ANNUAL REVIEWS

Annual Review of Psychology

Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices

M.K. Ward[1] and Adam W. Meade[2]

[1] Tacoma, Washington, USA; email: mk@neatquestion.com

[2] Department of Psychology, North Carolina State University, Raleigh, North Carolina, USA; email: awmeade@ncsu.edu

Annu. Rev. Psychol. 2023. 74:577–96

First published as a Review in Advance on August 16, 2022

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-040422-045007

Keywords

careless responding, insufficient effort responding, survey methodology, invalid responding, random responding, lazy respondents

Abstract

Surveys administered online have several benefits, but they are particularly prone to careless responding, which occurs when respondents fail to read item content or give sufficient attention, resulting in raw data that may not accurately reflect respondents' true levels of the constructs being measured. Careless responding can lead to various psychometric issues, potentially impacting any area of psychology that uses self-reported surveys and assessments. This review synthesizes the careless responding literature to provide a comprehensive understanding of careless responding and ways to prevent, identify, report, and clean careless responding from data sets. Further, we include recommendations for different levels of screening for careless responses. Finally, we highlight some of the most promising areas for future work on careless responding.

Contents

INTRODUCTION

Survey data are frequently used to draw research conclusions, make personnel decisions, set policies, assess mental health care, measure quality of life, make criminal forensic evaluations, and more (Anderson & Ferrell 2010, Bassett et al. 2017, Conijn et al. 2019, Schneider et al. 2018). Since 1999, there have been widespread movements in survey research to increasingly collect data online based on the numerous advantages of online surveys such as lower costs, increased accuracy, and the ability to require responses, among other benefits. One drawback of online surveys, however, is that they are particularly prone to careless responding.

Careless responding occurs when respondents fail to read or give sufficient attention to item content, resulting in data that may not accurately reflect respondents' true levels of the constructs being measured (Meade & Craig 2012, Ward & Meade 2018). There are different forms of careless responding such that careless response manifests in survey data in a variety of ways, making it particularly difficult to prevent and challenging to identify.

Careless responses in survey data can lead to a variety of psychometric issues. Careless responding can create spurious within-group variability and lower reliability and can either attenuate or strengthen correlations (Huang et al. 2014). Careless responding can increase risks of type 1 or type 2 errors in hypothesis testing and lower the quality of factor analytic solutions (Huang et al. 2014, Woods 2006). Further, careless responding is widely prevalent (Bowling et al. 2016, Curran 2016, Meade & Craig 2012, Ward & Pond 2015, Ward et al. 2017). Careless responding has the potential to impact all areas of psychology that use self-reported surveys and assessments, particularly when administered online.

Careless responding: survey behavior wherein respondents select a response with little or no regard for the item content

It is important to protect the quality of survey data upon which conclusions, predictions, and real-world decisions are based. Despite both the concern that careless responses are a known threat to the quality of survey data and the wide range of contexts potentially impacted by low-quality survey data, there is no widespread, consistent, and comprehensive understanding of careless responding or of best practices to address it.

Several subdisciplines in psychology have recognized the need to attend to careless responding, with publications in mental health (Conijn et al. 2019), addiction research (Godinho et al. 2016), methods and job satisfaction (Kam & Meyer 2015), personality (Arias et al. 2020), and other areas of psychology that use online surveys to collect data. There has been an enormous amount of research related to careless responding in the last decade, but very few papers have synthesized this literature into a comprehensive list of best practices. Consequently, research about careless responding has become scattered. This review takes stock of the careless responding literature to help a wide audience (beyond psychometricians and research psychologists) understand (*a*) the effects of careless responding, (*b*) the reasons it occurs, and (*c*) best practices to prevent and address careless responding. We build on previous reviews (e.g., Arthur et al. 2021) by conducting a more in-depth analysis focusing exclusively on careless responding. This focus enables a detailed discussion about prevention, which is arguably the preferred way to address careless responding. Finally, we move beyond summarizing indicators of careless responding and provide practical guidance for those working across the range of subdisciplines in psychology.

Impression management: the processes by which people control how others perceive them

Socially desirable responding: survey behavior wherein the respondent evaluates item content and chooses responses unmatched to their true level of the construct to be viewed favorably by others

WHAT IS CARELESS RESPONDING, AND WHY IS IT A PROBLEM?

Careless responding is one of several types of errors wherein an item response does not match the respondent's underlying trait or attitude. While survey responses might not match the underlying level of the latent construct for several reasons, researchers have made a distinction between these errors based on whether participants are responding to the item content or not (Nichols et al. 1989). Careless responding occurs when participants are not basing their response on the item content, and it can occur when a respondent does not read an item, does not understand an item, or is unmotivated to think about what the item is asking. This type of error is different from other errors, such as faking, malingering (see Berry et al. 1992), impression management, and socially desirable responding (Paulhus 1984), in which the respondent reads and evaluates the item content and chooses a response that does not match their true level of the construct measured by the item. Perhaps because careless responding is a concern to a wide variety of researchers across many disciplines, there are many names for the phenomenon, including random response (Beach 1989), protocol invalidity (Johnson 2005), and insufficient effort responding (Huang et al. 2012). We prefer the term careless responding because it is intuitive, is more accurate than random responding, and avoids construct proliferation.

Concerns about careless responding have been around for decades. For example, Haertzen et al. (1963) developed a scale consisting of 30 items containing repeated items or negative versions of otherwise identified items to comprise a validity scale for determining the consistency of responses to a 550-item assessment. Similar scales exist in widely used measures, such as the MMPI-2 Variable Response Inconsistency scale and the True Response Inconsistency scale.

Causes of Careless Responding

Several reasons have been put forth to explain why careless responding occurs, and these reasons fall into three broad categories: characteristics of the survey, the person, and the context.

Characteristics of the survey. Survey design, which includes survey length and instructions, is part of the respondent's survey environment (Berry et al. 1992, Gibson 2016). Toward the end of surveys, participants tend to respond more quickly and with less variation, and they are more likely to self-report careless responding (Berry et al. 1992, Brower 2018, Galesic & Bosnjak 2009).

Therefore, the assumption is that participant motivation will wane over the course of a lengthy survey and this will increase careless responding (Meade & Craig 2012). However, a short survey does not always guarantee less careless responding compared to a longer version. In a student sample, Gibson (2016) found no differences in rates of careless responding in short versus long survey conditions. Some forms of careless responding are influenced by explicitly warning participants in the survey instructions that their responses will be statistically screened for low quality or that rewards for the survey will be withheld if careless responding is detected (Huang et al. 2012, Ward & Pond 2015).

Characteristics of the person. Respondent characteristics that influence careless responding include interest in the survey topic, personality traits, and attitude toward the survey. Respondents with low interest in the survey are more likely to carelessly respond on long (not short) surveys (Brower 2018). Levels of interest can be changed by pressuring people to complete the survey or by enticing potential respondents with high compensation. Increasing pressure to participate (e.g., by making the survey mandatory) may push respondents to be uninterested in the survey topic itself and may consequently increase careless responding (Meade & Craig 2012). Ward & Meade (2018) found that instructions can create cognitive dissonance, resulting in increased interest and reduced careless responding. As expected, findings suggest that low interest corresponds with more careless responding and higher interest with more careful responding.

Those who display careless responding behavior in one task are more likely to display careless behavior on later tasks (Bowling et al. 2016, Camus 2015), suggesting that careless responding partially stems from individual traits. Lower scores on conscientiousness, agreeableness, extraversion, and emotional stability have been associated with more careless responding (Bowling et al. 2016, Dunn et al. 2018, Ward et al. 2017). However, other research has failed to replicate these findings (Camus 2015). Furnham et al. (2015) examined Hogan Personality Inventory traits and found that agreeable, stable, prudent, and ambitious individuals were less likely to respond carelessly. Agreeable individuals who take a survey may be driven by the desire to help to respond accurately and usefully (Dunn et al. 2018). More careless responding tends to occur in individuals who are excitable, cautious, skeptical, and reserved (Furnham et al. 2015) or prone to boredom (Dunn et al. 2018). Barber et al. (2013) found that people who reported less self-control were more likely to respond carelessly, and that the amount of energy the respondent exerted on the survey mediated this relationship.

Context. Aspects of the context in which respondents take a survey may influence careless responding. These include social norms and environmental distractions. In online surveys, where there is less direct interaction between respondents and survey administrators (Johnson 2005), social norms may not be available to signal appropriate survey behavior. Such norms may be activated with regard to the levels of interaction between survey administrators and respondents (Ward & Meade 2018, Ward & Pond 2015). However, the strength and generalizability of this effect appear limited, given that some null findings were reported by studies designed to leverage social norms to reduce careless responding (Ward & Meade, 2018, Ward & Pond 2015).

Environmental distractions are the elements of a respondent's surroundings that may impede their ability to provide careful responses. In online studies, the uncontrolled survey environment may include distractions from other people, media content, or multitasking (Meade & Craig 2012). To date, initial evidence suggests that distractions in a respondent's environment increase careless responding during the survey (Carrier et al. 2009).

Generally speaking, the causes of careless responding may have indirect effects on careless responding, with motivation and ability thought to mediate these relationships—though such indirect effects have not been empirically tested. Characteristics of the survey, person, and context have been discussed as antecedents to motivation. When they lower motivation to respond carefully, careless responding is more likely (Rios et al. 2017). More specifically, there can be low participant interest and other motivational reasons (Schwarz 1999). Additionally, respondents may not feel like it is their responsibility to respond carefully (Ward & Meade 2018), so that a respondent's goal may be survey completion rather than response accuracy. Survey behaviors that reflect motivational issues and result in careless responding are multitasking (Zwarun & Hall 2014) and satisficing (Barge & Gehlbach 2012). Taken together, characteristics of the respondent, the survey context, and the interaction between the two can increase or decrease the motivation and/or ability to respond attentively. This combination, in turn, can lead to different levels and types of careless responding.

Longstring: refers to consecutive identical responses or identical patterns of response to survey items

Response time: the time required by a respondent to answer a single question or complete a questionnaire

Careless Responding Data Patterns

Patterns in survey data can reveal whether careless responding may have occurred. Specifically, there are three major ways that careless responding shows up in survey data: invariability, fast responses, and inconsistency. In this section, we describe each in turn.

Invariability. Previous research has consistently identified at least two distinct types of careless responding, which gives rise to distinct patterns of response options (Kam & Meyer 2015, Maniaci & Rogge 2014, Meade & Craig 2012). The first is called invariability, also known as longstring (Johnson 2005) or straight-lining. This type of careless responding data pattern refers to consecutive identical responses or identical patterns of response (e.g., A, B, A, B, etc.) to survey items. For input types such as dropdown boxes, simply leaving one hand on a keyboard number key and using the other hand to tab to the following response is the fastest way to enter survey responses. If radio buttons are used, choosing the same response to consecutive items is the least distance to travel. Thus, invariable responding is one of the fastest ways to complete a survey.

Invariable responses are relatively easy to identify, as they stand out even with a visual scan of the data. Given the relatively easy identification of longstring responses, this is typically not a strategy used by careless respondents who care about being identified. In other words, if a respondent is completing a survey for pay or credit of some sort and believes that their lack of effort will jeopardize that reward, they are more likely to engage in a different pattern of careless response. Put differently, invariable responders do not seem to care much if the researcher knows they were careless in their responses.

Fast responses. Careless responses, in some cases, may be accompanied by fast times for survey responses. In cases of extremely fast responses, it is nearly impossible for respondents to have read, understood, and responded accurately to the survey items. For this reason, it is possible to set a minimum time threshold to identify respondents that are very unlikely to have responded diligently. Response time tracking can be done at the survey level or at the recommended level of response time per survey page (Bowling et al. 2016). As with invariability, excessively fast responses are relatively easy to identify if the survey is set to record elapsed time. Unless time tracking is explicitly conveyed to respondents, respondents with very fast responding patterns are likely unaware that their response time is being measured and that their rushing can therefore be detected. Although using response time is intuitive and can help detect impossibly fast responses, it fails to detect careless responding in some circumstances. For example, if a respondent is multitasking or distracted, stops for a moment during the survey, or pauses the survey, response time will fail to detect a very fast response. Recorded survey page time can detect extreme swings in response times on each page and can therefore be more sensitive to this than overall survey time.

Inconsistency. A more typical pattern of responses seen with careless responses can be described as inconsistent careless responding. Generally, this type of careless response generates data that do not match patterns that would be expected based on theoretical/logical grounds or trends in the data. For instance, if two items are nearly identical, or correlate very highly, we would logically expect a very similar response on these two items from every participant. Inconsistent careless responders often generate responses that fail to meet an expected level of consistency. Some careless responders may believe their careless responses may be more difficult to identify if they randomly choose a response near the scale midpoint, while others may randomly choose from all possible response options. Careless respondents may follow different strategies when responding, perhaps even within the same survey.

We also note that it is rare to encounter respondents who respond carelessly from the beginning to the end of the survey. While research on the topic is scarce, we believe that respondents are more likely to shift in and out of attentiveness (Baer et al. 1997, Berry et al. 1992, Meade & Craig 2012). For instance, a respondent may begin relatively diligently on a long survey, but they may tire, lose interest, or get distracted and thus finish the survey more carelessly. Indeed, survey length is associated with an increased probability of carelessness (Bowling et al. 2021, Brower 2018, Gibson & Bowling 2020, Ward et al. 2017).

Why Care About Careless Responding?

There are two main reasons to care about careless responding—specifically, that it is widely prevalent and that its presence introduces error. Thus, reliability, correlation, factor analysis, construct validity, hypothesis testing, and/or error estimates can be impacted by careless responding. Existing research has uncovered the variable psychometrics effects of careless responding and established its pervasiveness.

Prevalence. One reason to care about careless responding is that it is probably present in all survey data: It is very common. Any time respondents have incentives to finish the survey under conditions of low effort, careless responding is likely to occur. A further reason to look at careless responding in survey data is that the extent of careless responding varies wildly. Rates of careless responding can range from 1% to 50% (Brühlmann et al. 2020, Curran et al. 2010, Ehlers et al. 2009, Goldammer et al. 2020, Gough & Bradley 1996, Johnson 2005, Meade & Craig 2012, Ward & Meade 2018), and this range may widen as future studies across different contexts and respondents report on careless responding. This wide discrepancy is due in part to the lack of a clear consensus regarding the specific methods that should be used to identify careless responding. While we are not aware of any large meta-analyses on careless responding across all areas of research, a recent review focusing on crowdsourcing data for alcohol research found that the mean prevalence rate of careless responding across 48 studies was 11.7% of the sample (Jones et al. 2022). This number is similar to what both Meade & Craig (2012) and Kurtz & Parrish (2001) found for similar low-stakes conditions. While there is no single best guess as to how much careless responding is likely to be present in any one data set, we know that there is likely to be at least some careless responding.

Empirical effects of careless responding. Perhaps the best way to envision the effects of careless responding is to first understand the basics of psychometric theory upon which surveys rest. Generally, a survey is trying to measure one construct, either by extensive multi-item scales (e.g., personality) or by single-item measures (e.g., opinion polling). In all cases, the goal is to accurately measure the underlying construct with as little error as possible (i.e., to reliably measure the construct). Careless responding can be thought of as a source of error in the measurement process. The nature and extent of this error depend on the type and amount of careless responding. The most common type of careless responding involves a lack of consistency (e.g., random-like responding) that will necessarily introduce random error into the measure, which will have a deleterious effect on some psychometric properties of the data (Thompson 1975). We will focus next on the effects that the introduction of random error via careless responding may have on survey results.

Reduced reliability. One logical effect of introducing random error is the reduced reliability of psychometric measures. Under classical test theory, the reliability of a measure is essentially the extent to which it is free from random error. As a result, when careless responding is nearly random, random error increases and the reliability of measures necessarily decreases. This effect has also been demonstrated empirically (Arias et al. 2020).

Attenuated correlations. Another notable effect of increasing the amount of error present in measures is to attenuate correlations among the constructs measured by the survey. By definition, a random error does not correlate with anything, and as a result, we can expect attenuated correlations among variables when random error is increased in the measures of those variables. As one example, Kam (2019) found lower correlations between self and peer reports in the full sample than in the sample in which careless respondents were removed.

Factor analysis. Because careless responding can attenuate (or at times inflate) correlations among survey variables, this can, in turn, affect factor analytic results. Factor loadings may be attenuated, and factor structure may not be accurately recovered (Arias et al. 2020, Huang et al. 2014, Woods 2006). Both Kam (2019) and Huang et al. (2015) found that factor analysis is more likely to recover the correct number of factors when careless respondents are removed from the sample. Similarly, Arias et al. (2020) found that confirmatory factor analysis model fit was better after careless responders were removed.

Construct validity. Establishing construct validity for a measure is a complex process requiring many types of evidence (e.g., content validity, factor analysis). However, a key component of establishing construct validity is evaluating a construct in relation to other variables via correlations (i.e., convergent and discriminant validity). As careless responses can compromise correlations among variables, construct validity evidence can also be affected by careless responding (Kam 2019, Kam & Meyer 2015).

Hypothesis testing. Given the results summarized thus far, it should be no surprise that careless responding can also affect hypothesis testing. Maniaci & Rogge (2014) conducted a series of studies investigating the effects on hypothesis testing of removing careless responders from a sample. They found a notable increase in the power of regression analyses in the cleaned sample compared to the full sample. We have already detailed how careless responses can affect hypotheses involving correlations, and we add here that reduced reliability and increased error can also impact hypotheses related to means.

Interestingly, the effects on hypothesis testing are less straightforward than the effects of careless responding on point estimates (e.g., means, correlations). The statistical power of hypothesis testing depends in part on the size of the sample involved in the test. Moreover, while removing extreme careless responders can increase power by removing sources of error in the data (Maniaci & Rogge 2014), the effect of removing participants who are only somewhat careless is less clear, as the reduction in sample size also serves to decrease power. For this reason, it is likely to be beneficial to remove from the sample only those respondents who exhibit the highest amount of carelessness. Thus, when different indicators of careless responding are inconsistent in identifying carelessness for a given respondent, it may sometimes be better to leave the respondent in the sample to increase sample size and statistical power. Similarly, it is not recommended to rely on a single indicator of some careless response indices (such as outlier analysis or consistency analysis). In such situations, one can run analyses with and without the careless respondent in question to verify there is no effect on the results.

Bogus item: item describing logical impossibilities that is inserted into a survey to detect agreement and therefore careless responses

Systematic error and inflated correlations. Although the discussion so far has focused on the effects of random error on attenuated correlations, it is also possible for careless responding to introduce systematic errors that have the effect of inflating correlations. Huang et al. (2015) illustrate that careless responding can cause biased estimates in means, leading to spurious correlations between variables in some instances. This may be particularly true for measures of constructs on which most diligent respondents are well above or below the mean on two particular constructs. If diligent respondents are well above the mean, for example, but careless respondents tend to respond closer to the midpoint or below on both constructs, then the careless responses serve as influential outliers inflating the correlation between the two constructs. Thus, it is difficult to unambiguously predict the effects of careless response on a given study.

WHAT TO DO ABOUT CARELESS RESPONDING

There are two primary approaches to handling careless responding: prevention and removal. The best approach prevents careless responses in the first place; however, it is not realistic to prevent all careless responding. Researchers are often working with archival data sources, so it is also necessary to understand the best practices associated with identifying careless respondents for potential removal from the data.

Typology of Careless Responding Identification Indices

Several authors have developed typologies of methods to identify careless responding. The first distinction that can be made is between a priori methods, in which items, scales, timers, and other features are built into the survey data collection beforehand, and post hoc methods that are based on collected survey data (Meade & Craig 2012).

A priori methods. Methods in this category involve considerations made before data collection concerning the survey content, survey features, or some other element.

Instructed response items. Instructed response items are one of the most straightforward methods of detecting careless responses. These items deviate from the rest of the survey content by asking respondents to make a specific response (e.g., "To ensure data quality, please choose strongly disagree for this item"). Typically, an extreme response is requested, as such responses tend to be rare in general and thus less likely to be chosen by a careless responder. An advantage of instructed response items is that there is little ambiguity in scoring. An absence of the instructed response indicates careless responding on the part of the respondent (Meade & Craig 2012). A disadvantage is that such items are relatively easily identifiable, and savvy respondents may look for such items or use computing methods to search and identify these items.

Bogus items. Bogus items are like instructed response items in that they are specific items inserted into the survey to detect careless responses. Such items generally ask about something impossible or extremely improbable (e.g., "I am paid biweekly by leprechauns"; Meade & Craig 2012). An advantage of bogus items over instructed response items is that they tend to blend into the survey better. Participants who complete many surveys for pay on crowdsourcing websites may watch for instructed response items, whereas bogus items tend to be harder to spot with software or visual scan.

A disadvantage is that attentive respondents may resent such items as irrelevant or as a waste of time. Additionally, there is more ambiguity regarding the scoring of such items. For instance, it is not clear whether a "somewhat disagree" response for an impossible item should be scored as careless, so the researcher must decide. Importantly, Curran & Hauser (2019) did a think-aloud study with bogus items and found that sometimes highly diligent respondents would agree with impossible items. For instance, they found that respondents might agree with an item like "All my friends say I would make a great poodle" because they had been told that they were loyal friends and they associated loyalty with dogs in general. Thus, these items may sometimes result in false-positive identification of careless respondents.

Self-report. Self-report items ask the participant the extent to which they were careful and engaged during the survey process, typically accompanied by an assurance that whatever study reward is offered will still be delivered. These could be an individual item (e.g., "Should we use your data?") or a multi-item scale (see Meade & Craig 2012). Such scales or items are relatively straightforward and require minimal data analysis skills. An obvious disadvantage is that respondents may not be honest about their careless responding, particularly if completing the survey for monetary compensation.

Response time. Although response time must be computed after data collection is complete, we include response time as an a priori method because most survey software does not capture response time by default. Thus, consideration must be given before data collection begins. Response time can be nonobvious to respondents, and it is thus more difficult for them to elude detection compared to instructed response options. Importantly, it is recommended that researchers capture response time at the survey page level rather than the survey level. Study level measures may lead to false-negative results if the respondent takes a break or responds more slowly. Page time has led to much better results when using 2 seconds per item as a cutoff criterion to identify careless responding (Bowling et al. 2016).

Post hoc methods. Post hoc methods have several advantages over a priori methods. First, they can be computed for almost all data sets, even archival data. Second, as they are computed from the data sometimes using advanced statistical methods, they make avoiding detection more difficult. Third, they often are intuitive in the way that careless response is assessed. Methods considered here are a diverse group and can be further broken down into categories. We provide an overview of the most commonly used methods here, and Curran (2016) provides more details on the computation of many of the indices discussed in this section.

Invariability. Invariable responses can be detected by computing a longstring index (e.g., maximum number of consecutive identical responses) or other indices that examine invariability, such as the frequency of choosing each response option. Marjanovic et al. (2015) proposed computing the standard deviation of each person's responses to examine the extent to which respondents are excessively consistent. Dunn et al. (2018) have argued that variance of each person's response is preferable as it allows for the detection of some types of invariance based on pattern, whereas longstring only detects cases where the same response is given consecutively. While some software does automate the computation of longstring, for those without access to it within-person standard deviation may be easier to compute.

Outlier analysis. Mahalanobis distance is a multivariate outlier analysis that assesses the extent to which the respondent is an outlier from the rest of the sample. Mahalanobis distance is easy to compute in most software and can be computed for any data set, which is not always the case for careless responding indices. Mahalanobis distance may also be adept at identifying multiple types of careless responses; for instance, even careless responders using scale midpoints to avoid standing out may be outliers for items for which the remainder of the sample provided extreme responses. Finally, Mahalanobis distance uses all survey items, which is not the case with most other methods, so it makes better use of the data available.

Mahalanobis distance: a multivariate outlier analysis that assesses the extent to which the respondent is an outlier in the sample

Semantic synonyms/antonyms: item pairs based on theoretically similar content, where weak correlations between those item pairs indicate careless responding

Consistency indices. Consistency indices make up perhaps the largest group of indicators of careless response, as they are highly intuitive and can be computed for most data sets. Consistency analyses can be further divided into indices where item responses are grouped based on theoretical grounds and indices based on statistical grounds. Several consistency indices make use of a within-person correlation. Essentially, these correlations involve isolating the responses of a single respondent, then computing a correlation coefficient using only those data in which the respondent's item responses make up the rows of data, and the columns are determined based on the index in question. For instance, an index may call for creating several pairs of items based on the similarity of item content. If 10 item pairs are created, responses to one of the items in the pair can be grouped into column A and responses to the other item into column B so that there are 10 rows of data and 2 columns. The assignment of an item in the pair to column A or B is typically arbitrary.

Theoretically grouped consistency indices. These indices are computed by grouping or pairing items based on their content. For instance, items measuring the same trait and keyed in the same way may be paired, and then consistency can be computed as the extent to which responses to this pair of items are similar. This is highly intuitive: If an existing and well-validated scale is used to assess a given construct, we would expect responses to items measuring that scale to correlate highly. In some cases, items measuring the same trait but keyed in different ways may be grouped (for instance, a negatively coded item may be paired with a positively coded item measuring the same trait). An advantage of indices in this category is that unlike many others, they do not assume that most of the sample is responding diligently to the survey. These indices can be used even when most data are careless.

Even-odd consistency (sometimes called personal reliability). Jackson (1976, as cited in Johnson 2005) proposed an index in which established scales are broken into subscales for use as a consistency index. This index relies on the fact that established scales typically have undergone a validation process in which items within a given scale have been shown to correlate highly. This index involves dividing scales in the survey into subscales (e.g., a 10-item scale is divided into two 5-item subscales using an even-odd process to assign items to subscales), and these subscales computed from several scales are used to compute a within-person correlation. The advantage of this index is that by using multi-item subscales, the reliability of each indicator is higher than it is in approaches that use a single-item indicator. A downside is that the approach requires that the survey include several well-established scales as part of the content.

Semantic synonyms/antonyms. Using semantic synonyms/antonyms involves creating item pairs based on theoretically similar content (Kurtz & Parrish 2001). For instance, a multi-item scale might be divided into several item pairs based on the expected similarity of the responses. Once item pairs are formed, a within-person correlation is computed across the item pairs. Semantic antonyms constitute a similar index but involve pairing items with the expected opposite response; one of the two items is then reverse coded, and the within-person correlation is computed (Johnson 2005).

Polytomous Guttman error: error that occurs when a respondent strongly agrees with a more extreme item while agreeing less with a more moderate item

Statistically based consistency indices. Meade & Craig (2012) proposed indices highly similar to the semantic synonyms/antonyms indexes but suggested forming pairs based on the observed sample correlation rather than on theory. The advantage of this approach is that sometimes items on ostensibly different scales can be highly correlated, potentially increasing the number of item pairs (and thus the reliability of the indicator). A second advantage is that the process can be automated via computer code and thus readily implemented with no work on the part of the researchers. One caveat of this approach is that some threshold of minimal correlation must be set to establish two items as a pair. Meade & Craig (2012) used a correlation of 0.6, but this number can be adjusted up or down depending on the observed correlations in the data. For instance, it may be better to use a lower threshold to include more items in the index, but item responses can no longer be expected to be similar below some unknown threshold. Also, this approach requires that most of the sample be composed of diligent responders.

Person-total correlation. Person-total correlation is analogous to the inverse of the item-total correlation commonly used in psychometrics. With person-total correlation, each person's responses to survey items are correlated with the average responses for all other persons in the sample (Curran 2016). The correlation sample size is then the number of items in the analysis.

Polytomous Guttman errors. The concept of a Guttman error arose first in dichotomous tests, in which a Guttman error occurs when an examinee gets a relatively easy item incorrect while getting more difficult items correct. Expanding this to polytomous survey data, a polytomous Guttman error occurs when a respondent has a strong agreement with a more extreme item (based on the overall sample mean response) while having a weaker agreement with a more moderate item (Curran 2016).

Person-fit item response theory models. Person-fit item response theory (IRT) models can be thought of as a more sophisticated version of polytomous Guttman errors. These models, which are quite technical, examine the likelihood of witnessing the observed pattern of responses given what is known about the nature of the data. Thus, like in Guttman errors, poor person fit arises when a respondent agrees with more extreme statements while simultaneously disagreeing with similar or less extreme statements. Several such indices can be computed, and Beck et al. (2019) provide a good overview.

Resampling. Like in several indices in this section, in resampling items or subscales are treated as pairs, and a correlation is computed across the sets of pairs. With these indices, the choice of which item in the pair is assigned to subset A or B to compute the correlation is arbitrary. In indices such as the even-odd consistency, the assignment of items to a subscale is similarly arbitrary (following an even-odd pattern). A more robust estimate can be gained by resampling, that is, repeating the process many times using different items assigned to different subsets and then averaging the results (Curran 2016). An analogy is that any one split-half correlation will be less robust than an estimator that combines many possible split-half correlations to estimate reliability, as any one sampling contains large amounts of sampling error that can be reduced by replication. Although this process is not yet common, it should produce better estimates of carelessness.

Deciding Whether a Respondent Is Careless

The process of judging whether a respondent is careless differs based on the type of indicator in question. An advantage of instructed response items is that the decision is relatively easy: Either the correct response was chosen, or the respondent was careless. However, even on items such as these, a degree of decision making must come into play. For instance, if several instructed response items are embedded in the survey, how many incorrect responses indicate that the respondent should be removed from the data? Similarly, response time recommendations for removing careless respondents are currently 2 seconds per item, timed at the page level (Bowling et al. 2016). As with instructed responses, there are still decisions to be made about the threshold beyond which responses are too fast to retain a respondent. A case could be made for setting a threshold of even a single missed response, but if there are many such indicators and only a single one was missed, the respondent was likely mostly diligent, and the loss of power associated with removal may offset any benefits stemming from the reduction of error associated with removing the respondent.

Most indices require considerably more decision making. For instance, judging carelessness based on longstring can be more complex because the expected number of identical responses can vary depending on the nature of what is measured by the survey. If there is a grouping of highly related items measuring the same or highly correlated constructs, we might expect many identical responses to these items. Conversely, if consecutive items consist of reverse-coded items or measures of independent constructs, many identical consecutive responses may be extremely unlikely. Some researchers have recommended using a specific number of consecutive items (e.g., 6 to 14) that varies depending on the responses, given that extreme responses may be less likely for some constructs (Huang et al. 2012). Longstring, like many other indices, can also be used with an empirically derived cutoff score that can be more closely tailored to the researchers' data.

Similarly, decisions about how many items should be paired in indices such as psychometric synonyms must balance the need to set a high cutoff value (e.g., 0.70) for correlated responses with the need to have enough item pairs to build a reliable indicator. An additional consideration is the implicit assumption of many of these methods that most of the sample is responding diligently. For instance, correlations between similar items will not be very high when most of the sample is careless. As a result, indices like psychometric synonyms/antonyms cannot be computed. Similarly, outlier analysis and person-total correlation are also predicated on the notion that careless responses are rarer than diligent ones.

As of now, there are no clear guidelines regarding how to identify careless respondents through the various indicator variables that are available. Essentially, each researcher must weigh the pros and cons of determining carelessness and of removing careless respondents. For example, if the accurate estimation of means and standard deviations is important and the available sample size is large, a researcher may decide to remove more cases to ensure a cleaner sample. On the other hand, if sample sizes are smaller and the analyses are somewhat more robust (e.g., correlations), then perhaps removing only the most egregious careless responders is warranted.

It is worth noting that beyond some threshold of impossibility, longer response times (or less longstring responding, for that matter) do not imply more diligent responses. Thus, indices of careless responding should not be considered continuous indicators of data quality. Instead, they more closely resemble clinical measures in which a flag is raised when an indicator exceeds some threshold. For instance, a longstring value above a high threshold (e.g., 30) likely indicates a careless response, but a longstring value of 4 does not imply better data quality than a longstring value of 6. Understanding this is important for researchers interested in the study of correlates of careless response, for example. Most indices of careless response should be scored as dichotomous variables, as they rarely have meaning across the continuum of scores.

Removal of Careless Responding

Most suggestions for handling careless responders once they are identified involve removing them from the sample. There are relatively few studies that have tried to establish the best criteria for removing participants from the sample. Meade & Craig (2012) conducted a simulation study to determine the efficacy of different indicators in identifying careless responses. They obtained the most promising results from a mixture model analysis that considered multiple indicators. For single indicators, longstring was effective at identifying invariant responses, whereas Mahalanobis distance and even-odd consistency were among the best for identifying random responses.

Yentes (2020) also conducted a simulation study to evaluate not only some of the careless response indicators but also methods of identifying cut scores on these indicators. Yentes found good support for setting a longstring cutoff score of 0.4 standard deviations above the sample mean. He also found support for setting a Mahalanobis cutoff score of 0.5 standard deviations above the sample mean, resulting in a sizable loss of diligent responders. He found it was difficult to provide a generalizable cut score for even-odd consistency across the study conditions.

Patton et al. (2019) conducted a pair of simulation studies evaluating the efficacy of the person-fit IRT statistics on large scales. They found support for an interactive procedure in which careless respondents were identified and removed from the sample, then indices were estimated again. Goldammer et al. (2020) used a different approach in which they instructed a portion of the sample to be careless using random responses. They found promising results for Mahalanobis distance, psychometric synonyms, psychometric antonyms, and even-odd consistency. They found poor results for longstring, but their sample was not instructed to provide invariant-type careless responses.

Statistical Power and Sample Representativeness

As mentioned earlier, researchers should not assume that strict enforcement of careless response screening indices is always the best option. In many cases, by doing so the available sample size may be decreased dramatically, by 50% or more. Instead, it may be preferable in some situations to retain data in which respondents appear to be only occasionally careless in order to preserve sample size.

Also, as some researchers have pointed out, a strict screening runs the risk of removing a portion of the sample that systematically varies from the remaining portion of personality characteristics and perhaps other unknown variables (Bowling et al. 2016, Ward et al. 2017). However, if the data are truly careless, little will be gained by retaining those responses in the data set.

Recent Innovations

As this burgeoning literature evolves, recent studies have explored innovative ways of identifying and preventing careless responding. In this section, we discuss innovations at the cutting edge of careless responding research.

Eye tracking. Recently, Brower (2020) used eye tracking in a laboratory experiment to better understand the nature of careless responding. While this approach would be impractical in most circumstances, the study did corroborate that respondents vary considerably in their approaches to reducing the survey burden.

Prevention of careless responding. Although it may not be possible to prevent all careless responses, there is general agreement that prevention is preferable to the removal of careless participants (Bowling et al. 2016, McGonagle et al. 2015). Empirical investigations into preventing careless responding have centered on survey instructions, levels of proctoring, and—to a lesser extent—tweaks to survey design.

Survey instructions. Instructions that make the consequences of careless responding salient to respondents can prevent some forms of careless responding but risk an unpleasant survey experience. Ward & Meade (2018) improved response consistency, accuracy, and reported interest in the survey by asking respondents to sign their initials next to statements indicating they understood what it meant to carefully complete the survey. Inconsistency and inaccuracy can be reduced by using survey instructions that warn respondents or make them feel hypocritical about careless responding (e.g., Gibson & Bowling 2020, Huang et al. 2012, Meade & Craig 2012, Rauti 2017). Compared to warnings, using more positive approaches, such as offering rewards in exchange for careful responding (Gibson & Bowling 2020, Ward & Meade 2018) or increasing the social influence of survey administrators on respondents (Ward & Meade 2018), has been less effective.

Levels of proctoring. Proctoring and visual indicators of proctoring appear to effectively prevent some forms of careless responding. Francavilla et al. (2019) found that different levels of proctoring (i.e., remote online un-proctored, remote online virtually proctored, and in-person classroom proctored) impacted specific indicators of careless responding but had no effect on a composite score of all flags across careless responding indices. Compared to un-proctored surveys, virtually proctored surveys showed less careless responding on bogus items, and in-person proctoring showed less careless responding on bogus items and higher self-reported diligence on the survey. The relationship between in-person proctoring and self-reported diligence was fully mediated by environmental distractions, suggesting that environmental distractions need to be addressed to win the attention of survey respondents.

Survey design features. Two survey design features that have been tested for their impact on careless responding are visual elements of the survey webpages and interactivity of the survey messages. Ward & Pond (2015) reduced careless responding by displaying a virtual human to respondents. Although this was not empirically tested, it may be that the virtual human seemed like it was virtually proctoring because it appeared to pay attention to what the respondent was doing, and instructions warned respondents that answer quality would be monitored. Gibson (2019) tested four types of interactive warnings: a no warning group, a noninteractive warning, an interactive threatening message, and an interactive encouraging message. There was no effect of interactivity of survey messages on careless responding.

Overall, evidence suggests that the most effective ways to prevent careless responding involve ensuring that survey respondents believe that carelessness is being monitored and that they will experience consequences based on the quality of data they provide. Although there is ample space to improve prevention strategies, this budding area of research confirms that it is possible to inhibit careless responding before it occurs.

RECOMMENDATIONS

Evidence of the problematic effects of careless responding tends to spark two questions from scientists and practitioners across subdisciplines of psychology: Do I need to do anything about careless responding in my project, and if so, what can I do? To answer the first question, researchers should look for risk factors that suggest careless responding could reasonably occur. To answer the second question, they should consider constraints of the survey, sample, and context.

Table 1 Recommendations for three levels of screening for careless responding

Screening level	A priori screening	Post hoc screening
Minimal	■ Instructed response items ■ Page-level response time	■ Invariance analysis (longstring index or within-person variance) ■ Multivariate outlier analysis (such as Mahalanobis distance)
Moderate	■ Instructed response items ■ Page-level response time ■ Optional bogus items	■ Invariance analysis (longstring or within-person variance) ■ Multivariate outlier analysis ■ Two consistency indicators (e.g., psychometric synonyms, even-odd index)
Extensive	■ Instructed response items ■ Page-level response time ■ Bogus items	■ Invariance analysis (longstring or within-person variance) ■ Multivariate outlier analysis ■ Three consistency indicators (e.g., psychometric synonyms, even-odd index) ■ Person-fit indices ■ Resampling statistics ■ Use of sequential removal (i.e., removing invariant responders prior to computing other indices) ■ Latent class analysis or mixture model

When Careless Responding Is Likely to Be a Problem

Ideal conditions for careless responding would be a long survey comprised of over 100 repetitive items, administered online, about a topic the respondent finds uninteresting and/or irrelevant, and with no apparent consequences for careless responding. Respondents in the sample would be prone to boredom, extraverted, and low on agreeableness, emotional stability, and self-control. Further, completing the survey would be mandatory. If any of these characteristics of the survey, respondent, or context are present, then we strongly urge the use of at least minimal screening for careless responding and we advise using moderate or extensive screening.

How to Address Careless Responding

We recommend that both prevention and identification be used wherever possible and appropriate. To date, research supports multiple ways to prevent careless responding, including setting clear expectations for careful responding, explaining the consequences of careless responding, asking respondents to explicitly commit to being careful, and showing respondents that the quality of their responding will be actively monitored (i.e., proctor the survey where possible) (Meade & Craig 2012, Ward & Meade 2018, Ward & Pond 2015). Rather than using one way to prevent careless responding, it is generally advisable to use a combination of prevention techniques without making the survey experience excessively punitive.

Building on past research, we provide recommendations for data screening at three levels (minimal, moderate, and extensive; see **Table 1**), matched to researchers with variable levels of motivation and ability and to situations in which the researcher can design elements into the survey beforehand. In some cases, research novices will be limited in their ability to manipulate the data and use more advanced software. In other cases, archival data will prevent the ability to design the survey.

Minimal Screening

Generally speaking, we believe most researchers should be engaging in at least moderate-level screening for careless responses. However, we realize there are times when the researcher may be a novice to research methods or may not have access to, or the ability to use, software that allows for more thorough screening. In these cases, we recommend that researchers conduct no less than minimal screening for careless responses prior to further data analysis. We include in the table an "a priori screening" column relevant only to situations in which the researcher has control over the design process. These options are not meant to preclude the use of post hoc methods, but they are presented separately because archival data sets do not allow for the a priori methods to be used. In other words, whenever possible, researchers should use both a priori and post hoc methods to screen data.

If a priori options are possible, we strongly recommend the use of instructed response items and page-level response time tracking. Instructed response items are easy to include in survey design, and even novices can screen data based on instructed responses. Similarly, page time tracking is a feature of several commercially available software tools, though computation of elapsed time between pages can be slightly more difficult. These two tools will provide a minimal screening of the most egregiously careless respondents with relatively little effort. The downside is that these two approaches are unlikely to identify all careless respondents, especially those who are savvy (e.g., those who complete dozens of surveys for compensation). We also recommend screening based on outliers and invariant responders, as mentioned below.

Post hoc options are available for archival data and for more thorough screening of data even when survey design is possible. At a minimum, we urge screening based on outliers and invariant responses. We recommend outlier analysis (such as the Mahalanobis distance) because it is built into most software and is already typically part of the data analysis process. Mahalanobis distance also has a built-in chi-square statistic providing a ready-made cutoff score for decision making. Thus, even novice researchers can conduct this analysis. If a researcher is able, we also recommend the longstring index, or within-person variance if it is easier for the researcher to compute, to screen for invariable responding. Invariable responders are typically a small percentage of careless responders but can have the most biasing effect on the data.

Moderate Screening

We start by recommending the a priori approaches described in our minimal screening section—the use of instructed response items and page-level response time. If the researcher expects to collect data with savvy survey takers (i.e., those who regularly complete surveys for pay), we recommend adding bogus items, as savvy careless respondents often watch for and identify instructed response items. It is important to design bogus items so they are not amusing, as this has tempted student samples to agree with items as a way to be funny (Meade & Craig 2012). The researcher will need to decide which responses are acceptable for bogus items, and post hoc analyses will then identify nonsensical responses to bogus items.

In addition to the post hoc methods previously recommended (invariance analysis and multivariate outlier analysis), we strongly recommend computing at least two measures of consistent response. Previous research has found good results with the even-odd index (Goldammer et al. 2020, Meade & Craig 2012), though that index requires the inclusion of several multi-item constructs using well-validated scales. We also recommend psychometric synonyms, as empirical relationships may identify more potential item pairs than other approaches.

Extensive Screening

Our recommendations for a priori options are the same as for moderate-level screening. We believe that the most thorough screening will fully utilize post hoc methods.

We do not believe most researchers need to engage in extensive screening of their data for careless responding. However, for conditions in which this might be desirable (i.e., very large samples where respondents may be motivated to respond with minimal effort), we would advise several additional steps post hoc. First, we recommend including a few more indicators of careless responding. For instance, additional consistency indices, including person-fit IRT statistics, can be useful additional indicators.

Perhaps more importantly, we recommend doing more with the available indicators. For instance, it is advantageous to use a sequential process in which invariant responders are removed prior to computing correlations to identify psychometric synonyms, outliers, and other indices of careless responding. In their simulation study, Patton et al. (2019) found promising results using an iterative process to identify and remove careless responders using a person-fit measure. Similarly, using resampling for many careless response indicators can provide a more accurate and robust estimate of the extent to which a respondent is careless than a single computation.

Lastly, we believe that using more advanced modeling can provide a more comprehensive view of careless responding. For instance, latent class analysis or mixture modeling using the careless response indices as indicators can provide guidance as to the likelihood of different types of careless responding in the data set (Meade & Craig 2012, Yentes 2020). Although development is still ongoing, the R package Careless (Yentes & Wilhelm 2018) provides many tools needed to automate the computation of careless response indices.

Reporting

We strongly encourage future studies to report how careless responding was addressed. Screening for careless responding is an extension of more conventional data cleaning, and we recommend writing up what was done to prevent, identify, and remove a careless response or respondent. This can be included in the methods section or supplementary materials. Finally, if preserving statistical power is critical, then we encourage researchers to run their analyses with and without careless responding. This is to determine its effects on the data, report any meaningful differences in the results, and justify any reluctance to remove respondents flagged for careless responding.

CONCLUSION

Across psychology's subdisciplines, a growing number of studies are screening data for careless responding or researching ways to prevent, identify, and/or remove careless responding. Survey instructions and design can impact careless responding with variable effects. From such studies, it is clear that nearly any sample involving students or professional survey-takers (e.g., MTurk, Prolific, Qualtrics samples) would benefit from screening data for careless responses. Consequently, there is growing recognition of the importance of this topic across domains.

Careless responding is consistently found across survey data sets, manifesting in the data in different ways and at variable rates. Prudent researchers gauge the risks of careless responding given the characteristics of the respondents, the survey design, and the context or survey environment. After risks are considered, an appropriate approach to address careless responding can be selected. Recent research has made substantial progress on scoring careless responding indicators, and clues to prevention have appeared. Further insights can be gleaned both from studies focused on understanding careless responding and from studies that use survey data and describe prevention and screening strategies. Researchers, regardless of their subdiscipline, can accelerate the development of this exciting area by reporting ways they address careless responding in their survey data. To that end, our hope is that this review will inspire and enable survey administrators and researchers to address careless responding and propel this exciting research area into the future.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Anderson A, Ferrell W. 2010. Assessment of qualifications needed by environmental health graduates entering private-sector employment. *J. Environ. Health* 72:14–20

Arias V, Garrido L, Jenaro C, Martínez-Molina Arias B. 2020. A little garbage in, lots of garbage out: assessing the impact of careless responding in personality survey data. *Behav. Res. Methods* 52(6):2489–505

Arthur W, Hagen E, George F. 2021. The lazy or dishonest respondent: detection and prevention. *Annu. Rev. Organ. Psychol. Organ. Behav.* 8:105–37

Baer R, Ballenger J, Berry D, Wetter M. 1997. Detection of random responding on the MMPI–A. *J. Pers. Assess.* 68(1):139–51

Barber LK, Barnes CM, Carlson KD. 2013. Random and systematic error effects of insomnia on survey behavior. *Organ. Res. Methods* 16(4):616–49

Barge S, Gehlbach H. 2012. Using the theory of satisficing to evaluate the quality of survey data. *Res. High. Educ.* 53(2):182–200

Bassett J, Cleveland A, Acorn D, Nix M, Snyder T. 2017. Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assess. Eval. High. Educ.* 42(3):431–42

Beach D. 1989. Identifying the random responder. *J. Psychol.* 123(1):101–3

Beck MF, Albano AD, Smith WM. 2019. Person-fit as an index of inattentive responding: a comparison of methods using polytomous survey data. *Appl. Psychol. Meas.* 43(5):374–87

Berry D, Wetter M, Baer R, Larsen L, Clark C, Monroe K. 1992. MMPI-2 random responding indices: validation using a self-report methodology. *Psychol. Assess.* 4(3):340–45

Bowling NA, Gibson AM, Houpt JW, Brower CK. 2021. Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organ. Res. Methods* 24(4):718–38

Bowling NA, Huang J, Bragg C, Khazon S, Liu M, Blackmore C. 2016. Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *J. Pers. Soc. Psychol.* 111(2):218–29

Brower C. 2018. *Too long and too boring: the effects of survey length and interest on careless responding*. PhD Thesis, Wright State Univ., Dayton, OH

Brower C. 2020. *What are you looking at? Using eye-tracking to provide insight into careless responding*. PhD Thesis, Wright State Univ., Dayton, OH

Brühlmann F, Petralito S, Aeschbach L, Opwis K. 2020. The quality of data collected online: an investigation of careless responding in a crowdsourced sample. *Psychol. Methods* 2:100022

Camus K. 2015. *Once careless, always careless? Temporal and situational stability of insufficient effort responding (IER)*. PhD Thesis, Wright State Univ., Dayton, OH

Carrier L, Cheever N, Rosen L, Benitez S, Chang J. 2009. Multitasking across generations: multitasking choices and difficulty ratings in three generations of Americans. *Comput. Hum. Behav.* 25(2):483–89

Conijn J, Franz G, Emons W, de Beurs E, Carlier I. 2019. The assessment and impact of careless responding in routine outcome monitoring within mental health care. *Multivar. Behav. Res.* 54(4):593–611

Curran PG. 2016. Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* 66:4–19

Curran PG, Kotrba L, Denison D. 2010. *Careless responding in surveys: applying traditional techniques to organizational settings*. Paper presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA

Curran PG, Hauser K. 2019. I'm paid biweekly, just not by leprechauns: evaluating valid-but-incorrect response rates to attention check items. *J. Res. Pers.* 82:103849

Dunn A, Heggestad E, Shanock L, Theilgard N. 2018. Intra-individual response variability as an indicator of insufficient effort responding: comparison to other indicators and relationships with individual differences. *J. Bus. Psychol.* 33(1):105–21

Ehlers C, Greene-Shortridge TM, Weekley JA, Zajack MD. 2009. *The exploration of statistical methods in detecting random responding*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA

Francavilla N, Meade A, Young A. 2019. Social interaction and internet-based surveys: examining the effects of virtual and in-person proctors on careless response. *Appl. Psychol.* 68(2):223–49

Furnham A, Hyde G, Trickey G. 2015. Personality and value correlates of careless and erratic questionnaire responses. *Pers. Individ. Differ.* 80:64–67

Galesic M, Bosnjak M. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opin. Q.* 73(2):349–60

Gibson A. 2016. *Stop with the questions already! The effects of questionnaire length and monetary incentives on insufficient effort responding*. Master's Thesis, Wright State Univ., Dayton, OH

Gibson A. 2019. *Stop what you're doing, right now! Effects of interactive messages on careless responding*. PhD Thesis, Wright State Univ., Dayton, OH

Gibson A, Bowling N. 2020. The effects of questionnaire length and behavioral consequences on careless responding. *Eur. J. Psychol. Assess.* 36(2):410–20

Godinho A, Kushnir V, Cunningham J. 2016. Unfaithful findings: identifying careless responding in addictions research. *Addiction* 111(6):955–56

Goldammer P, Annen H, Stöckli P, Jonas K. 2020. Careless responding in questionnaire measures: detection, impact, and remedies. *Leadersh. Q.* 31(4):101384

Gough H, Bradley P. 1996. *California Psychological Inventory*. Palo Alto, CA: Consult. Psychol. Press

Haertzen C, Hill H, Belleville R. 1963. Development of the addiction research center inventory (ARCI): selection of items that are sensitive to the effects of various drugs. *Psychopharmacologia* 4(3):155–66

Huang JL, Bowling N, Liu M, Li Y. 2014. Detecting insufficient effort responding with an infrequency scale: evaluating validity and participant reactions. *J. Bus. Psychol.* 30(2):299–311

Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. 2012. Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* 27(1):99–114

Huang JL, Liu M, Bowling N. 2015. Insufficient effort responding: examining an insidious confound in survey data. *J. Appl. Psychol.* 100(3):828–45

Jackson DN. 1976. *The appraisal of personal reliability*. Paper presented at the Meeting of the Society of Multivariate Experimental Psychology, University Park, PA

Johnson J. 2005. Ascertaining the validity of individual protocols from Web-based personality inventories. *J. Res. Pers.* 39(1):103–29

Jones A, Earnest J, Adam M, Clarke R, Yates J, Pennington C. 2022. Careless responding in crowdsourced alcohol research: a systematic review and meta-analysis of practices and prevalence. *Exp. Clin. Psychopharmacol.* 30(4):381–99

Kam CCS. 2019. Careless responding threatens factorial analytic results and construct validity of personality measure. *Front. Psychol.* 10:1258

Kam CCS, Meyer J. 2015. How careless responding and acquiescence response bias can influence construct dimensionality. *Organ. Res. Methods* 18(3):512–41

Kurtz J, Parrish C. 2001. Semantic response consistency and protocol validity in structured personality assessment: the case of the NEO-PI-R. *J. Pers. Assess.* 76(2):315–32

Maniaci M, Rogge R. 2014. Caring about carelessness: participant inattention and its effects on research. *J. Res. Pers.* 48:61–83

Marjanovic Z, Holden R, Struthers W, Cribbie R, Greenglass E. 2015. The inter-item standard deviation (ISD): an index that discriminates between conscientious and random responders. *Pers. Individ.* 84:79–83

McGonagle A, Fisher G, Barnes-Farrell J, Grosch J. 2015. Individual and work factors related to perceived work ability and labor force outcomes. *J. Appl. Psychol.* 100(2):376–98

Meade A, Craig S. 2012. Identifying careless responses in survey data. *Psychol. Methods* 17(3):437–55

Nichols D, Greene R, Schmolck P. 1989. Criteria for assessing inconsistent patterns of item endorsement on the MMPI: rationale, development, and empirical trials. *J. Clin. Psychol.* 45(2):239–50

Patton J, Cheng Y, Hong M, Diao Q. 2019. Detection and treatment of careless responses to improve item parameter estimation. *J. Educ. Behav. Stat.* 44(3):309–41

Paulhus D. 1984. Two-component models of socially desirable responding. *J. Pers. Soc. Psychol.* 46(3):598–609

Rauti C. 2017. *Assessing the effects of survey instructions and physical attractiveness on careless responding in online surveys*. Master's Thesis, Univ. Windsor, Windsor, Can.

Rios J, Guo H, Mao Liu O. 2017. Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not? *Int. J. Test.* 17(1):74–104

Schneider S, May M, Stone A. 2018. Careless responding in internet-based quality of life assessments. *Qual. Life. Res.* 27(4):1077–88

Schwarz N. 1999. Self-reports: how the questions shape the answers. *Am. Psychol.* 54(2):93–105

Thompson A. 1975. Random responding and the questionnaire measurement of psychoticism. *Soc. Behav. Pers.* 3(2):111–15

Ward MK, Meade A. 2018. Applying social psychology to prevent careless responding during online surveys. *Appl. Psychol.* 67(2):231–63

Ward MK, Meade A, Allred C, Pappalardo G, Stoughton J. 2017. Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Comput. Hum. Behav.* 76:417–30

Ward MK, Pond S. 2015. Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Comput. Hum. Behav.* 48:554–68

Woods C. 2006. Careless responding to reverse-worded items: implications for confirmatory factor analysis. *J. Psychopathol. Behav. Assess.* 28(3):186–91

Yentes RD. 2020. *In search of best practices for the identification and removal of careless responders*. PhD Thesis, N.C. State Univ., Raleigh

Yentes RD, Wilhelm F. 2018. Careless: procedures for computing indices of careless responding. *R Package Version 1.1.3*. **https://cran.rproject.org/web/packages/careless/index.html**

Zwarun L, Hall A. 2014. What's going on? Age, distraction, and multitasking during online survey taking. *Comput. Hum. Behav.* 41:236–44