described by Meade & Craig (2012), was used as a measure of individual consistency; and finally, Mahalanobis' distance served as a multivariate outlier measure. These indices were computed using the careless R package (Yentes, 2016). Following recommendations from Curran (2016), maximum longstring was computed first, on the raw response vectors. Then reverse worded items were reverse coded, and both even-odd consistency and Mahalanobis' distance were computed on these reverse coded response vectors.

Classifier Performance. As mentioned previously, the focus of this paper is on developing an evidence-based process for arriving at a final determination of whether a survey respondent should be considered careless. The general form of this problem is known as binary classification. While there are a variety of named metrics for evaluating the performance of binary classifiers (e.g. sensitivity, specificity, accuracy, etc.), classifier performance was operationalized using Bookmaker Informedness which, conceptually, is a quantification of how informed a predictor is with respect to a given condition (Powers, 2011). Informedness is computed as:

$$B = 1 - fnr - fpr$$

Where *fnr* is the false negative rate, which is equal to the proportion of false negatives resulting from prediction, divided by the proportion of real negatives in the sample, and *fpr* is the false positive rate, which is equal to the proportion false positives resulting from prediction, divided by the proportion of real positives in the sample. Informedness was preferred over other metrics because it is unbiased with respect to its treatment of positive and negative cases, and confidence intervals for it are easily computed using the Standard Error (Powers, 2011).

$$SSE_B = (1 - B)^2$$

$$SE = \sqrt{[SSE_B/(N - 1)]}$$

An additional reason for preferring Informedness to other measures is because of its relationship with ROC analyses, which are also sometimes used for model evaluation. A roc curve is a plot of the sensitivity of a model on the y axis, against 1-specificy on the x axis. The diagonal of this plot generally represents chance performance, and a model performance is interpreted in terms of the total area between the curve and the line representing chance. Importantly, the line is plotted across all possible thresholds (cut scores) for the predictor, thus informedness is defined for every point in the line, and by maximizing informedness we arrive at a basis for deciding upon a cut score.

Proposed Analyses

Hypothesis 1. The purpose of hypothesis 1 was to determine whether removing data from survey respondents that match a longstring pattern of careless responding, prior to computing the outlier index of careless responding, would improve its informedness. In order to test this hypothesis, the long string and outlier indices will both be computed for each full sample. Then, longstring careless respondents will be flagged and removed, and the outlier index will be re-computed on

the reduced sample. In both cases, the top performing cut scores, as determined by the results of Research Question 1 will be applied. Performance of the outlier index will be assessed by comparing the informedness of the outlier index computed on the full sample, with the outlier index computed on the reduced sample. If the confidence intervals of informedness are non-overlapping between the two computations of the outlier index, and the outlier index computed after cleaning a dataset using the longstring index performs better, in at least 75% of the samples, then hypothesis 1 will be supported.

Hypothesis 2. The purpose of hypothesis 2 was to determine whether removing data from survey respondents that match a longstring pattern of careless responding, prior to computing the even/odd index of careless responding, would improve its informedness. In order to test this hypothesis, the long string and even/odd indices will both be computed for each full sample. Then, longstring careless respondents will be flagged and removed, and the even/odd index will be re-computed on the reduced sample. In both cases, the top performing cut scores, as determined by the results of Research Question 1 will be applied. Performance of the even/odd index will be assessed by comparing the informedness of the even/odd index computed on the full sample, with the even/odd index computed on the reduced sample. If the confidence intervals of informedness are non-overlapping between the two computations of the even/odd index, and the outlier index computed after cleaning a dataset using the longstring index performs better, in at least 75% of the samples, then hypothesis 2 will be supported

Research Question 1. Research question 1 is proposed to determine how well a survey respondent's number of standard deviations from the mean works, with respect to different indices of careless responding, when used as a cut score. In order to answer this question, the Longstring, Outlier, and Even/Odd indices of careless responding will be computed, and left in

their raw form. For each index of careless responding, an ROC plot will be generated, depicting its performance when using standard deviations from the mean of that index, ranging from 1 to 3, in .1 standard deviation steps, as a cut score. Overall performance of standard deviations as a basis for cut-scoring will be assessed by examining the maximum value of informedness, averaged across all samples. If informedness is better than chance in at least 75% of samples, then using standard deviations from mean index scores may be recommended. Additionally, if there is relative stability across samples for the value of SD at which informedness is maximized, then that may serve as a data point in favor of the generalization of a best practice for using standard deviations to create a cut score.

Research Question 2. Research question 2 is proposed to assess whether information from different indices of careless responding can be combined to improve upon the accuracy of individual indices. In order to answer this question, a variety of different combination methods will be computed and assessed. Table 2 contains a full list of the methods that will be tested. As with hypothesis 1, informedness will be computed for each of the different combination methods. If the average informedness, across the simulated samples, of any of the combination methods is higher than the average informedness of the top performing cut scores from research question 1, then Research question 2 will be answered in the affirmative.

Research Question 3. Research question 3 is proposed to determine whether latent profile analysis (LPA) can be usefully applied as a classifier of careless respondents. To answer this research question, a three-class LPA will be applied to respondents standardized scores on the three raw indices of careless responding. Respondents will then be assigned to a class using either the probability of class membership output by the model, or the nearest neighbor of their score vector, with respect to the mean vectors of the classes output by the model. To achieve a

binary classification, an algorithm will designate one class as careful respondents, and the other two will be assumed to represent careless respondents. The effectiveness of this process will be evaluated by comparing its informedness, averaged across all samples, to the accuracy of the methods previously described.

References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American psychologist*, *57*(12), 1060-1073.

Ashton, M. C., Lee, K., & Goldberg, L. R. (2007). The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, *42*(8), 1515-1526. doi:10.1016/j.paid.2006.10.027

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of Random Responding on the MMPI--A. *Journal of personality assessment*, *68*(1), 139-151. doi:10.1207/s15327752jpa6801_11

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods*, *9*(1), 3-29. doi: 10.1037/1082-989X.9.1.3

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

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739-753. doi: 10.1111/ajps.12081

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who Cares and Who Is Careless? Insufficient Effort Responding as a Reflection of Respondent Personality. doi: 10.1037/pspp0000085

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software,* 48(6), 1-29. doi:10.18637/jss.v048.i06

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66,* 4-19. doi:10.1016/j.jesp.2015.07.006

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.

Evans, R. G., & Dinning, W. D. (1983). Response consistency among high F scale scorers on the MMPI. *Journal of Clinical Psychology*. doi:10.1002/1097-4679(198303)39:2<246::AID-JCLP2270390217>3.0.CO;2-9

Francavilla, N. M. (2016). *Examining the Influence of Virtual and In-Person Proctors on Careless Response in Survey Data*. Unpublished Manuscript. North Carolina State University.

Haertzen, C. A., Hill, H. E., & Belleville, R. E. (1963). Development of the Addiction Research Center Inventory (ARCI): selection of items that are sensitive to the effects of various drugs. *Psychopharmacologia, 4*(3), 155-166. doi:10.1007/BF02584088

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55. doi:10.1080/10705519909540118

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*(1), 99-114. doi:10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828. doi:10.1037/a0038510

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, *39*(1), 103-129. doi:10.1016/j.jrp.2004.09.009

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of applied psychology*, *74*(4), 657. doi:10.1037/0021-9010.74.4.657

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. doi:10.1016/j.jrp.2013.09.008

Meade, A. W., & Craig, S. B. (2012) Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. doi:10.1037/a0028085

Meade, A. W., & Pappalardo, G. (2013). Predicting careless responses and attrition in survey data with personality. In *28th Annual Meeting of the Society for Industrial and Organizational Psychology, Houston, TX*.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2 (1)*, 37-63. dc.identifier.uri: https://hdl.handle.net/2328/27165

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological bulletin*, *140*(6), 1411. doi:10.1037/a0037428

Saari, L. M., & Scherbaum, C. A. (2011). Identified employee surveys: Potential promise, perils, and professional practice guidelines. *Industrial and Organizational psychology*, *4*(4), 435-448. doi: 10.1111/j.1754-9434.2011.01369.x

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods*, *29*(2), 274-279. doi:10.3758/BF03204826

Thompson, A. H. (1975). Random responding and the questionnaire measurement of psychoticism. *Social Behavior and Personality: an international journal*, *3*(2), 111-115. doi:10.2224/sbp.1975.3.2.111

Thompson, L. F., & Surface, E. A. (2007). Employee surveys administered online attitudes toward the medium, nonresponse, and data representativeness. *Organizational Research Methods*, *10*(2), 241-261. doi: 10.1177/1094428106/294696

Thompson, L. F., Surface, E. A., Martin, D. L., & Sanders, M. G. (2003). From paper to pixels: Moving personnel surveys to the Web. *Personnel Psychology*, *56*(1), 197-227. doi:10.1111/j.1744-6570.2003.tb00149.x

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554-568. doi:10.1016/j.chb.2015.01.070

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186. doi: 10.1007/s10862-005-9004-7

Wunder, R. S., Thomas, L. L., & Luo, Z. (2010). Administering assessments and decision-making In Farr, J.L., & Tippins, N. T. (Eds.), *Handbook of employee selection*, 377-398. New York, New York: Routledge.

Yentes, R. D. (2016). Careless: Procedures for Computing Indices of Careless Responding. R Package version 1.0 URL: https://github.com/ryentes/careless

Yentes, R. D., Foster L. L., Meade, A. W., & Pond, S. B. (2017). *Attention and Data Quality in Online Surveys.* Paper presented at the 32nd annual meeting of the Society for Industrial and Organizational Psychology, Orlando, Florida

Table 1

Table 1

List of IPIP Hexaco Facets used in simulating data

Factor	Facet	+ items	- items
Honesty/Humility	Sincerity	1	9
Honesty/Humility	Fairness	5	5
Emotionality	Anxiety	5	5
Emotionality	Dependence	10	0
Extraversion	Liveliness	8	2
Agreeableness	Forgiveness	4	6
Agreeableness	Patience	5	5
Conscientiousness	Perfectionism	8	2
Openness to Experience	Inquisitiveness	6	4
Openness to Experience	Unconventionality	5	5

Note: +/- Refer to the wording of the item with respect to the construct

Table 2

List of Combination methods to attempt for research question 2

Method	Description
Any Flag - LS First	Longstring computed and applied first, then Mahalanobis distance and even/odd are computed and applied. Any flag is sufficient for removal
Any Flag - Simultaneous	All indices computed and applied at the same time. Any flag is sufficient for removal
All Flags	All indices computed and applied at the same time. Respondents are only flagged as careless if they are flagged by all three metrics
LS or Agree - LS first	Longstring computed and applied first, then Mahalanobis distance and even/odd are computed and applied. Respondents are flagged as careless if longstring flags, or if Mahalanobis distance and even/odd both flag them as careless
LS or Agree - Simultaneous	All indices computed and applied at the same time. Responds are flagged as careless if they are flagged by longstring, or if they are flagged by both mahalanobis distance and even/odd.

APPENDICES

Appendix A

Complete source code for this dissertation is located in a private repository on Github. Adam Meade has access, and I will happily grant access to any committee members who wishes to track its evolution and development following the proposal, just send me an email and your preferred email address for github. After I've defended the dissertation, the repository will be made public.