J Bus Psychol (2018) 33:559–577
DOI 10.1007/s10869-017-9514-9

ORIGINAL PAPER

Dirty Data: The Effects of Screening Respondents Who Provide Low-Quality Data in Survey Research

Justin A. DeSimone[1] · P. D. Harms[1]

Published online: 2 September 2017

Abstract The purpose of this study is to empirically address questions pertaining to the effects of data screening practices in survey research. This study addresses questions about the impact of screening techniques on data and statistical analyses. It also serves an initial attempt to estimate descriptive statistics and graphically display the distributions of popular screening techniques. Data were obtained from an online sample who completed demographic items and measures of character strengths (N = 307). Screening indices demonstrate minimal overlap and differ in the number of participants flagged. Existing cutoff scores for most screening techniques seem appropriate, but cutoff values for consistency-based indices may be too liberal. Screens differ in the extent to which they impact survey results. The use of screening techniques can impact inter-item correlations, inter-scale correlations, reliability estimates, and statistical results. While data screening can improve the quality and trustworthiness of data, screening techniques are not interchangeable. Researchers and practitioners should be aware of the differences between data screening techniques and apply appropriate screens for their survey characteristics and study design. Low-impact direct and unobtrusive screens such as self-report indicators, bogus items, instructed items, longstring, individual response variability, and response time are relatively simple to administer and analyze. The fact that data screening can influence the statistical results of a study demonstrates that low-quality data can distort hypothesis testing in organizational research and practice. We recommend analyzing results both before and after screens have been applied.

Keywords Data screening · Survey research · Research methods · Data analysis · Research design · Insufficient effort responding

Dr. Smith is interested in researching character strengths. She decides to design her study using a series of self-report questionnaires. For reasons of convenience, Dr. Smith enrolls anyone she can in the study. Four of her participants are Kyle, Jane, Kate, and Henry.

Kyle is a college sophomore with better things to do than fill out a survey. In fact, Kyle's only motivation for participation is to get extra credit for his psychology course. Kyle does not even bother to read the questionnaire content, opting instead to select the "middle" option for each item. After about 2 min of filling in bubbles, Kyle hands in his completed survey and walks away with his extra credit.

Jane is a recent graduate who took Dr. Smith's course when she was in college. She is incredibly busy but decides to participate as a favor to her former professor. Unfortunately, she is late for a meeting and does not have time to complete the questionnaire as thoroughly as she would like. Jane still wants Dr. Smith to think highly of her, so she knows she cannot leave questions blank or respond the same way to each item. Jane's compromise is to randomly select responses to make it appear as if she were responding thoughtfully.

Kate is majoring in psychology and fascinated with all of the fun tests she gets to take when she participates in psychological research. Kate's intellectual curiosity prompts her to "figure out" each test as quickly as she can. Kate is quite proud of her ability to determine each test's purpose within the first few items. After that, it is simply a matter of selecting the

✉ Justin A. DeSimone
jadesimone@cba.ua.edu

[1] Department of Management, University of Alabama, 361 Stadium Drive, Tuscaloosa, AL 35487-0025, USA

responses that will make her appear to possess high levels of all of the character strengths examined in Dr. Smith's study.

Henry is an incredibly paranoid student who believes that Dr. Smith secretly works for a nefarious government intelligence agency. Henry is convinced that this study represents an attempt by the government to profile all university students across the country. But Henry is smarter than the government officials. He knows that if he can make himself appear inconsistent, the government will not be able to use his information for their reprehensible purposes. Henry pays close attention to the content of each item to ensure that he selects different responses to similar questions.

Kyle and Jane respond without considering the content of the items. Kate and Henry respond in accordance with their beliefs about how the questions will be scored and used. All four participants seem unconcerned with providing responses that accurately reflect their self-perceptions of character strengths. Regardless of whether the participants are engaging in content nonresponsivity or content-responsive faking (see, Nichols, Greene, & Schmolck, 1989), the responses provided by each participant reflect constructs that are not relevant to the research Dr. Smith is attempting to conduct (see, Gallen & Berry, 1996; Nichols et al., 1989; Paolo & Ryan, 1992). Additionally, each of the four participants may be responding in a manner consistent with a different irrelevant construct.

Should Dr. Smith use the data provided by Kyle, Jane, Kate, and Henry? If Dr. Smith were to eliminate one or more of these participants from his study, on what grounds should she make this decision? What statistical techniques are available for the identification of low-quality data? Would the elimination of inattentive or purposively deceptive participants affect the performance of the questionnaire? The present paper addresses these (and other) questions using data from an online sample.

Although the above examples may be extreme, variations on these themes are common in research. Self-report surveys are common in research on individual differences (Clark & Watson, 1995; Schwarz, 1999). Participants in survey research may exhibit a variety of response styles (Couch & Keniston, 1960; Cronbach, 1950). Some of these styles involve exerting little effort when responding to questionnaires (Kurtz & Parrish, 2001; Meade & Craig, 2012). Respondents like Kyle and Jane exhibit insufficient effort responding (Liu, Bowling, Huang, & Kent, 2013). Specifically, Kyle has an invariant response style and Jane has a random response style.

Kate and Henry's responses cannot be labeled as "insufficient effort." Indeed, both of these participants put much thought and effort into their responses. Nevertheless, their data may be problematic in that these data do not reflect their honest self-perception of their standing on the constructs being measured. Kate is "faking good" or exhibiting socially desirable responding (Edwards, 1957). Henry, on the other hand, is providing duplicitous or disingenuous responses. Even though Kate and Henry are exerting effort, they are not providing Dr. Smith with meaningful data.

Throughout this paper, we will use the term "low-quality data" (LQD) to describe responses that fall into these categories. LQD may take many forms in survey data, including insufficient effort (e.g., random or invariant) or deceptive (faking good or intentionally dishonest) responses. We make no claim that all forms of LQD are equally egregious or harmful to research. However, we assume that all forms of LQD are less desirable and more harmful than honest, thoughtful, and effortful responses.

Investigators have been aware of variations in data quality for more than half of a century, but only a handful of articles directly compare the performance of screening procedures (e.g., Berry et al., 1991; Clark, Gironda, & Young, 2003; Huang, Curran, Keeney, Poposki, & DeShon, 2012; Paolo & Ryan, 1992). Early studies focused nearly exclusively on the development and evaluation of linear composite data screens used to detect random responding on the Minnesota Multiphasic Personality Inventory, a measure consisting of hundreds of dichotomous items used to detect clinical pathologies. Recent work has focused on comparing or categorizing screening techniques for use in survey data (Bowling et al., 2016; Credé, 2010; DeSimone, Harms, & DeSimone, 2015; Huang et al., 2012; Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014; Meade & Craig, 2012).

Despite its long research history and recent renewed interest, many questions about data screening remain unanswered. For example, researchers disagree about the prevalence of LQD, which screening techniques are most effective, and the impact of screening on the results of statistical analyses. The purpose of this paper is to identify and empirically address research questions pertaining to data screening practices. Additionally, we examine distributional characteristics of various screening indices and provide recommendations for survey design and data analysis.

Methods for Detecting LQD

There are many methods available with the potential to identify various forms of LQD (see, Curran, 2016; DeSimone et al., 2015). Some of these methods involve the direct assessment of response quality, others involve unobtrusive observation of respondent behavior patterns, and others require the calculation of statistical indicators.

Direct Methods of Detecting LQD Direct assessment of response quality involves the insertion of items into a questionnaire to determine whether or not respondents are exerting sufficient effort. There are three types of items that can be included in a survey to directly evaluate data quality, including self-reported effort items (e.g., "I carefully considered each item before responding"; Berry et al., 1992; Costa & McCrae, 1997), "bogus items" (e.g., "I was born on February 30"; Bagby, Gillis, & Rogers, 1991; Huang, Bowling, Liu, & Li, 2015), and "instructed items (e.g., "Please mark 'slightly agree' for this item"). Participants are flagged as potentially providing LQD if they indicate their responses are untrustworthy, provide illogical responses to bogus items, or fail to follow instructions in instructed items.

The major strength of direct data screening techniques lies in their face validity. It is difficult to rationalize retaining a participant's data when that participant admits to exerting minimal effort, when it is obvious that (s)he was not paying attention, or when (s)he fails to comply with instructions. The primary weakness of these methods is their transparency. There is ample evidence to suggest that motivated participants are both able and willing to "fake good" on transparent self-report measures (Ellingson, Sackett, & Hough, 1999; Leary & Kowalski, 1990; Rosse, Stecher, Miller, & Levin, 1998; Snell, Sydell, & Lueke, 1999; Zickar & Robie, 1999). Participants who "fake good" or respond in a manner consistent with social desirability or demand characteristics will likely also be willing to respond to self-reported effort items in similar ways.

Bogus and instructed items are relatively straightforward. Attentive participants will be able to determine the purpose of these items and provide the "correct" responses. Bogus and instructed items are useful for identifying inattentive participants, but may be less suitable for identifying respondents who intentionally distort their responses. However, if survey respondents know one another (or communicate online), it is possible that some participants may be "warned" about the presence and location of these items in a survey (Chandler, Mueller, & Paolacci, 2014).

Unobtrusive Methods of Detecting LQD Unobtrusive approaches involve evaluations of participant behavior during the administration of the survey. Unobtrusive screens do not require modifying the survey, are typically undetectable to respondents, and are relatively simple to compute. The three major unobtrusive methods involve recording response time (Behrend, Sharek, Meade, & Wiebe, 2011; Berry et al., 1992), the number of consecutive identical responses provided by the respondent ("longstring"; Behrend et al., 2011; Huang et al., 2012; Meade & Craig, 2012), and individual response variability (IRV; Dunn, Heggestad, Shanock, & Nels, in press). Response time can be calculated on an entire questionnaire or page-by-page, with the latter being more useful when attempting to identify sporadic or local random responding. The longstring index can be calculated separately for each response option (see, Costa & McCrae, 2008) and is amenable to use with various response formats when a set of consecutive items share a common response format (e.g., Likert, true/false, forced choice). IRV is conceptually related to the longstring index and is defined as the standard deviation of a participant's responses to all items on a questionnaire (Dunn et al., in press). Participants who respond to items too quickly, exceed a predetermined number of consecutive identical responses, or exhibit low response variability are flagged as potential LQD.

The major complication in implementing a response time screen involves deciding what cutoff to use. There is little data available to identify norms for response time, as these norms will vary based on factors such as item length, participant reading speed, and response format. Absent such norms, researchers may opt to flag respondents based on a logical cutoff point (e.g., faster than 2 s/item; Huang et al., 2012) or proportions of LQD estimated in the literature (Dunn et al., in press; Johnson, 2005; Meade & Craig, 2012).

Like the response time index, the long string and IRV indices do not have a well-defined cutoff. Although Costa and McCrae (2008) provide scale-specific cutoffs for each response option, the theoretical rationale for why it should require more invariant "agree" responses than "strongly disagree" responses to indicate LQD is not well justified. Dunn et al. (in press) did not specify a cutoff value for IRV but note that the screening technique works best when calculated on a set of 25 to 150 items. The range of possible IRV values will vary as a function of the number of items on a scale and the number of response options available to respondents for each item. The cutoff chosen should likely be dependent on the nature of the scale. For example, the cutoff value for a uniformly positively worded scale measuring a single homogenous construct should likely be higher than the cutoff score for a multidimensional scale or one containing both positively and negatively worded items (DeSimone et al., 2015).

Statistical Methods of Detecting LQD Statistical methods require the calculation of indices based on individual response patterns. Although statistical techniques are also unobtrusive, they are considered a separate category due to the computational effort involved. While response time, longstring, and IRV can be calculated in a single step (i.e., by calculating a count or standard deviation), the computation of statistical screens relies on more advanced statistics and multiple steps. Statistical screens also require the analyst to make decisions such as how to identify conceptually similar items, how to divide a test, or which norm(s) to use.

Outliers are commonly flagged as potential LQD (Aguinis, Gottfredson, & Joo, 2013; Anscombe & Guttman, 1960). One outlier identification method appropriate for use with survey data is the Mahalanobis distance (*D*; Mahalanobis, 1936), a measure of the multivariate distance between an individual's response vector and the average response vector for all participants who took the questionnaire (Meade & Craig, 2012; Stevens, 1984). Respondents who are furthest from the average response vector are flagged as potential LQD.

Other statistical methods typically involve attempts to quantify the consistency with which respondents answer survey items. Examples of these methods include psychometric synonyms or antonyms (Bruehl, Lofland, Sherman, & Carlson, 1998; Schinka, Kinder, & Kremer, 1997; Wetter, Baer, Berry, Smith, & Larsen, 1992) and "personal reliability" (Jackson, 1976). Psychometric synonyms and antonyms empirically identify item pairs that have the largest positive or negative (respectively) inter-item correlations to determine whether participants are responding similarly to items with conceptually similar content. Personal reliability examines the similarity with which respondents answer items on two halves of the same test (e.g., first/last half or even/odd items). Inconsistent respondents are flagged as potential LQD.

Statistical approaches have many of the same benefits and drawbacks as unobtrusive approaches in that they are undetectable to respondents, yet lack well-defined cutoffs. Consistency-based indices are effective at identifying random responding (Berry et al., 1991; Pinsoneault, 2007). *D* is computationally complex but seems to slightly outperform other screening techniques when careless responses are randomly distributed (see, Meade & Craig, 2012). Personal reliability estimates require the two halves of the test (front/back, even/odd) to be comparable and can only be used in unidimensional tests (DeSimone et al., 2015).

Selection of Screening Techniques

The multitude of available screening options may seem daunting, but there are a number of guidelines researchers or practitioners can use to determine the most appropriate screening technique(s). It is important to emphasize that not all screens are appropriate for use in all studies. The suitability of screening techniques depends on survey design and methodology (Curran, 2016; DeSimone et al., 2015; Dunn et al., in press). For example, time-based screens require survey administrators to measure the time required for each respondent to complete the survey. As a result, time-based screens are easier to implement in computer-based surveys than paper-and-pencil surveys. Self-report and bogus or instructional items require insertion of additional items into a survey. Introducing new items may disrupt survey flow or affect respondents' perceptions of the survey. The longstring and IRV indices assume that consecutive items share the same response scale. Personal reliability coefficients are most appropriate when there are a large number of unidimensional scales containing a large number of items.

Each screening technique differs in calculation and relies on somewhat different assumptions. As a result, each screening technique is designed to be sensitive to a slightly different form of LQD. For example, the longstring and IRV indices would certainly flag Kyle's survey-long string of "middle" responses. Additionally, Kyle's speed of completion would likely appear conspicuous to a researcher examining a response time indicator. Since Jane was pressed for time, she also may be flagged by a time-based screen. Jane's random responding may be flagged using a bogus or instructed item screen or negatively impact her personal reliability.

Participants like Kate may employ cognitive shortcuts such as skimming the items or trying to "figure out" the test. If Dr. Smith used homogeneous scales in which all items were scored in the same direction, then Kate's responses would likely all fall on the extreme ends of the response options. If this were the case, then the longstring index would flag Kate's responses as LQD. If Kate's responses were dissimilar to the mean sample responses, then *D* would identify her responses as LQD. Henry's paranoid pattern of responses has the potential to be flagged by a number of indices of LQD. His intentional inconsistency may be detected through the use of personal reliability or psychometric synonyms. His intentional dissimilarity to the normative response pattern would be captured by *D*.

As Schmitt and Stults (1985) note, LQD may take multiple forms. Participants may demonstrate insufficient effort by responding randomly or invariantly. Other participants may exhibit response biases such as acquiescence, social desirability, or demand characteristics. These biases may yield responses that are a function of both content-dependent and content-irrelevant constructs (cf. Cronbach, 1950; Frederiksen, 1965; Messick, 1960; Orne, 1962).

Research Questions

The presence of participants who exert insufficient effort or content-irrelevant variance in their responses is undesirable in survey research. No single screen is capable of flagging each of these types of LQD. Therefore, researchers and practitioners can benefit from developing a better understanding of the types of screening techniques available, how they interact with one another, and the effects that they have on data and analysis. This paper examines five research questions in an effort to better understand the impact of using various screening techniques on the data analysis process.

The Prevalence of LQD The prevalence of LQD is a difficult but important issue to address. Estimates of the prevalence of careless responding vary widely. Costa and McCrae (1997) report that less than 1% of respondents admit to answering dishonestly. Other estimates are much higher, including rates of partial random responding as high as 60% (Berry et al., 1992). Most estimates range between 5 and 15% (see, Meade & Craig, 2012). Estimates of LQD rates are likely to be a function of the screening techniques employed by the data analyst. As a result, we will attempt to determine how estimated rates of LQD may be influenced by the selection of screening techniques.

> *Research question 1*: Do data screening techniques differ in the proportion of participants flagged?

Are multiple data screens redundant in the participants they flag? There are many reasons why participants may provide LQD and many tools with which the researcher can identify LQD. Although each tool uses a different technique to identify potentially problematic response patterns, some rely on similar logic. For example, psychometric synonyms and personal reliability are both designed to identify participants who demonstrate response inconsistency. As a result, we expect the psychometric synonyms and personal reliability indices to be positively correlated. In contrast, the longstring index identifies participants who demonstrate *too much* response consistency and would be expected to correlate negatively with psychometric synonyms and personal reliability. Similarly, the longstring index should be related to IRV, as long sequences of identical responses will yield lower response variance.

Because conceptually similar screens differ in the operationalization of LQD, no screening technique is expected to be entirely redundant with any other technique. One method of determining whether screening indices are redundant in identifying LQD is to examine the proportion of participants who are flagged by multiple screening techniques. Another method of identifying potential redundancy in LQD identification involves examining the correlations between various screening indices. If these techniques are redundant in functionality, then researchers should expect high inter-screen correlations and most of the flagged participants to be captured by multiple techniques.

Huang et al. (2012) reported inter-screen correlations that ranged from 0.18 to 0.69, suggesting moderate overlap between the self-report, response time, longstring, psychometric antonym, and personality reliability screening indices. The present analysis will supplement this analysis through the inclusion of bogus items, instructed items, IRV, and *D*.

> *Research question 2*: How many participants are flagged by multiple screens? What is the relationship between data screening techniques?

What is the distribution of screening indices? To evaluate the performance of various screening techniques, it is imperative that researchers understand the distributional characteristics of each screening index. This represents a first step in understanding the strengths and weaknesses of each technique. If a researcher plans to cull LQD participants, then it is important to ensure that appropriate cutoffs are selected for each index used. Like any statistical technique, it is important to select a value that maximizes the number of hits and correct rejections while minimizing false positives and false negatives. Although each screening technique can be used to make a dichotomous decision (i.e., flag or do not flag), each LQD index examined in this paper can also be expressed as a polytomous or continuous variable.

As noted above, a major limitation of many data screening techniques is that a little rationale exists for selecting cutoff values. The percentage of participants flagged is a function of the cutoff score set for the technique. With the exception of *D*, cutoff values for screening techniques are necessarily subjective and arbitrary. Previous literature has used cutoff values for most screening techniques, but none of these values is set in stone and the appropriateness of each may be influenced by aspects of data collection (e.g., length of a survey, balance of positively and negatively worded items; see, DeSimone et al., 2015).

This paper is not intended to definitively answer the question of what cutoff scores are appropriate for each screening technique. In fact, we discourage readers from searching for a "perfect" cutoff value of any statistic and encourage them instead to weigh multiple pieces of evidence when deciding whether or not to remove any given participant. However, knowledge of the distributional characteristics of screening indices can help inform these decisions, as can knowledge pertaining to the relative standing of respondents on each index.

> *Research question 3*: What are the distributional characteristics of each data screening index?

How does LQD impact results? There is a general consensus that LQD should be identified and excluded from analysis (Berry et al., 1992; Credé, 2010; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Johnson, 2005; Liu et al., 2013; McGrath, Mitchell, Kim, & Hough, 2010; O'Rourke, 2000; Wilkinson and Task Force on Statistical Inference, 1999). There is disagreement, however, about the potential consequences of LQD. Some researchers claim that the inclusion of participants who exhibit content-nonresponsivity or content-responsive faking may have negligible effects on psychometric properties and convergent validity estimates (Costa & McCrae, 1997; Kurtz & Parrish, 2001; McCrae et al., 1989). Other research demonstrates that the inclusion of these participants is associated with changes in factor structure (Schmitt & Stults, 1985; Woods, 2006), decreases in validity estimates (Hough et al., 1990; Liu, Huang, Bowling, & Bragg, 2013), and decreases in estimates of internal consistency and structural equation modeling (SEM) fit (Huang et al., 2012). Research that fails to find these effects typically relies on substantially smaller samples than the research that finds effects, indicating that statistical power may play an important role in elucidating the impact of LQD on research results.

Researchers disagree about the proportion of LQD required to negate the utility of an instrument. Gallen and Berry (1997) note that low levels of partial random responding may not affect the interpretability of a test. Schmitt and Stults (1985) use simulated data to determine that LQD rates between 5 and 10% are required to impact the factor structure of a test that includes negatively worded items. Clark et al. (2003) claim that LQD rates of 8 or 9% may not impact test scores on the MMPI while higher levels (around 15%) are sufficient to do so.

In our opinion, even minimal amounts of LQD are undesirable because data containing construct-irrelevant variance, which contaminates study results, is introduced into a dataset. Therefore, even minor changes due to LQD can distort the veracity of the results as these changes call into question the trustworthiness of the data (and the conclusions drawn using those data). The current analysis empirically examines the impact of various data screening techniques on inter-item correlations, inter-scale correlations, estimates of internal consistency, and relationships between demographic characteristics and character strengths.

> *Research question 4*: Can the use of data screening affect scale characteristics or relationships between items and scales?
>
> *Research question 5:* Can the use of data screening affect the results of statistical tests?

Method

Data

To address these five research questions, a series of self-report questions was administered online to a group of participants recruited through Amazon's Mechanical Turk. Participants were compensated $1 for taking part in the study, which included demographic questions, the International Personality Item Pool version of the Values in Action Inventory (IPIP-VIA), and a brief measure of character strengths. Of the 382 participants recruited, 337 (88.2%) met the inclusion criteria, which included US citizenship, being currently employed, and being older than 18 years of age. Participants who answered at least 90% of the questions on each measure were included in the analyses. Three hundred and seven participants (80.4%) met this criterion and were included in the analysis sample. The average age of participants was 33.62 (s = 10.62), and the analysis sample comprised 159 (51.8%) males, 147 (47.9%) females, and 1 individual who did not disclose his or her gender.

Measures

The IPIP-VIA (Goldberg et al., 2006) contains 213 items intended to measure 24 character strengths. Each character strength subscale contains between seven and eleven items. Like all IPIP items, each item is self-rated on a scale from 1 (very inaccurate) to 5 (very accurate). Sixty-six items are reverse scored, and internal consistency estimates for the subscales range from 0.70 to 0.91. The items and scoring key for the IPIP-VIA are available at http://ipip.ori.org/newVIAKey.htm. The presentation of IPIP-VIA items was randomized for each participant. As a result, the calculation of some screening indices (e.g., longstring) was not possible. Additionally, the utility of unobtrusive and statistical data screens may be mitigated by differences in item order to the extent that inter-item proximity influences responses and inter-item correlations. Although bogus items, instructed items, reaction time and self-report data quality items were still relevant to this measure, the majority of data screening indices were calculated on the shorter measure (Abbreviated Character Strengths Test (ACST)).

The ACST (Vanhove, Harms, & DeSimone, 2016) is a 24-item self-report measure of character strengths. Each item is rated on a scale from 1 (never/rarely) to 11 (always), and none are reverse scored. Each ACST item is intended to correspond to 1 of the 24 IPIP-VIA character strengths, and the 24 items are organized into six subscales based on the theoretical work of Peterson and Seligman (2004).[1] Each ACST subscale contains three to five items and internal consistency values range from 0.70 to 0.84. ACST items were presented to all participants in the same order, so most screening indices were computed using the ACST data.

A set of demographic items was also administered, but most of these items did not directly influence our analyses. These items asked participants to report their residency, employment status, gender, age, and ethnicity. The age and gender items were used to demonstrate the influence of data screening on statistical results (research question 5).

Missing Data

Missing data impacts screening indices in different ways (DeSimone et al., 2015). Researchers interested in employing screening techniques can impute missing data or (more conservatively) use pairwise or listwise deletion. The use of pairwise deletion relies on the assumption of data missing completely at random (Allison, 2009) and is inconsistent with the notion of construct-irrelevant variance as understood in LQD. In the current analyses, listwise deletion would have decreased the sample substantially (to 223 participants, or 58.4% of the original sample), so regression-based imputation was used to estimate the values of missing data when necessary.[2] Imputed responses were rounded to the nearest whole number in order to mimic plausible response selections.

[1] It is noteworthy that subsequent work has failed to replicate this six-factor structure but has also failed to consistently support an alternative factor structure using confirmatory analysis (Vanhove et al., 2016).

[2] Using listwise deletion did not substantially change the results of the analysis. For example, the percentage of participants flagged differed by 2.7% or less for each screening technique and correlated higher than 0.99 with percentages computed using the data imputation technique.

Data Cleaning Methods and Cutoffs

All unobtrusive and statistical screening indices were calculated using the ACST because the ACST items were administered in the same order for all respondents. Each technique used in this study has been demonstrated to be appropriate for use with ordinal, interval, or ratio data, including dichotomous and Likert-type items. As noted above, there is no perfect cutoff score for any data screening technique. However, previous literature has identified cutoffs for each technique (listed below) that are applied, where necessary, in the current analyses.

Self-report Items Four self-report questions were appended to the end of the questionnaire. These questions asked participants to indicate the frequency of answering questions honestly, responding without carefully reading the questions (reverse-scored), putting thought into survey responses, and using little effort when selecting answers (reverse-scored). Each item was rated on a five-point scale ranging from "very rarely" to "very often." Lower scores indicated potential LQD. For example, a score of 2.0 means that, on average, the participant indicated that (s)he responded thoughtfully and effortfully only "somewhat rarely." Participants with an average score below 4.0 were flagged as potential LQD.

Bogus and Instructed Items Two bogus items ("I was born on February 30" and "I have exactly 354 best friends") and two instructed items ("Please indicate option [X] for this question") were inserted into the IPIP-VIA survey. Consistent with Bagby et al.'s (1991) recommendations, respondents were flagged if they responded incorrectly to any of these items. Due to this cutoff recommendation and the conceptual similarity of the bogus and instructed item techniques, these two methods were combined in the analyses. Higher values indicate LQD because more bogus or instructed items were "missed." For example, a score of 2 indicates that a participant missed two items and a score of 0 means they missed none.

Response Time The response time screen was computed using the average number of seconds required to complete each item. For example, a score of 1.0 indicates that the participant required 1 s/item while a score of 2.5 indicates that the participant required 2.5 s/item. Lower scores indicate less time required to complete each item, with very low scores indicating LQD. For the purposes of data cleaning, Huang et al. (2012) suggested screening participants who required less than 2 s/item. A single measure of total time elapsed was provided by the software, so participants who completed the entire survey in a time faster than 506 s were flagged. This cutoff score reflects an average of 2 s/item across the entire survey (including the IPIP-VIA, ACST, demographic questions, and self-report data quality questions).

Longstring The longstring index was defined as the maximum number of consecutive invariant responses provided by a respondent. Higher scores indicate longer sequences of invariant responding and, therefore, LQD. For example, a score of 10 indicates that the participant indicated at least one string of ten consecutive identical responses. According to Costa and McCrae (2008), participants who indicate consecutive strings of at least six "strongly disagrees," nine "disagrees," ten "neither agree nor disagrees," fourteen "agrees," or nine "strongly agrees" should be flagged. Huang et al. (2012) revised these estimates to seven, seven, twelve, ten, and eight, respectively. Since response options in the present data did not identically conform to the format used by the aforementioned analyses, we selected a single value for the cutoff across all response options. Participants who invariantly responded to at least nine items in any scale set were flagged. The cutoff of nine invariant responses was chosen because it reflects the median of Costa and McCrae's (2008) analysis and is close to the mean (8.80) of Huang et al.'s (2012) analysis.

IRV IRV was calculated as the standard deviation of responses for each participant. Lower IRV values indicate LQD as they are associated with less variance in item responses. For example, an IRV value of 2.7 indicates that the participant had a standard deviation of 2.7 across his or her responses to the 24 items on the ACST. Based on the number of items and response options, the maximum IRV value for the ACST is 5.11 and the minimum is zero. IRV is a relatively new technique, and no cutoff score has been specified to date. Consistent with Dunn et al. (in press), the 31 participants (approximately 10%) with the lowest IRV scores were flagged.

Psychometric Synonyms Psychometric synonyms were identified by examining the inter-item correlation matrix. Item pairs with correlations above 0.60 were identified as psychometric synonyms (Meade & Craig, 2012). Items that were identified in two or more synonym pairs were assigned to the pair with the largest correlation. Five pairs of psychometric synonyms were identified (Q1/Q2, Q3/Q8, Q4/Q5, Q10/Q11, and Q13/Q14). Each participant's psychometric synonym coefficient was computed by correlating his or her responses to Q1, Q3, Q4, Q10, and Q13 with his or her responses to Q2, Q8, Q5, Q11, and Q14. Lower scores indicate LQD in the form of response inconsistency.

To remain consistent with previous literature, participants with psychometric synonym coefficients below 0.22 were flagged. This cutoff value was calculated by using a weighted average of the mean psychometric synonym coefficients for the two "careless responding" latent classes identified in Meade and Craig's (2012) analysis. Due to the nature of the 24-item survey (all items were positively worded and scored in the same direction), no psychometric antonyms could be identified.

Personal Reliability Jackson's (1976) personal reliability coefficient was computed by correlating the average score on even items with the average score on odd items for each subscale of the ACST. Lower scores indicate LQD in the form of response inconsistency. Personal reliability was computed using a within-person correlation between the vector of even response averages and the vector of odd response averages adjusted for double length using the Spearman-Brown prophesy formula (Brown, 1910; Spearman, 1910). Consistent with Johnson (2005), participants were flagged if their corrected personal reliability coefficient did not exceed 0.30.

Mahalanobis *D* *D* values were calculated using the formula $D=\sqrt{\left(\overrightarrow{X_i}-\overrightarrow{\overline{\overline{X}}}\right)^T*\mathrm{COV}_{XX}^{-1}*\left(\overrightarrow{X_i}-\overrightarrow{\overline{\overline{X}}}\right)}$, where $\left(\overrightarrow{X_i}-\overrightarrow{\overline{\overline{X}}}\right)$ represents the vector of mean-centered item responses for participant i and $\mathrm{COV}_{XX}{}^{-1}$ represents the inverted covariance matrix of all items. Larger deviation from the normative response pattern yields higher D values and is considered a potential indicator of LQD. A single D statistic was computed for each of the participants using all ACST items. The squared value of D follows a chi-square distribution with degrees of freedom equal to the number of items used in the calculation of D. Participants were flagged if their D^2 value placed them in the highest 5% of the chi-square distribution.

Results and Discussion

Research Question 1: Do Data Screening Techniques Differ in the Proportion of Participants Flagged?

Table 1 lists the proportions of participants flagged by each screening technique using the cutoff values indicated in the "Method" section. Results indicate that there are differences in the proportion of participants flagged by each technique. The self-report technique identified the lowest proportion (4.9%) while the psychometric synonym technique identified the highest (40.4%). Employing all seven techniques would reduce the sample to 74 respondents (retaining only 24.1% of participants).

Four screening techniques (bogus/instructed items, response time, longstring, and IRV) provide estimates consistent with the 5 to 15% range typically found in previous literature, though IRV was forced to fall within this range by design (Dunn et al., in press). Two others (self-report at 4.9% and D at 17.9%) were relatively close to the 5 to 15% range. Psychometric synonyms and personal reliability flag substantially more participants than the other techniques. As noted above, both of these indices are related to response consistency and rely on the identification of conceptually similar items. Cutoff criteria for psychometric synonyms and personal reliability may be too liberal for the current study. While all indices vary in utility as a function of study characteristics, consistency-based indices may be especially sensitive to scale attributes (e.g., homogeneity of item content).

Practically, these results indicate that the selection of screening techniques will influence the number of participants who are flagged and/or eliminated from the analysis. If researchers or practitioners plan to use the established cutoff scores noted above, they should anticipate that direct and unobtrusive screens will flag 5 to 10% of their sample while consistency-based indices (e.g., personal reliability, psychometric synonyms) will flag a much larger proportion of respondents. We suggest that practitioners and future researchers carefully consider both the type of screening techniques and the expected performance of those techniques when designing a survey.

Research Question 2: How Many Participants Are Flagged by Multiple Screens? What Is the Relationship Between Data Screening Techniques?

Screening indices are minimally correlated with one another, and about half of flagged participants trigger multiple screens. Table 2 indicates that, of the 233 individuals identified by screening techniques, almost half (115) were only identified by a single technique, whereas 77, 34, 5, and 2 respondents were identified by two, three, four, and five screening techniques, respectively. These results suggest small overlap between most screening techniques with moderate overlap between a few.

Table 3 displays an alternative assessment of the overlap between screening techniques by reporting Pearson correlations between the values of screening indices. This method is not influenced by cutoff values because scores on the screening indices were not dichotomized for the purposes of determining which participants were flagged. Instead, the screening index values were used in their continuous or polytomous form. The pattern of correlations is consistent with the contention of low to moderate overlap between most screening techniques.

Two of the largest correlations are between the two consistency indices (psychometric synonyms and personal reliability at − 0.30) and the two invariance indices (longstring and IRV at 0.43). These results are consistent with the similarity in the rationale behind the two pairs. Psychometric synonyms and personal reliability are both intended to flag inconsistent responses while longstring and IRV are both intended to flag invariant responses.

A very large negative correlation was observed between IRV and D, indicating that respondents flagged by one these

Table 1 Number of participants screened using each technique

Screening technique	Number of participants screened	Percentage of participants screened
Self-report	15	4.9
Bogus/instructed	23	7.5
Response time	22	7.2
Longstring	31	10.1
Individual response variability	31	10.1
Psychometric synonyms	124	40.4
Personal reliability	100	32.6
Mahalanobis D^2	55	17.9
All screens combined	233	75.9

techniques are unlikely to be flagged by the other. It is difficult to speculate about the reasons for this strong negative relationship on the basis of a single sample and analysis, but this result is likely related to the variance of the mean response vector. Future research should examine whether this strong negative correlation exists in other samples and model this relationship as a function of distributional characteristics of the mean response vector.

The longstring index was negatively correlated with all indices except IRV, bogus/instructed items, and self-reported effort. This indicates that invariant respondents were more likely than others to miss a bogus or instructed item or self-identify as providing LQD, but also that the longstring index has the potential to complement other unobtrusive and statistical screens well because it will identify a different set of LQD respondents. This is unsurprising as the longstring index flags overly consistent participants while consistency-based indices flag inconsistent participants.

Finally, the self-reported effort screen was only statistically related to the bogus/instructed item and response time screens, indicating that respondents may be aware that they are not paying attention or responding too quickly. The lack of correlation between self-report and other indices indicates that participants may not be capable of perceiving their own response variance, inconsistency, or deviation from the norm (as, these are indicated by other indices such as longstring, IRV, personal reliability, psychometric synonyms, and *D*).

Overall, these results indicate that practitioners and researchers have some flexibility when it comes to selecting screening techniques. With a few exceptions (e.g., IRV and longstring or *D*), the screening techniques are not highly correlated with each other, indicating that multiple screens are not redundant in identification of LQD. As a result, we encourage readers to consider the selection of screening techniques as a part of survey design. Combined use of complementary techniques can flag multiple types of potentially problematic data.

Research Question 3: What Are the Distributional Characteristics of Each Data Screening Index?

Table 4 contains the minimum, maximum, mean, standard deviation, skewness, and kurtosis of each of the screening indices. Most screening techniques share two important distributional properties. First, there is substantial variability in each screening index, which indicates a wide range of scores on each data screening index. Second, most mean values lie within the range of normal (unflagged) values. Our results indicate that the average participant reports high effort, fails no bogus or instructed items, spends 5.37 s on each item, has a maximum of four to five consecutive identical responses, has a standard deviation of responses around 1.84, has a positive correlation between psychometrically similar items, and does not have a significantly different response vector from the sample mean.

In order to better understand the relationship between cutoff scores and proportions of participants identified by the various screening techniques, cumulative distributions of screening index scores are presented in Fig. 1. To our knowledge, this is the first graphical depiction of screening index distributions. As a result, we urge caution when interpreting or generalizing these results before additional data are reported.

Table 2 Number of screens participants failed

Number of screening techniques	Number of participants	Percentage of participants
0	74	24.1
1	115	37.5
2	77	25.1
3	34	11.1
4	5	1.6
5	2	0.7

Table 3 Correlations between screening indices

	SR	BI	RT	LS	IRV	PS	PR	MD
SR	1							
BI	0.21*	1						
RT	0.13*	0.06	1					
LS	0.11	0.16*	− 0.12*	1				
IRV	− 0 .15*	0.16*	0.07	0.43*	1			
PS	− 0.02	0.06	− 0.01	− 0.09	− 0.03	1		
PR	0.07	0.06	0.01	− 0.03	0.02	0.30*	1	
MD	− 0.03	− 0.02	0.05	− 0.22*	− 0.84*	0.02	0.09	1

Note. N = 288–307 (some PS and PR indices cannot be computed due to invariant responding). All screening indices scored in the same direction (overlap indicated by a positive correlation)

SR self-report, *BI* bogus and instructed, *RT* response time, *LS* longstring, *IRV* individual response variability, *PS* psychometric synonym, *PR* personal reliability, *MD* Mahalanobis D^2

$^*p < 0.05$

Each distribution appears to be quadratic or cubic, indicating that each screening index has at least one inflection point at which it becomes substantially more or less effective. Most screens have an inflection point corresponding to the 5 to 15% range of respondents we would expect to engage in LQD based on estimates from previous research.

The practical implications of these results are difficult to establish in a single study. The fact that mean values lie within the normal range is good news, as it suggests that most of our participants are responding in a manner consistent with what we would expect from thoughtful and effortful respondents. We hope that documenting the distributional characteristics of each technique may lead to improved normative data and a better sense of which respondents are providing LQD. These results can be combined with previous and future research to provide researchers and practitioners with a better understanding of how we should expect respondents to act and screening indices to perform.

Research Question 4: Can the Use of Data Screening Affect Scale Characteristics or Relationships Between Items and Scales?

Data screening can influence estimates of internal consistency, inter-item correlations, and inter-scale correlations, but the influence varies by technique and is small to moderate for most techniques. In order to assess the impact of data screening on correlation matrices, standardized root mean-square residual (SRMR; Bentler, 1995) values were used to compare correlation matrices computed before and after employing various screening techniques. Tables 5 and 6 contain SRMR matrices for inter-item and inter-scale correlations, respectively. As noted in Table 1, each screening technique flags a different set and proportion of respondents. Accordingly, each scenario represented in Tables 5 and 6 compares two identical correlation matrices computed on different subsets of the sample. The implementation of each screening technique resulted in a different set of flagged participants and, accordingly, different sample sizes.

SRMR is a common tool for comparing correlation matrices (to provide evidence of model fit) in SEM. At a basic level, SRMR examines the average change in non-redundant cells of a correlation matrix. Lower values of SRMR indicate more

Table 4 Descriptive statistics for each screening index

Screening technique	Minimum value	Maximum value	Mean	Standard deviation	Skewness	Kurtosis
Self-report	3.00	5.00	4.74	0.45	− 1.77*	2.33*
Bogus/instructed	0.00	4.00	0.12	0.47	4.91*	27.49*
Response time	0.83 s/item	76.59	5.37	6.54	6.87*	60.84*
Longstring	1.00	24.00	4.49	4.25	2.92*	9.36*
Individual response variability	0.00	4.05	1.84	0.80	0.22	− 0.03
Psychometric synonyms	− 1.00	1.00	0.31	0.48	− 0.58*	− 0.56
Personal reliability[a]	− 21.12	0.99	− 0.17	2.46	− 5.40*	35.09*
Mahalanobis D^2	3.03	110.67	23.92	17.30	1.60*	3.15*

Note. Due to the use of the Spearman-Brown prophesy formula, correlations of − 0.34 or lower will yield a personal reliability coefficient lower than − 1.00

s/item seconds per item

*$p < 0.05$, statistically different from zero

[a] If all personal reliability coefficients less than − 1.00 are replaced with values of − 1.00, the minimum value becomes − 1.00, the mean becomes 0.29, the standard deviation becomes 0.65, the skewness becomes − 0.96*, and the kurtosis becomes − 0.46. If an equivalent adjustment is made for both positive and negative correlations, the mean becomes 0.34, the standard deviation becomes 0.56, the skewness becomes − 0.83*, and the kurtosis becomes − 0.60*

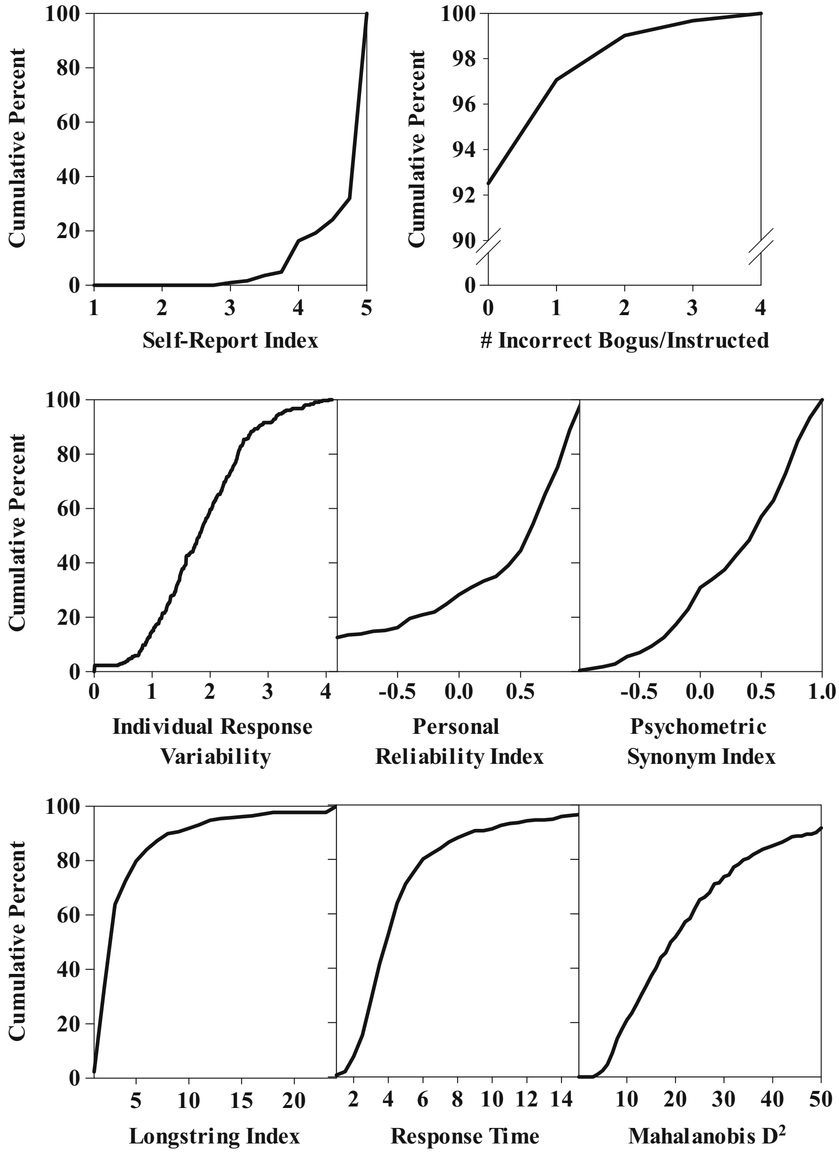

Fig. 1 Cumulative probability distributions for values of each data screening index

similarity in the correlations matrices, and cutoff values of 0.07 or 0.08 are commonly used in SEM to indicate dissimilarity (Hu & Bentler, 1999; Yu, 2002). Recent research has demonstrated the utility of using SRMR outside the context of SEM to compare correlation matrices across time or samples (DeSimone, 2015).

Each cell in Tables 5 and 6 represents the SRMR statistic resulting from a comparison of inter-item or inter-scale correlation matrices computed after eliminating flagged respondents using the corresponding screening techniques. For example, in Table 5, comparing the inter-item correlation matrix where no screening techniques were used with the matrix where all techniques were used yields an SRMR of 0.12 (0.15 for inter-scale correlations in Table 6). This level of matrix inequivalence indicates moderate changes in the inter-item and inter-scale correlation matrices. Alternatively, comparing the inter-item correlation matrix computed after eliminating respondents who were flagged by longstring with the matrix computed after eliminating participants flagged with IRV yields an SRMR of 0.02 (0.01 for inter-scale correlations in Table 6). This value indicates a negligible difference in inter-item and inter-scale correlation matrices, suggesting that choosing between longstring and IRV will not strongly influence the estimation of these relationships.

The results in Tables 5 and 6 demonstrate the general extent to which implementation of each screening technique changes the inter-item and inter-scale correlations. Compared with the unscreened data, results indicate that implementing the longstring, IRV, and *D* indices is associated with the largest changes in inter-item and inter-scale correlation matrices while the other screening indices produced negligible changes in both. Combining all available screening techniques also yields

Table 5 SRMR matrix for inter-item correlation matrices

	None	SR	BI	RT	LS	IRV	PS	PR	MD	All
None	0									
SR	0.05	0								
BI	0.03	0.02	0							
RT	0.04	0.02	0.02	0						
LS	0.09	0.05	0.06	0.05	0					
IRV	0.09	0.04	0.06	0.05	0.02	0				
PS	0.05	0.09	0.08	0.09	0.13	0.13	0			
PR	0.05	0.08	0.07	0.08	0.12	0.12	0.05	0		
MD	0.11	0.15	0.14	0.14	0.19	0.19	0.08	0.09	0	
All	0.12	0.10	0.10	0.10	0.10	0.10	0.14	0.12	0.18	0

Note. Based on 24 × 24 inter-item correlation matrices

None no screens employed, *SR* self-report, *BI* bogus and instructed, *RT* response time, *LS* longstring, *IRV* individual response variability, *PS* psychometric synonym, *PR* personal reliability, *MD* Mahalanobis D^2 , *All* all screens employed

a correlation matrix that is different from the full-sample correlation matrix.

We also examined the extent to which data screening influences internal consistency. To assess changes in internal consistency, Cronbach's (1951) alpha was computed for each of the six subscales of the ACST as well as the overall test before and after employing each of the screening techniques. Results are presented in Table 7. The effect of data screening on alpha coefficients ranged from negligible (changes of 0.00) to moderate (changes of 0.11) and varied by screening technique. Eliminating participants flagged by the self-report, bogus/instructed, response time, longstring, and IRV techniques tended to reduce estimates of alpha, whereas the use of psychometric synonyms, personal reliability, and *D* tended to increase estimates of alpha. The combined use of all available screening techniques yielded mixed results, with notable increases (changes between 0.05 and 0.11) in some subscales and minor decreases (changes between 0.00 and 0.06) in others.

Table 6 SRMR matrix for inter-scale correlation matrices

	None	SR	BI	RT	LS	IRV	PS	PR	MD	All
None	0									
SR	0.04	0								
BI	0.03	0.02	0							
RT	0.04	0.01	0.01	0						
LS	0.08	0.04	0.06	0.05	0					
IRV	0.09	0.05	0.07	0.06	0.01	0				
PS	0.03	0.07	0.05	0.06	0.11	0.11	0			
PR	0.03	0.03	0.02	0.03	0.06	0.07	0.04	0		
MD	0.07	0.11	0.09	0.10	0.15	0.16	0.05	0.09	0	
All	0.15	0.11	0.13	0.12	0.08	0.08	0.16	0.13	0.21	0

Note. Based on 6 × 6 inter-scale correlation matrices

None no screens employed, *SR* self-report, *BI* bogus and instructed, *RT* response time, *LS* longstring, *IRV* individual response variability, *PS* psychometric synonym, *PR* personal reliability, *MD* Mahalanobis D^2 , *All* all screens employed

The most pronounced decrease in alpha was associated with the longstring and IRV indices. This is consistent with the idea that these indices flag overly consistent responding, resulting in a decrease in alpha in scales with homogeneous directionality of items (such as the ones used here). In contrast, the most consistent increases in alpha were produced by the screening techniques that encourage response consistency (psychometric synonyms, personal reliability) or homogeneity between respondents (*D*).

The major contribution of these results is the demonstration that data cleaning can lead to changes in the ways in which items operate within a scale. Internal consistency and inter-item correlations may change to a minor or moderate extent depending on which screening techniques are employed. These changes may be reflected in factor structure, inter-scale correlations, and correlations with other measures or criteria. This finding underscores the importance of carefully selecting the screening techniques used in data cleaning. If screening can impact results, then it is important to ensure we are flagging the appropriate respondents to maximize the probability that accurate and meaningful data are analyzed while LQD are flagged and eliminated from the sample.

Research Question 5: Can the Use of Data Screening Affect the Results of Statistical Tests?

Consistent with previous research (e.g., Maniaci & Rogge, 2014), our results indicate that data screening can have a moderate to large impact on statistical results. In order to assess the impact of screening techniques on statistical results, two basic analyses were conducted. First, scores on ACST subscales and the overall scale were correlated with age (Table 8). Second, gender differences were computed using independent-sample *t* tests (Table 9).

While age and gender may not be the most theoretically interesting variables, there are a few advantages to examining these relationships. First, since demographic data was captured at the beginning of the survey, it is plausible that most participants were paying attention when responding to these items, meaning that age and gender may be less susceptible to low-quality responding in the current study design.

Second, our goal for this analysis is to investigate whether eliminating suspected LQD can affect conclusions about statistical relationships. The analysis of age and gender represent some of the most basic inferential statistical techniques (correlations and independent-sample *t* tests, respectively) available to researchers. If these analyses are affected by data

Table 7 Internal consistency estimates before and after employing screening techniques

Screening technique	S1	S2	S3	S4	S5	S6	Entire test
None	0.85	0.72	0.76	0.76	0.74	0.73	0.93
Self-report	0.83	0.68	0.73	0.75	0.70	0.70	0.92
Bogus/instructed	0.84	0.69	0.74	0.75	0.71	0.71	0.92
Response time	0.83	0.68	0.73	0.73	0.70	0.71	0.92
Longstring	0.81	0.61	0.69	0.72	0.68	0.66	0.90
Individual response variability	0.81	0.64	0.70	0.71	0.67	0.66	0.90
Psychometric synonyms	0.87	0.76	0.83	0.82	0.77	0.75	0.94
Personal reliability	0.88	0.76	0.83	0.84	0.79	0.79	0.94
Mahalanobis D^2	0.90	0.78	0.81	0.81	0.81	0.79	0.95
All screens combined	0.90	0.66	0.83	0.87	0.70	0.73	0.91

Note. Internal consistency estimates based on Cronbach's alpha. S1 to S6 contain five, four, three, three, four, and five items, respectively; the entire test contains 24 items

*S1*subscale 1

screening, then it is likely that more advanced analyses will be as well.

Previous research has established the existence of moderate gender and age differences in character strengths (Littman-Ovadia & Lavy, 2012; Ruch et al., 2010). The existence of these relationships ensures the presence of variance to explain. However, a moderate to strong relationship between LQD and either age or gender may serve as a confound by introducing plausible alternative explanations for changes in statistical results. Fortunately, age and gender were negligibly or minimally related to all LQD indices. Specifically, correlations with age ranged from − 0.18 (bogus/instructed) to 0.20 (self-report), with only the bogus/instructed and self-report indices being statistically significant. All other indices had correlations of 0.10 or less with age. While age did not account for a substantial amount of variance (4.0% or less) in any screening index, readers may wish to interpret the bogus/instructed and self-report results in Table 8 with some caution. Independent-sample *t* tests did not reveal a gender difference in any of the screening indices, with Cohen's *d* values ranging from 0.01 (psychometric synonyms, personal reliability, and *D*) to 0.22 (longstring).

Correlations with age varied as a function of the screening technique used. Eliminating participants on the basis of bogus and instructed items had a minimal impact on correlations with age (changes of 0.00 to 0.02), while *D* resulted in slightly larger changes (0.01 to 0.06). Employing all screens yielded some of the largest changes (0.02 to 0.15).

Screening techniques had a more pronounced impact on gender differences. Use of individual screening techniques changed the Cohen's *d* estimates from negligible (0.00) to moderate (0.11) amounts. These changes occasionally impacted estimates of statistical significance but never changed the direction of the relationship. Combined use of the seven screening techniques had a larger impact on effect size estimates (changes in *d* ranged from 0.09 to 0.65), impacted statistical significance and even changed the direction of the effect in two analyses (e.g., subscale 4 changed from −.11 to 0.54).

Table 8 Correlations with age before and after employing screening techniques

Screening technique	S1	S2	S3	S4	S5	S6	Entire test
None	0.16*	0.14*	0.15*	0.15*	0.20*	0.24*	0.22*
Self-report	0.13*	0.11	0.12*	0.13*	0.18*	0.23*	0.20*
Bogus/instructed	0.16*	0.13*	0.14*	0.17*	0.21*	0.25*	0.22*
Response time	0.12*	0.11	0.12*	0.14*	0.18*	0.22*	0.19*
Longstring	0.17*	0.15*	0.17*	0.16*	0.22*	0.25*	0.25*
Individual response variability	0.12*	0.10	0.12*	0.13*	0.18*	0.22*	0.19*
Psychometric synonyms	0.15*	0.14	0.21*	0.24*	0.24*	0.26*	0.25*
Personal reliability	0.18*	0.17*	0.17*	0.17*	0.24*	0.23*	0.24*
Mahalanobis D^2	0.17*	0.16*	0.21*	0.17*	0.18*	0.29*	0.24*
All screens combined	0.12	0.12	0.29*	0.30*	0.26*	0.33*	0.31*

S1 subscale 1

**p* < 0.05

Table 9 Gender differences (Cohen's *d* values) before and after employing screening techniques

Screening technique	S1	S2	S3	S4	S5	S6	Entire test
None	− 0.19	− 0.25*	− 0.44*	− 0.11	− 0.21	− 0.30*	− 0.31*
Self-report	− 0.17	− 0.22	− 0.42*	− 0.11	− 0.20	− 0.29*	− 0.30*
Bogus/instructed	− 0.09	− 0.16	− 0.43*	− 0.10	− 0.15	− 0.24*	− 0.23*
Response time	− 0.17	− 0.24*	− 0.42*	− 0.10	− 0.18	− 0.32*	− 0.30*
Longstring	− 0.09	− 0.16	− 0.40*	0.00	− 0.13	− 0.21	− 0.22
Individual response variability	− 0.10	− 0.16	− 0.40*	− 0.03	− 0.12	− 0.24*	− 0.23
Psychometric synonyms	− 0.10	− 0.19	− 0.49*	− 0.05	− 0.32*	− 0.39*	− 0.32*
Personal reliability	− 0.12	− 0.16	− 0.35*	− 0.05	− 0.11	− 0.24	− 0.22
Mahalanobis D^2	− 0.23	− 0.32*	− 0.45*	− 0.06	− 0.35*	− 0.35*	− 0.36*
All screens combined	0.32	0.07	− 0.29	0.54*	− 0.12	− 0.08	0.07

Positive values indicate that males have higher scores than females

S1 subscale 1

*$p < 0.05$, statistical significance of independent-sample *t* test

If screening techniques can influence the results of the basic analyses like correlations and *t* tests then they can certainly influence the results of more complex analyses (e.g., SEM or meta-analyses that rely on correlational results). The results pertinent to research questions 4 and 5 indicate that data cleaning can influence the both psychometric estimates and statistical results. These findings underscore the importance of considering data screening a part of research design, thoughtfully selecting the appropriate screening techniques, and implementing these techniques in a fair and accurate manner in an effort to ensure that we retain appropriate data and eliminate LQD.

Summary

Data screening techniques differ in the methods used to eliminate participants. Some screening techniques emphasize consistency (e.g., psychometric synonyms, psychometric antonyms, personal reliability). Some techniques emphasize conformity to normative response patterns (e.g., *D*). Other techniques emphasize the identification of random or invariant response behaviors (e.g., response time, longstring, IRV). Each screening technique differs in assumptions, purpose, and computation.

The results from these analyses are consistent with the idea that participants typically respond to survey items thoughtfully, but a subset of respondents engage in LQD. The index inter-relationships empirically confirm that, while there is some overlap, each screening technique has the potential to identify a different set of respondents. Indeed, each index flagged some participants that no other index identified. Most of the inter-screen correlations were small to moderate, which should assuage concerns about redundancy in LQD identification. Accordingly, the percentages of participants flagged by each index also indicate differences between screening techniques. There is notable variance in the percentage of participants identified by each screening technique, and, although about half of the flagged participants are identified by multiple screens, relatively few are identified by more than two.

Data screening can have considerable impact on the results of a study, though the magnitude of these effects depends on both choice of screening techniques and the nature of the dataset. Most individual screening techniques have only minor effects on the inter-item and inter-scale correlation matrices as well as small to moderate effects on the relationship of the scale with demographic variables. On the other hand, combining screening techniques may have a moderate to substantial impact on internal consistency and statistical results.

The impact of data screening techniques on research is somewhat complex. The proportion of respondents flagged, changes in internal consistency estimates, and changes in statistical results are a function of the screening techniques used as well as their respective cutoff scores. Although more data would certainly be welcome, it seems as though the cutoff scores used in this study for self-report, bogus/instructed, response time, longstring, and *D* may be appropriate, while those used for psychometric synonyms and personal reliability may require further consideration.

Implications for Survey Design and Data Analysis

While data cleaning practices can improve data quality, it is important to note that screening is stochastic in nature. Data cleaning requires judgment on the part of the researcher to determine which techniques are most appropriate and how best to implement them (e.g., which cutoff scores to use). It is unlikely that any technique will perform perfectly, as false positives and negatives may still exist. While we discourage researchers and practitioners from assuming all data are useful and of sufficient quality, we also acknowledge that no post hoc screening procedures can completely “fix” LQD. None of these techniques can change LQD into high-quality responses. At their best, screening techniques may allow us to identify and eliminate some construct-irrelevant responses. Therefore, it is important for survey researchers to recruit suitable participants and use appropriate instructions and incentives to encourage honest and effortful responding.

We echo earlier calls (e.g., DeSimone et al., 2015; Huang et al., 2012; Meade & Craig, 2012) that encourage researchers to carefully consider which screening techniques to use in the course of study design. For example, bogus or instructed items are relatively simple to add to a questionnaire when respondent attentiveness is suspected to be a problem. *D* is most appropriate when researchers expect relatively homogeneous response patterns in the target population (i.e., outliers are suspect). The longstring and IRV indices are most appropriate when some survey items are reverse scored or multiple constructs are assessed in a group of items.

Due to their intuitive appeal and ease of calculation, it may be advisable to primarily rely on direct and unobtrusive screening techniques. Based on the graphs in Fig. 1 and the proportions of respondents flagged, it seems as though statistical screens are quite sensitive to cutoff score selection and may vary in effectiveness with survey characteristics. Although *D* is appropriate in some situations (see, Meade & Craig, 2012), it can be more difficult to justify than the direct or unobtrusive techniques. *D* compares individual responses to mean response values, which may be inappropriate when data are contaminated (e.g., when test administration procedures are inconsistent or a large proportion of participants engage in LQD).

We caution researchers and practitioners against arbitrary selection of screening techniques. When possible, it is advisable to use multiple complementary screening techniques in order to maximize the efficiency of the screening process (Bruehl et al., 1998; Meade & Craig, 2012). Our analysis reveals minimal overlap between most techniques, indicating that the use of multiple data screens is not redundant with respect to LQD identification. Prior to collecting data, it is advisable to consider the types of techniques available and the type of LQD (e.g., invariance, counternormative responses, random responding) each technique is intended to flag. For example, survey administrators may plan to employ bogus or instructed items to flag random responders, a longstring or IRV index to flag response invariance, and a consistency-based screen to flag response inconsistency. Researchers and practitioners should consider screening an integral part of the data collection process. Like other aspects of design and methodology, decisions about screening should be deliberate, defensible, and not based solely on convenience.

After data collection is complete, examination of graphical representations of the data may serve to identify aberrant response patterns (Anscombe, 1973). The current analyses reveal that respondents are more capable of identifying their own low-quality responding when responding quickly or inattentively. They are less likely to identify inconsistency in their responses. Due to low inter-correlations between screening indices, the use of self-report items, bogus items, instructed items, response time, and either the longstring or IRV index can be combined to capture different types of LQD. However, when the insertion of items will disrupt the flow of a study or potentially influence the structure of a measure it may be necessary to rely solely on unobtrusive indices such as longstring, IRV, or response time.

The use of data screening techniques benefits from high-quality survey design. Because screening techniques use different methods and potentially screen different participants, researchers and practitioners should attempt to understand or predict which type of LQD is likely to be problematic in the context of their study. This can be accomplished through pilot testing in order to gain a better understanding of the tendencies and characteristics of the population under investigation. For example, if participants seem to be completing the survey quickly, introduce a response time screen. If invariant responding is prevalent, introduce a longstring screen. Alternatively, if a survey designer is familiar with similar survey data and test administration procedures, it may be possible to anticipate which types of LQD are most likely based on previous experience. Additionally, researchers and practitioners can monitor data as it is being collected in an ongoing effort to better understand which types of LQD are most prevalent. However, it is not advisable to change survey instructions or add items (e.g., self-report, bogus/instructed) after data collection has begun as these changes may alter or disrupt survey flow for a subset of survey respondents.

Researchers should screen participants prior to analysis and eliminate participants suspected of providing LQD (Berry et al., 1992; Hough et al., 1990). In doing so, it is important for researchers to exercise transparency in reporting results. Researchers should always be clear about which data screening indices and cutoff values were used. When possible, cutoff values should be justified on the basis of previous literature or the proportion of respondents flagged. The rationale for each screen should be explained and results should be presented for both the post-screening and pre-screening data so that readers can better understand if and how screening decisions affected the results of analyses (McCrae et al., 1989; Stevens, 1984).

Limitations and Future Directions

Early research on the development and use of screening techniques relied heavily on the questions of how and why each was appropriate. Future research should continue to establish norms and best practices for the use of each technique by examining how each interacts with characteristics of the methodology, sample, and analysis.

The results of this analysis represent only a single application of data screening. It is unlikely that these results will generalize to all measures and survey research designs. The relationship between study characteristics (e.g., survey design) and the performance of screening indices is fertile ground for subsequent research, especially research that manipulates design characteristics such as survey format or length. It would also be interesting to determine if sample characteristics (e.g., participation incentive, student/non-student participants) influence the performance of data screening techniques.

The online data collection and accordant computer administration of the ACST and IPIP-VIA can be considered both a benefit and limitation to this study. This mode of administration facilitated the measurement of response time and allowed the ability to easily measure and compute the various screening indices. The relatively high rate of LQD allowed us to examine the simultaneous use of multiple screening techniques. On the other hand, we were unable to examine the effects of administration mode on LQD, as doing so would require additional samples using paper-and-pencil or offline computer administration of an identical questionnaire. The drawbacks of "crowdsourced" data collection are well-documented (Harms & DeSimone, 2015), but future LQD research should attempt to address differences in the prevalence and effects of LQD on results as a function of survey administration mode.

Another limitation is that the current study relied exclusively on cross-sectional self-report surveys. Although this methodology is representative of the majority of survey-based research in the social sciences, data screening techniques may apply equally well to other forms of data collection. For example, the use of data screens in longitudinal research (especially experience sampling methodology) may indicate the point at which a participant becomes bored with a study. Data screening can also be used to identify judges who may be exerting minimal effort when rating the behaviors or performance of others.

Relatedly, the self-report format of our data collection effort yielded a situation in which all items were susceptible to the influence of LQD. Although this is ideal for examining the performance of screening indices, it is not optimal for the purposes of examining the relationship of LQD with other constructs. Emerging research has demonstrated the relationship of LQD with personality, boredom proneness, and aggression (e.g., Bowling et al., 2016; Dunn et al., in press). Other research has demonstrated that response patterns such as differences in responses to positively and negatively worded items or individual susceptibility to data collection method may have predictive value (Chen, Watson, Biderman, & Ghorbani, 2016).

Additionally, the data analyzed in this study may not be optimal for some screening techniques. For example, the fact that all ACST items were positively scored mitigates the utility of the longstring screen (see, DeSimone et al., 2015). This scale characteristic also did not allow for the computation of psychometric antonyms. Additionally, the ACST contains 24 items, which may be too few to generate sufficient variability in IRV values (Dunn et al., in press). Some recently proposed screening techniques could not be included in the present analysis due to the nature of the sample and data. Guttman errors and IRT-informed response probabilities are gaining popularity (see, Curran, 2016; Niessen, Meijer, & Tendeiro, 2016). However, as these techniques are computed using various operationalizations of item difficulty and/or discrimination, they are most appropriate for use with unidimensional scales containing many items (Meijer, Monelaar, & Sijtsma, 1994). Absent a much larger sample or normative data, estimates of item parameters would be too unstable to be useful for computing these indices (De Ayala & Sava-Bolesta, 1999; DeMars, 2003). Niessen et al. (2016) examined the cutoff values and relationships of these indices with some of the techniques assessed in the current paper. Future research should examine the distributional characteristics of these techniques as well as their potential impact on item inter-relationships or the results of statistical analyses (research questions 3, 4, and 5 above).

It is important to understand the expected values of each data screen. To this end, the current analyses complement and extend Huang et al. (2012) in providing descriptive statistics and cumulative percentage distributions for each data screen. However, the limited results from existing literature are insufficient for producing normative data on each screening technique. There are still many unanswered questions about LQD and survey research in general. For example, how long should we expect a participant to spend answering a self-report question? What is the expected longstring value in a uniformly scored scale as opposed to a scale containing both positively and negatively scored items? How are different analytic techniques influenced by the various screening techniques? We hope that future research can address some of these questions in an effort to help researchers and practitioners make the most of their data.

Conclusions

This study addressed some important questions about LQD and data screening practices. A side-by-side examination of multiple screening techniques revealed some facts about the nature of data screening. Estimates of the prevalence of LQD in survey research are a function of both the screen(s) used and cutoff points chosen. The cutoff points found in previous literature perform well for some screens (self-report, bogus and instructed items, response time, longstring, and *D*). However, cutoff points for consistency-based techniques (e.g., psychometric synonyms, personal reliability) seem too liberal and are likely sensitive to sample and survey design characteristics. We encourage researchers to exercise caution when employing consistency-based screens until future research rigorously examines cutoff criteria for these indices.

Due to the small amount of overlap between screening indices, researchers should not hesitate to employ multiple screening techniques to identify LQD and screen participants prior to data analysis. Researchers are encouraged to supplement unobtrusive techniques with direct techniques whenever possible. The use of response time, longstring, or IRV and either self-report, bogus, or instructed items can assist researchers in identifying different forms of LQD in their data. However, we caution readers against selecting screening techniques arbitrarily or based on convenience. Instead, we encourage survey designers to thoughtfully consider data screening as part of the research design process.

Finally, the choice of data screening techniques can impact the performance of items and measures as well as the results of a study. While most techniques have negligible effects on inter-item and inter-scale correlation matrices, longstring, IRV, and *D* can noticeably impact these analyses. Screening techniques that discourage response homogeneity (e.g., bogus/instructed items, longstring, response time) can decrease alpha estimates, while screens that encourage response consistency (e.g., psychometric synonyms, personal reliability, *D*) can increase alpha estimates. Statistical results such as correlations and Cohen's *d* estimates can also change noticeably as the result of data screening. We encourage transparency in screening processes and urge researchers to report which screens were used and which cutoff values were selected. We also recommend that researchers report study results before and after screening.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*, 270–301.

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 72–90). Los Angeles: Sage.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*, 17–21.

Anscombe, F. J., & Guttman, I. (1960). Rejection of outliers. *American Society for Quality, 2*, 123–147.

Bagby, R. M., Gillis, J. R., & Rogers, R. (1991). Effectiveness of the Millon clinical multiaxial inventory validity index in the detection of random responding. *Psychological Assessment, 3*, 285–287.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research, 43*, 800–813.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

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.

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

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. *Journal of Personality and Social Psychology, 111*, 218–229.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Bruehl, S., Lofland, K. R., Sherman, J. J., & Carlson, C. R. (1998). The variable responding scale for detection of random responding on the multidimensional pain inventory. *Psychological Assessment, 10*, 3–9.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*, 112–130.

Chen, Z., Watson, P. J., Biderman, M., & Ghorbani, N. (2016). Investigating the properties of the general factor (M) in bifactor models applied to the big five or HEXACO data in terms of method or meaning. *Imagination, Cognition, and Personality: Consciousness in Theory, Research, and Clinical Practice, 35*, 216–243.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

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

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

Costa, P. T., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment* (pp. 179–198). London: SAGE.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Aggreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*, 151–174.

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

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3–31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3–19.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*, 275–288.

DeSimone, J. A. (2015). New techniques for evaluating temporal consistency. *Organizational Research Methods, 18*, 133–152.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171–181.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Nels, T. (in press). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. Journal of Business and Psychology. Available from https://link.springer.com/article/10.1007/s10869-016-9479-0. Accessed 11 April 2017.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166.

Frederiksen, N. (1965). Response set scores as predictors of performance. *Personnel Psychology, 18*, 225–244.

Gallen, R. T., & Berry, D. T. R. (1996). Detection of random responding in MMPI-2 protocols. *Assessment, 3*, 171–178.

Gallen, R. T., & Berry, D. T. R. (1997). Partially random MMPI-2 protocols: When are they interpretable? *Assessment, 4*, 61–68.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology, 8*, 183–190.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis. Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311.

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

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828–845.

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

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

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

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*, 34–47.

Littman-Ovadia, H., & Lavy, S. (2012). Character strengths in Israel. *European Journal of Psychological Assessment, 28*, 41–50.

Liu, M., Bowling, N. A., Huang, J. L., & Kent, T. A. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist, 51*, 32–38.

Liu, M., Huang, J. L., Bowling, N. A., & Bragg, C. (2013). *Attenuating effect of insufficient effort responding on relationships between measures*. Paper presented at the 28th annual conference for the Society for Industrial and Organizational Psychology, Houston, TX.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*, 49–55.

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.

McCrae, R. R., Costa, P. T., Grant, W., Barefoot, J. C., Siegler, I. C., & Williams Jr., R. B. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine, 51*, 58–65.

McGrath, R. E., Mitchell, M. K., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450–470.

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

Meijer, R. R., Monelaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Messick, S. (1960). Dimensions of social desirability. *Journal of Consulting Psychology, 24*, 279–287.

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.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1–11.

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

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Paolo, A. M., & Ryan, J. J. (1992). Detection of random response sets on the MMPI-2. *Psychotherapy in Private Practice, 4*, 1–8.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. New York, NY: Oxford University Press.

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

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Ruch, W., Proyer, R. T., Harzer, C., Park, N., Peterson, C., & Seligman, M. E. P. (2010). Values in action inventory of strengths (VIA–IS): Adaptation of the German version and the development of a peer-rating form. *Journal of Individual Differences, 31*, 138–149.

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

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

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resources Management Review, 9*, 219–242.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

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

Vanhove, A. J., Harms, P. D., & DeSimone, J. A. (2016). The abbreviated character strengths test (ACST): A preliminary assessment of test validity. *Journal of Personality Assessment, 98*, 536–544.

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

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

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

Yu, C. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation. Retrieved from http://statmodel2.com/download/Yudissertation.pdf on February 18, 2014.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551–563.