Journal of Research in Personality 48 (2014) 61–83

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Caring about carelessness: Participant inattention and its effects on research

Michael R. Maniaci, Ronald D. Rogge *

Department of Clinical and Social Sciences in Psychology, University of Rochester, United States

ARTICLE INFO

Article history:
Available online 7 October 2013

Keywords:
Attentiveness
Random responding
Inattentive responding
Careless responding
Validity scales

ABSTRACT

The current studies examined the adverse effects of inattentive responding on compliance with study tasks, data quality, correlational analyses, experimental manipulations, and statistical power. Results suggested that 3–9% of respondents engaged in highly inattentive responding, forming latent classes consistent with prior work that converged across existing indices (e.g., long-string index, multivariate outliers, even–odd consistency, psychometric synonyms and antonyms) and new measures of inattention (the Attentive Responding Scale and the Directed Questions Scale). Inattentive respondents provided self-report data of markedly poorer quality, sufficient to obscure meaningful regression results as well as the effects of experimental manipulations. Screening out inattentive respondents improved statistical power, helping to mitigate the notable drops in power and estimated effect sizes caused by inattention.

1. Introduction

An oft-neglected issue underlying much research is that not all respondents pay sufficient attention when completing self-report measures. Such responding could introduce error into a dataset, potentially decreasing power and obscuring results. Inattention is sometimes blatant and easy to address (e.g., removing participants from analyses if they exhibit suspiciously fast reaction times or below-chance performance on a task). This logic, however, is only rarely applied to inattentive responding on self-report scales as this form of inattention is more subtle and therefore more difficult to measure. The current studies were designed to identify excessive inattention using a multi-method approach, exploring its impact on: (1) compliance with common study tasks, (2) quality of self-report data, (3) correlational and experimental analyses and (4) statistical power, in order to determine the potential scope of this problem and the degree to which addressing it might improve statistical analyses. Estimated rates of inattention in the existing literature have varied widely, from 3% to 46% of respondents (e.g., Berry et al., 1992; Johnson, 2005; Meade & Craig, 2012; Oppenheimer, Meyvis, & Davidenko, 2009). In part, this wide range of estimates is due to a lack of clarity on how best to measure inattentive responding and on what thresholds correspond to unacceptably error-ridden data. The present research used indicators of non-compliance, data quality, and statistical power as criteria for comparing methods of measuring inattention and establishing concrete, practical thresholds for researchers to use to screen their data.

1.1. Forms of non-compliance

It is common for a small portion of participants to exhibit poor attention and effort in research. For example, subjects with excessively short reaction time latencies on implicit measures like the Implicit Association Test (IAT; Greenwald, Nosek, & Banaji, 2003) are routinely excluded from analyses. Although such practices are common in research using reaction time paradigms and experimental manipulations, this logic has not typically been extended to research utilizing self-report methods. This discrepancy is likely not because researchers believe that experimental manipulations or reaction time measures are more prone to non-compliance than self-report measures; rather, it is simply easier to identify non-compliance on such tasks. As researchers typically do not screen for inattention on self-report scales, the prevalence and impact of such problematic responding is largely unknown.

1.2. Inattentive responding as a distinct construct

Although a number of constructs are occasionally lumped together under the heading of *validity scales* (e.g., socially desirable responding, faking good, faking bad, random responding), the present research focuses on a specific form of invalidity: inattention

* Corresponding author. Address: Department of Clinical and Social Sciences in Psychology, University of Rochester, Meliora Hall, RC Box 270266, Rochester, NY 14627-0266, United States.

E-mail address: rogge@psych.rochester.edu (R.D. Rogge).

0092-6566/$ - see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jrp.2013.09.008

when completing self-report measures. This form of inattention is distinct from other types of invalidity. For instance, the response sets of faking good, faking bad and social desirability imply a motivation to present oneself in a particular manner. Ironically, these forms of invalidity may be *negatively* related to inattentive responding because presenting oneself in a particular manner requires carefully attending to questions (Meade & Craig, 2012). In contrast, inattentive responding corresponds to a lack of motivation to present oneself in a certain manner, and should therefore contribute little more than error variance to analyses. Extreme levels of inattention conceptualized in this manner are consistent with the extremely inattentive latent class identified by Meade and Craig (2012) as comprising approximately 9% of an undergraduate sample, and with what Nichols, Greene, and Schmolck (1989) called "content nonresponsivity." Although inattention could be correlated with individual differences, in the current study we view inattentive responding as a proximal behavior enacted during the completion of research studies. We therefore conceptualize it as more of a transitory (state) phenomenon, allowing for the possibility that the same individual might provide high levels of attention in one study (e.g., a short and particularly interesting study) but insufficient levels of attention in other studies.

1.3. Approaches to measuring inattention

1.3.1. Infrequency and inconsistency scales

Much of the work examining inattentive responding on self-report measures has been conducted in the development of clinical assessment batteries like the Personality Assessment Inventory (PAI; Morey, 1991) and the Minnesota Multiphasic Personality Inventory (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The length of these inventories (typically containing several hundred items) demands high levels of sustained attention from respondents, necessitating the development of scales designed to identify problematic responding. Two types of validity scales within these clinical batteries (infrequency and inconsistency scales) assess inattentive responding, with a focus on identifying extreme (and therefore problematic) levels. Infrequency scales (e.g., the infrequency scale of the PAI, the "bogus" items of Meade & Craig, 2012) are made up of items that elicit nearly identical (highly skewed) responses from most respondents (e.g., "I have been to every country in the world"). Respondents receive higher scores on infrequency scales for each increasingly unlikely response across the set of items, and cut-scores are used to identify inattentive responding excessive enough to yield an invalid protocol. Inconsistency scales (e.g., the VRIN scale of the MMPI, the inconsistency scale of the PAI) are made up of pairs of items with nearly identical content that are presented in opposing halves of a survey (e.g., "I am an active person" paired with "I have an active lifestyle"). Absolute differences in responses are summed across the item pairs so that higher scores reflect more inconsistent responding. Thus, although these scales make use of self-report items, they do not ask subjects to report on their own levels of attention but instead use their responding behavior on a set of heterogeneous items to indirectly assess their attentiveness to item content. Such scales are effective at distinguishing randomly generated data from actual data (e.g., Bruehl, Lofland, Sherman, & Carlson, 1998; Pinsoneault, 2005), but have only rarely been implemented outside of the clinical instruments for which they were designed (e.g., Saavedra, Chapman, & Rogge, 2010).

1.3.2. Instructional manipulation check

More recently, Oppenheimer and colleagues (2009) developed the instructional manipulation check (IMC), a single item measure presented as a separate page in an online survey that uses critical instructions embedded at the end of a lengthy paragraph to assess participants' attentiveness to instructions. Although the IMC moderated the effectiveness of text-based manipulations, its usefulness is minimized by the fact that it only measures one form of inattention (skipping instructions) which is relatively common and therefore identifies a high proportion of participants (35–45%) as inattentive. As Oppenheimer and colleagues note, eliminating that many participants could potentially reduce power and bias results, leading them to suggest using the IMC as an intervention to encourage attentiveness rather than as a measure of inattention. Despite that suggestion, published studies using the IMC have simply excluded high proportions of inattentive respondents (e.g., Simmons & Nelson, 2006, Study 12). This highlights the need for a measure of attention with greater variability and specificity, potentially allowing for a smaller proportion of participants to be identified as excessively inattentive.

1.3.3. Post hoc indices

In addition to adding measures in order to identify highly inattentive respondents, researchers can also calculate post hoc indices of inattention using virtually any body of self-report items after data have been collected. Meade and Craig (2012) examined the convergence of several such indices, including highly correlated item-pairs (psychometric synonyms and antonyms assessing the consistency of responding), even–odd consistency (split-half reliabilities measured within respondents across scales), multivariate outlier distances (assessing statistically unlikely response patterns), long string analyses (measuring the tendency to choose identical answers in blocks of items), and time spent on the survey. Latent profile analyses of these indices along with an infrequency scale identified two main types of inattention: one reflecting more general inattentive responding and another marked by subjects frequently selecting the same answer for entire blocks of questions and consequently completing the survey in suspiciously short periods of time. Their analyses suggested an overall frequency of 10–12% of inattentive respondents in a sample of undergraduates.

1.4. Effects of inattention

Careless or inattentive responding might act as a source of measurement error that could obscure meaningful results. Identifying and removing inattentive respondents before data analysis could therefore offer a relatively easy method of decreasing error variance and increasing statistical power in research using self-report measures. Consistent with this assertion, Oppenheimer et al. (2009) found that previously reported experimental results with two different manipulations did not replicate in a group of participants identified as inattentive, although they did replicate among attentive participants. Similarly, inattentive responding can adversely affect correlational and factor analyses (Johnson, 2005; Meade & Craig, 2012; Woods, 2006), moderating findings even to the point of generating spurious results.

1.5. Addressing challenges to validity measures

In contrast to these findings, Piedmont, McCrae, Riemann, and Angleitner (2000) challenged the utility of validity scales as a set, demonstrating that scores on a diverse array of 13 different validity measures (including measures of inattention as well as distinct constructs, such as social desirability, faking good, faking bad) failed to moderate substantive results when formed into a heterogeneous composite. The methodological decisions underlying their non-significant results highlight four fundamental elements of the current approach. First, the current work focused exclusively on inattentive responding, as it is not clear that all forms of invalid responding would have equivalent effects on data quality and statistical power. Consequently, forming composites of disparate validity indices could obscure results. Second, the current work adopted a taxonometric approach focused on identifying only the most extreme inattentive respondents (whose inattention would have the most pernicious effects on data quality). This stands in contrast to prior work that either modeled inattentive responding as a continuous variable or that identified large proportions of the sample as problematic (e.g., Oppenheimer et al., 2009; Piedmont et al., 2000), potentially underestimating the effects of inattention on data quality (as only extremely high levels of inattention are likely to introduce sufficient randomness to obscure meaningful results). Third, the current work collected sufficiently large samples so that a reasonable number (e.g., 30–40) of excessively inattentive respondents could be identified and analyzed separately. Finally, the current work addressed a critical gap in the literature by specifically examining the practical utility of screening out extreme inattention (examining its effects on statistical power). Removing inattentive respondents reduces sample size, and prior research has not yet demonstrated that this data cleaning approach increases statistical power.

1.6. The current studies

The studies presented in this paper sought to extend prior work by exploring the correlates of inattention and examining the effects of inattention on compliance with study tasks, data quality, correlational analyses, experimental manipulations, and statistical power. Toward those ends, the studies also sought to develop effective methods of identifying extreme inattention, establishing practical thresholds of unacceptable inattention based on indices of non-compliance, data quality, and statistical power. Analyses across all three samples: (1) examined personality and motivational correlates of inattention, (2) evaluated the convergent validity of various indicators of inattention and non-compliance, (3) replicated the latent classes of inattention from Meade and Craig (2012), (4) examined the impact of inattention on data quality (i.e., internal consistency of scales, substantive correlational and experimental results) and statistical power, and (5) examined the potential gains in power afforded by the use of various inattention indices when cleaning data.

2. Study 1

2.1. Multi-method approach

One goal of Study 1 was to augment recent work investigating the prevalence of inattentive responding (Meade & Craig, 2012), taking a multi-modal approach that included assessing self-reported responding styles, previously published or new indices of inattention (multivariate distances, long-string indices, psychometric synonyms and antonyms, even–odd consistency, infrequency and inconsistency scales, directed questions), and 3 indicators of compliance with tasks common to psychological research (watching a video, spending sufficient time completing study tasks, completing a pronoun-identification priming task) to cross-validate estimated levels of inattention. We hypothesized that these various methods of assessing inattention would show moderate agreement and would conceptually replicate the latent classes of inattentive responding observed by Meade and Craig (2012).

2.2. The Attentive Responding Scale (ARS)

Given the concerns that have been raised about the utility of existing measures of inattentive responding (e.g., Piedmont et al., 2000), Study 1 also sought to validate a broadband measure of inattentive responding (consisting of an infrequency and an inconsistency subscale, consistent with prior approaches) that could offer researchers a parsimonious method of screening their datasets. Toward this end, a pool of 67 potential items (written for this project based on the style of existing inconsistency and infrequency scales) were evaluated in separate batches (prior to the current work) across 9 online samples comprising a total of 13,780 respondents. As mentioned above, although this approach involves self-report items, the items do not ask participants to report their own attentiveness. Rather, the items ask about unrelated content and it is the participants' response behavior on these items (e.g., selecting incredibly unlikely or inconsistent answers) that *indirectly* indicates higher levels of inattention.

As the current work took a taxonometric approach (focused on creating a tool to identify extreme cases of inattentive responding), we selected items based solely on their utility. The items were initially screened to select infrequency items with the greatest skew, inconsistency item-pairs with the highest levels of agreement, and all items for their relative abilities to distinguish between actual responses and computer-generated random data (simulating their abilities to identify extreme inattention). Finally, when possible, items were selected for more subtle content (e.g., "It feels good to be appreciated" rather than "My favorite hobbies are coin collecting and interpretive dance") to help mask the nature of the scale. This multi-stage screening process yielded a final set of 33 items (22 inconsistency items forming 11 item-pairs and 11 infrequency items, described below) that were heterogeneous in item content but unified by their joint ability to identify extreme inattention. Thus, instead of developing continuous scales of inattention (with a focus on internal consistency/factor structure), the development of the ARS focused on validating cut-scores for the purposefully diverse set of items shown to be most effective at identifying extreme inattention. This effectively rendered correlational analyses like internal consistency and factor analysis irrelevant given the lack of any correlational assumptions underlying this taxonometric approach. Because these two subscales measure inattention with different strategies (focusing on highly atypical or highly inconsistent responses, respectively), we analyzed them separately in addition to examining their joint effects.

2.3. Correlates of inattention

Study 1 also examined personality, demographic, and motivational correlates of inattentive responding to help characterize the types of individuals most likely to engage in this behavior. We hypothesized that the personality traits of conscientiousness and agreeableness would be associated with lower levels of inattentive responding given the attention to detail, helpfulness and willingness to please associated with these two traits (e.g., Digman, 1990). We further hypothesized that internal motivation for participating in the study (e.g., out of intrinsic interest) would be associated with lower levels of inattention whereas external motivation (for course credit or money) would be associated with higher levels of inattention.

2.4. Impact of inattention

Finally, Study 1 sought to examine the effects of inattention on research studies. We hypothesized that excessively inattentive respondents would show higher levels of non-compliance on common study tasks, provide self-report data of poor quality (as assessed by internal consistencies of well-validated scales), and fail to demonstrate regression findings replicated in the attentive respondents.

2.5. Study 1 method

2.5.1. Participants and procedure

Individuals had to be at least 18 years old and report having participated in previous online research to participate. A total of 674 individuals responded to the online survey, and were predominantly female (70%) and Caucasian (75%), 11% Asian, and 5% African American. Participants reported an average of 14.4 years of education (SD = 2.3) with a median household income range of $50,000–59,999. The mean age was 29.5 years (SD = 12.1). Participants were recruited from Mechanical Turk (46%), other online forums and advertisements (27%) and from a psychology participant pool with the incentive of extra credit (27%). Participants from Mechanical Turk were offered small monetary recruitment incentives (25 cents) for participating. All participants were also entered into a lottery for $50. All procedures and materials for this study (and for Studies 2, 3, and 4) were approved by an Institutional Review Board.

2.5.2. Measures of inattention

2.5.2.1. ARS infrequency subscale. A set of 11 infrequency items were worded to obtain highly skewed response distributions with a purposefully diverse range of content across the items to ensure that scores would not be strongly influenced by having a single atypical attitude. Participants rated each item on a 5-point Likert response scale (*not at all true* to *very true*) and answers were coded so that higher scores indicated increasingly less likely responses. Half of the infrequency items were presented near the beginning of the survey and the remaining items were presented near the end of the survey (the Supplementary appendix includes the complete set of items and scoring information).

2.5.2.2. ARS inconsistency subscale. A set of 22 inconsistency items (grouped into 11 pairs of items) were written such that the items within each pair were near identical in content (the Supplementary appendix and Table S1 include all item pairs). Item pairs were chosen to represent diverse content so that scores averaged across multiple item pairs would not be strongly influenced by ambivalence or inconsistency regarding any single topic. The two items from each pair were presented in opposite halves of the survey. Item pairs were scored by summing the absolute difference between paired responses so that high scores indicated greater inconsistency in responding.

2.5.2.3. Directed Questions Scale (DQS). To borrow yet another approach to identifying inattention, we wrote a set of 7 questions directing subjects to give specific answers (e.g., "*Please skip this question*," "*I read instructions carefully. To show that you are reading these instructions, please leave this question blank*," "*This is a control question. Mark 'Mostly True' and move on*"). The directed questions were distributed throughout the survey, with one directed question appearing on approximately every other page of the survey, embedded within blocks of items from other substantive scales. This scale was scored by summing the number of mistakes each subject made on these items to create scores ranging from 0 to 7.

2.5.3. Post hoc indices of inattention

2.5.3.1. Long string index. Following previous research (e.g., Johnson, 2005; Meade & Craig, 2012), we calculated the longest string of identical answers across the 44 items of the BFI for each respondent as a measure of patterned responding (i.e., repeatedly choosing the same answer regardless of item content).

2.5.3.2. Statistical indices of inattention. Following established methodology (e.g., Johnson, 2005; Meade & Craig, 2012), we calculated four standard statistical indices of inattention. First, **multivariate distances** were calculated on the 44 items of the Big Five Inventory (BFI) along with the 10 items of the Rosenberg Self Esteem Scale (RSES), described below. Second, the six distinct subscales of the BFI and RSES measures were each split into **even and odd** halves based on the order of item presentation within the survey. We then calculated averages for all of the scale halves, computed within-person correlations between those scale halves across the 6 scales, and applied the Spearman–Brown split-half formula. Third and fourth, **psychometric synonym** and **psychometric antonym** indices were created by using sets of item-pairs within the BFI and RSES scales showing the strongest positive (5 pairs with $r \geqslant .64$) and negative (5 pairs with $r \leqslant -.49$) correlations. For each index, within-person correlations across the 5 item pairs were calculated for each subject, reversing the direction of those correlations for the antonym pairs so that higher scores indicated greater consistency for both indices.

2.5.4. Measures of non-compliance with study tasks

2.5.4.1. Mistakes on pronoun task. To behaviorally assess how carefully subjects comply with a typical priming task, we asked participants to complete a task that has been used in previous research to prime collectivistic self-construal (e.g., Gardner, Gabriel, & Lee, 1999). Respondents read a paragraph about a trip to a city and clicked on the 20 pronouns it contained. The task was designed to be easy to complete and listed the specific words the subject was supposed to identify (e.g., *we*, *us*, *our*) within the instructions. The number of errors (both missing pronouns and clicking on words other than pronouns) was summed for each subject to reflect their lack of compliance.

2.5.4.2. Seconds of video skipped. At the end of the survey, participants watched a 2-min video clip displaying an abstract screen saver. Similar videos have been used as control conditions in studies manipulating affect (e.g., Gross & Levenson, 1995; Rottenberg, Ray, & Gross, 2007). The website surreptitiously recorded how long participants kept the video open before continuing onto the next page.

2.5.4.3. Time spent completing survey. The final page of the survey was programmed to record the time that had elapsed since the first page of the survey was opened by each participant. Although the vast majority of respondents completed the survey in under an hour, a small proportion of individuals (7.1%) took longer than 60 min (most likely due to leaving the survey and returning to complete it later). As these times represented univariate outliers, we truncated them to 60 min. Using these cleaned times, respondents took an average of 28.9 min (SD = 12.8) to complete the survey.

2.5.5. Self-report measures

2.5.5.1. Self-reported responding styles. Thirteen items were used to assess general tendencies in responding. Five items assessed **Self-Reported Careless Responding** (e.g., *how often do you: read each question carefully, pay attention to every question, take as much time as you need to answer the questions honestly* – all reverse coded). Three items assessed **Self-Reported Patterned Responding** (e.g., *how often do you: make patterns with the responses to a block of questions, use the same answer for a block of questions on the same topic [rather than reading each question]*). Three items assessed **Self-Reported Rushed Responding** (*how often do you: answer quickly without thinking, answer impulsively without thinking, rush through the survey*). Finally, two items assessed **Self-Reported Skipping of Instructions** in surveys (*how often do you: skim the instructions quickly, skip parts of the instructions*). Items were rated on a 7-point scale (1 = *never*, 4 = *about half the time*, 7 = *all of the time*), responses were averaged so that high scores reflected more problematic responding, and the scales demonstrated reasonable internal consistency ($\alpha_{careless}$ = .85, $\alpha_{patterned}$ = .86, α_{rushed} = .88, $\alpha_{skipping}$ = .79).

2.5.5.2. Self-reported motivations for participating in research. A set of 10 items used the same stem (*please rate the extent to which each of the following items is generally a reason you have participated in research*) to assess motivations for participating in psychology studies. Five of those items assessed **Internal Motivation** (*supporting research, gaining insight about myself, gaining insight about others or the world in general, for fun, general interest in the research*). Two items assessed participating **For Money** (*payment, entry into a lottery*). Two items assessed participating **Out of Boredom** (*just to mess around, to kill time*). Finally, a single item assessed participating **For Extra Credit** (*extra credit in a course*). Items were rated on a 6-point scale (*not at all* to *extremely*), and responses were averaged so that higher scores indicated higher levels of each type of motivation ($\alpha_{internal} = .88$, $\alpha_{money} = .68$, $\alpha_{boredom} = .78$).

2.5.5.3. Personality. The 44-item Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) was included to assess the constructs of extraversion, neuroticism, conscientiousness, openness to experience and agreeableness. The items were rated on 5-point scales, were averaged so that higher scores indicated higher levels of each personality trait, and demonstrated high levels of internal consistency (see Table 6).

2.5.5.4. Self-esteem. The 10-item Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) was included to assess self-esteem. The items were rated on 4-point scales, were averaged so that higher scores indicated higher levels of self-esteem, and demonstrated moderate levels of internal consistency (see Table 6).

2.5.5.5. Desirable responding. The 20-item Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984) was included to assess the constructs of self-deception and impression management. The items were rated on 7-point scales and, following the standard scoring method, extreme scores were given a 1-point value for each item and then summed to create total scores ranging from 0 to 10 for each scale.

2.6. Study 1 results and discussion

2.6.1. Self-reported responding styles

Descriptive statistics for the self-reported responding styles are presented on the right side of Table 2. As seen in the row of Table 2 labelled "Any problems," 36–84% of participants self-reported engaging in various problematic responding behaviors at least some of the time and, as seen in the row labelled "Highly inattentive," 3–19% of participants selected a response indicating that they engaged in these forms of problematic responding more than half of the time while completing self-report measures.

2.6.2. Measures of inattention and non-compliance

The measures of inattention and non-compliance demonstrated convergent validity with the self-report measures as they not only correlated significantly in the expected directions (see Table 1), but also yielded similar estimates of extreme inattention (Table 2). Consistent with previous work (e.g., Meade & Craig, 2012), a small proportion of participants (3–7%) exhibited notably high[1] levels of inattentive responding (e.g., missing more than half of the control questions or pronoun items, multivariate outliers).

2.6.3. Predictors of inattention

As hypothesized, the various indices of inattention and noncompliance were associated with lower levels of agreeableness and conscientiousness (Table 2). Exploratory analyses revealed that inattentive participants also tended to report less openness to experience and lower self-esteem. Finally, higher levels of inattention and noncompliance were associated with lower levels of social desirability and impression management, suggesting that inattention is a distinct construct from socially desirable responding, representing a novel source of error variance.

Consistent with our hypotheses, inattention was associated with lower reports of internal motivation and higher reports of external motivation (e.g., to earn extra credit). This is consistent with research indicating that response rates (another form of compliance) tend to be higher when people find the study more intrinsically interesting (see Edwards et al., 2002 for a review). Thus, motivations underlying participation predicted levels of inattention.

Inattention was also associated (albeit less consistently across indices) with several demographic variables. For instance, rates of inattention were higher (on one or more indicators) for respondents who were younger, less educated, or male. Participants recruited from the psychology participant pool did not differ from participants recruited via the Internet on the various indicators of inattention and non-compliance, although they did self-report greater inattentive responding and skipped more of the video than other participants.

2.6.4. The Attentive Responding Scale

2.6.4.1. Item-level analyses. The 11 infrequency items of the ARS demonstrated extremely high skew (with skew/SE ranging from 17.3 to 39.7), suggesting that respondents typically selected the same one or two answer choices on those items. The 11 infrequency items and the 22 inconsistency items (forming 11 item-pairs) also yielded strong effect sizes at identifying randomly generated data[2] (Cohen's *ds* ranging from 0.72 to 1.77). These items continued to demonstrate these psychometric properties in Studies 2 and 3. A summary of the obtained values is included in Table S1 in the online Supplementary material. In addition to the 33-items used in the ARS-33, we also created an 18-item version of the scale (ARS-18; containing 6 infrequency items and 12 inconsistency items (forming 6 item-pairs) that would add only about 1.5 min to a study) based on results from all three studies, offering researchers a shorter alternative. When selecting items for the ARS-18, we used the utility of each item/item-pair as our primary criteria. Thus, we took into account each item's ability to discriminate random vs. actual data, the direction of items (so as to have a balanced scale with items showing typically high or low responses), and the ability of each item to increase statistical power (based on resampling analyses similar to those presented in Study 3). Given the measurement benefits of longer scales (greater precision, lower noise; e.g., Funk & Rogge, 2007), we expected the longer ARS-33 to more accurately assess extreme inattentive responding and therefore to provide stronger improvements in data quality when used to screen datasets. Nonetheless, we also examined the ARS-18 (as well as the 6-item and 11-item infrequency subscales of the ARS-18 and ARS-33,

[1] To establish cut-scores for problematic inattention on the post hoc indices, we began with cut-scores established in the existing literature (e.g., Huang, Curran, Keeney, Poposki, & DeShon, 2012; Johnson, 2005), then cross-validated these suggested cut-scores using power analyses in Study 3, choosing cut-scores that optimized power for detecting manipulation effects. For 4 of these indices (multivariate distances, long string index, psychometric synonyms, and even-odd consistency) the power analyses suggested cut-scores that converged with the previous literature. On the remaining indices of inattention and non-compliance, we set the cut-scores to reflect inattention/non-compliance rates of $\geqslant 50\%$. The specific cut-scores are presented in Table 2.

[2] To evaluate the ability of the items to detect randomly generated data, a sample of 674 lines of random data (representing 50% of the resulting sample) was generated using a random number generation algorithm designed to produce random numbers with uniform distributions (using the same numeric response options presented to actual participants for each item).

Table 1
Correlations among measures of inattention and non-compliance in Study 1.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. ARS-33 inconsistency	1.00															
2. ARS-33 infrequency	.31**	1.00														
3. ARS-18 inconsistency	.81**	.18**	1.00													
4. ARS-18 infrequency	.28**	.94**	.17**	1.00												
5. Mistakes on 7 directed Qs	.26**	.72**	.14**	.65**	1.00											
6. Multivariate distances	.25**	.07	.23**	.05	.11**	1.00										
7. Long string index	.03	.33**	.02	.31**	.32**	−.15**	1.00									
8. Psychometric synonyms	−.19**	−.27**	−.14**	−.24**	−.19**	−.26**	−.04	1.00								
9. Psychometric antonyms	−.14**	−.20**	−.07	−.16**	−.23**	−.08	−.34**	.19**	1.00							
10. Even–odd consistency	−.22**	−.22**	−.19**	−.22**	−.19**	−.19**	−.13**	.27**	.21**	1.00						
11. Mistakes on pronoun task	.25**	.48**	.20**	.45**	.53**	.14**	.18**	−.17**	−.11**	−.23**	1.00					
12. Seconds of video skipped	.12**	.33**	.09*	.29**	.36**	.12**	.17**	−.16**	−.18**	−.12**	.24**	1.00				
13. Time spent on survey	−.08*	−.13**	−.05	−.11**	−.12**	−.06	−.05	.05	.05	.08*	−.14**	−.19**	1.00			
14. SR inattentive responding	.15**	.36**	.06	.34**	.42**	.01	.32**	−.15**	−.17**	−.18**	.22**	.34**	−.12**	1.00		
15. SR patterned responding	.17**	.49**	.10*	.43**	.53**	.03	.36**	−.16**	−.26**	−.20**	.35**	.39**	−.08*	.51**	1.00	
16. SR rushed responding	.08*	.24**	−.02	.21**	.28**	−.02	.15**	−.10*	−.11**	−.13**	.14**	.26**	−.10*	.56**	.52**	1.00
17. SR skipping instructions	.11**	.21**	.06	.19**	.23**	.10*	.11**	−.05	−.07	−.17**	.09*	.23**	−.03	.49**	.44**	.50**

Note: SR = self-reported.
* $p < .05$.
** $p < .01$.

respectively, and the 7-item DQS) to see if shorter measures might also perform well as tools to screen for extreme inattention.

2.6.4.2. Establishing cut-scores and identifying random data. Consistent with our taxonometric focus, our analyses validating the ARS and DQS scales focused on the utility of those scales for identifying extreme levels of inattention (validating optimized cut-scores and demonstrating the utility of those cut-scores) rather than the underlying correlational structure of those scales (e.g., internal consistency, factor structure) as we make no assumptions that the items will show strong correlations.[3] Toward this end, the ARS-33 and ARS-18 infrequency items were summed to create infrequency totals ($M_{ARS\text{-}33} = 3.35$, $SD_{ARS\text{-}33} = 4.47$; $M_{ARS\text{-}18} = 1.93$, $SD_{ARS\text{-}18} = 2.58$) and the absolute differences within inconsistency item pairs were summed to create inconsistency totals ($M_{ARS\text{-}33} = 5.24$, $SD_{ARS\text{-}33} = 2.71$; $M_{ARS\text{-}18} = 2.80$, $SD_{ARS\text{-}18} = 1.76$). Scores could theoretically range from 0 to 44 (for each ARS-33 subscale) and from 0 to 24 (for each ARS-18 subscale). Cut-scores were calculated for each subscale using response-operating curves (ROCs)[4] to obtain roughly the expected proportion of actual respondents identified as inattentive while maximizing the proportion of random data correctly identified (hit rate). The ROCs yielded cut-scores of 11.5 (infrequency) and 10.5 (inconsistency) for the ARS-33 (1.8 and 1.9 *SD*s above the mean), and cut-scores of 7.5 (infrequency) and 6.5 (inconsistency) for the ARS-18 (2.2 and 2.1 *SD*s above the mean). Across all three samples, these cut-scores were further validated by: additional ROC analyses in Study 2 yielding identical results, substantial effect sizes when comparing ARS attentive and inattentive respondents on the other indices of inattention/non-compliance, and resampling power analyses in Study 3. Mistakes on the DQS were summed so that scores on the DQS could theoretically range from 0 to 7 ($M = 0.32$, $SD = 1.00$). Similar ROC analyses across the three studies suggested a cut-score of 3 or more mistakes (2.7 *SD*s above the mean) to identify excessively inattentive respondents on the DQS.

Using these cut-scores, the ARS-33 and ARS-18 subscales identified 7.9% and 7.7% participants as potentially problematic due to

[3] It is worth noting that the items (and item pairs) of the ARS are actually indirect measures of a behavior that we are trying to capture – specifically a behavior that is completely unrelated to the actual content of those items. Because we are interested in inattentive responding rather than the responses to the items themselves, we purposefully wrote items with heterogeneous content. As a result, we did not anticipate responses on these heterogeneous items to correlate with one another as would be typical of most self-report scales. Furthermore, as we are interested only in identifying the most extreme cases of inattentive responding, we specifically selected infrequency items that demonstrated extreme levels of skew. Thus, the nature of these items make typical correlational methods for validating measures (e.g., factor analysis, corrected item-to-total correlations, Cronbach's alpha coefficients) less appropriate. As a result, we relied on the utility of each item or item-pair (the criterion validity) to identify the final ARS items and focused on validating cut-scores against concrete criteria (e.g., power gains) instead of using correlational methods to further validate the scales. Despite this taxonometric approach, the ARS infrequency and inconsistency scales demonstrated reasonable internal consistency in the highly inattentive respondents (i.e., those exceeding the cut-score for either subscale, who we would expect to show meaningful variance on these items; $\alpha_{\text{infrequency}} = .83$, $\alpha_{\text{inconsistency}} = .64$).

[4] ROCs are generated by classifying some status of interest (in this case randomly generated data) using every possible cut-score on a scale (in this case the ARS), and calculating the associated sensitivity (e.g., proportion of random data actually identified as invalid by the ARS) and specificity (e.g., proportion of non-random data actually identified as valid) for each of those cut-scores. By plotting the sensitivity by 1 – specificity, it is possible to identify the cut-score that minimizes classification mistakes (optimizing both sensitivity and specificity).

Table 2
Behaviorally measured and self-reported inattentiveness in Study 1.

	Existing indicators of inattention						Non-compliance with tasks			Self-reported responding styles			
	Mistakes on directed questions	Multivariate distances	Long string index	Psychometric synonyms	Psychometric antonyms	Even–odd consistency	Mistakes on pronouns	Seconds of video skipped	Time spent on survey	Careless	Patterned	Rushed	Skipping instructions
Descriptives													
Observed range	0–6	14–174	2–44	−1 to 1	−1 to 1	−1 to 1	0–23	0–118	6–60	1–6	1–7	1–7	1–7
Mean	.32	53.48	4.51	.73	.42	.76	2.24	16.37	28.91	1.96	1.40	2.24	2.62
SD	1.00	20.70	4.19	.34	.45	.33	4.57	33.82	12.76	.85	.82	1.03	1.36
Any problems (%)	14	–	–	–	–	–	55	25	–	81	36	83	84
Highly inattentive[a] (%)	5	4	6	4	3	7	6	14	6	4	3	8	19
Correlations with individual differences													
Extraversion	.00	−.13**	−.03	−.04	.21***	−.02	.04	−.03	−.01	−.01	−.06	.00	−.01
Agreeableness	−.21***	−.25***	−.17***	.18***	.13**	.18***	−.14***	−.10*	.13**	−.25***	−.16***	−.20***	−.17***
Conscientiousness	−.18***	−.23***	−.17***	.18***	.15***	.17***	−.09*	−.14***	.01	−.28***	−.24***	−.26***	−.18***
Neuroticism	.04	.23***	.08†	−.04	−.22***	−.16***	.02	.04	.02	.11**	.13***	.16***	.14***
Openness	−.19***	−.13**	−.11**	.12**	.15***	.24***	−.14***	−.17***	.08*	−.25***	−.18***	−.14***	−.12**
Self-deception	−.16***	.05	−.12**	.13**	.22***	.07†	−.06	−.11**	.03	−.30***	−.15***	−.28***	−.15***
Impression Mgmt	−.17***	.06	−.18**	.12**	.10*	.14***	−.07†	−.12**	.13**	−.34***	−.20***	−.24***	−.18***
Self-esteem	−.19***	−.28***	−.12**	.24***	.26***	.06	−.12**	−.14***	.02	−.22***	−.21***	−.19***	−.12**
Correlations with demographics and recruitment source													
Age	−.10**	−.02	−.09*	.06	.03	.09*	.00	−.15**	−.02	−.25**	−.12**	−.22**	−.18**
Years of education	−.03	−.02	−.06	.10**	.09**	.07	.04	−.06	−.10**	.00	−.03	.03	.03
Female	−.05	−.03	.01	.09*	.01	.04	−.06	−.08*	−.03	.02	−.06	.01	.04
Non-white	−.15***	−.16**	−.11**	.11**	.05	.17**	−.12**	−.13**	−.11**	−.12**	−.24**	−.09*	−.12**
UG subject pool	−.01	−.03	.02	.00	−.01	−.07	−.05	.10*	.12**	.24**	.06	.22**	.19**
Correlations with motivations for participating													
Internal motivation	−.02	.02	−.08*	.02	.04	.01	−.02	−.09*	.09*	−.27***	−.05	−.14***	−.06
For money	.04	.02	.04	−.02	.00	−.04	.05	−.01	−.10**	−.05	.11**	.06	−.03
Out of boredom	.18***	.02	.07†	−.12**	−.08*	−.17***	.14***	.07†	−.09*	.04	.17***	.11**	.15***
For extra credit	.11**	−.02	.03	−.03	−.02	−.08*	.04	.10**	.12**	.23***	.13***	.26***	.19***

[a] To establish cut-scores for problematic inattention/non-compliance, we began with cut-scores established in the existing literature for the post hoc indices (e.g., Huang et al., 2012; Johnson, 2005), then cross-validated these suggested cut-scores using power analyses in Study 3, choosing cut-scores that optimized power. For 4 of these indices (multivariate distances, long string index, psychometric synonyms, and even–odd consistency) the power analyses suggested cut-scores that converged with the previous literature. On the remaining indices of inattention and non-compliance, we set the cut-scores to reflect inattention/non-compliance rates of ⩾50%. Thus, we used cut-scores of significant multivariate distances ($\chi^2(54)$'s > 91.87, $p < .001$), 8 or greater for the long-string index, <−.03 for psychometric synonyms, <−.65 for psychometric antonyms, <.30 for the even–odd consistency, ⩾3 mistakes on directed questions, ⩽60 s watching video, ⩾10 mistakes on pronoun task, ⩽14.5 min spent on survey, and an average score of 4 or higher on the measures of self-reported responding styles (indicating an average response of "about half the time" or greater).

† $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 3
Ability of the ARS-18, ARS-33, and PAI scales to identify computer-generated random data in Studies 1, 2, and 3.

Validity scale	Hit rate for correctly identifying random data			Proportion of actual respondents identified as inattentive		
	ARS-33	ARS-18	PAI	ARS-33	ARS-18	PAI
Study 1						
Inconsistency subscale	.96	.86		.04	.04	
Infrequency subscale	.99	.91		.05	.05	
Combined	1.00	.99		.08	.08	
Study 2						
Inconsistency subscale	.95	.84	.66	.03	.04	.04
Infrequency subscale	1.00	.95	.89	.05	.05	.05
Combined	1.00	.98	.97	.07	.08	.09
Study 3						
Inconsistency subscale	.97	.86		.04	.03	
Infrequency subscale	.99	.91		.05	.03	
Combined	1.00	.99		.07	.06	

Note: Random data (representing 50% of the resulting sample) was generated for each sample using a random number generation algorithm designed to produce random numbers with uniform distributions.

excessive levels of inattention, respectively. The DQS identified 4.9% of participants as excessively inattentive. These proportions fall well within the estimates of inattentive responding suggested by the self-report and behavioral markers, are comparable to the rates of random inattention estimated by Meade and Craig (2012), and are similar to the proportions obtained in Studies 2 and 3 (Table 3). As seen in Table 3, when the two subscales for each version of the ARS were used jointly (identifying a row of data as problematic if it was flagged by either subscale), the ARS-33 and ARS-18 identified 100% and 99% of randomly generated data, respectively. Computer generated random data automatically provides answers to all questions. As a result, it would have provided answers to all of the DQS items (including the 5 items asking respondents to not provide answers), thereby inflating DQS scores in all random rows of data. Thus, computer-generated random data was not used to evaluate the DQS as it would have served as a problematic method of approximating inattention for that scale.

2.6.4.3. Links to other indices of inattention and non-compliance. As shown in Table 4, respondents identified as having problematic levels of inattention by the ARS-33: (1) made an average of more than 2 additional mistakes on the 7 questions directing specific answers, (2) had stretches of identical answers an average of 5–6 items longer, (3) made an average of 7–8 more mistakes on the pronoun task, (4) skipped an average of 36 s more of the 120 s video, (5) spent 6–7 fewer minutes completing the survey, and (6) exhibited significant mean differences on statistical indices of inattention in comparison to attentive respondents, all yielding generally large effect sizes (Cohen's *d*s ranging from .47 to 2.94). The ARS inattentive respondents also demonstrated significant differences on self-reported levels of careless, patterned, and rushed responding, and reported skipping instructions more frequently than other participants, yielding robust effect sizes (Cohen's *d*s ranging from .56 to 1.71).

The subscales of the ARS-33 and ARS-18 as well as the 7-item DQS revealed similar differences for participants they identified as excessively inattentive to those obtained with the combined ARS-33 or ARS-18 scales, yielding a large set of significant effects (Table 4). More specifically, the 11-item infrequency scale of the ARS-33, the 6-item infrequency scale of the ARS-18, and the 7-item DQS yielded comparable effect sizes to those obtained with the combined ARS-33 or ARS-18 scales on all of the indices of inattention examined (with the exception of the multivariate outlier index). This would suggest that it might be possible to identify excessive inattention with the addition of as few as 6 or 7 items to a survey. In contrast, the 22-item ARS-33 inconsistency subscale and the 12-item ARS-18 inconsistency subscale showed the strongest differences for the participants they identified as excessively inattentive on the multivariate outlier index (yielded fewer significant effects with generally smaller effect sizes on the remaining indices of inattention). As the multivariate outlier index essentially identifies extremely unlikely inconsistencies within self-report scales, these results help to contrast the type of inattentive responding captured by the inconsistency subscales from that identified by the infrequency and directed question scales.

Taken as a set, these results showed that individuals identified as inattentive by the ARS and DQS scales performed poorly across a wide range of indices assessing effort and care, suggesting strong levels of criterion validity for the scale. These results further suggested that inattentive respondents could potentially be expected to: (1) fail to read and respond carefully to item content, (2) select the same answer for a block of items regardless of item content, (3) put notably lower amounts of effort into even short and straightforward priming tasks, (4) skip significant portions of videos used as experimental manipulations, and (5) rush through online surveys. Thus, although these participants voluntarily chose to participate in a survey, the roughly 8% identified by the ARS and DQS scales as excessively inattentive seem to have been relatively disengaged in the process of completing the survey, putting forth very little effort on any of the tasks presented to them.

2.6.5. Latent profile analyses of inattention indices

We conducted a latent profile analysis to see if we could replicate the three latent classes of inattention obtained by Meade and Craig (2012) using the ARS-33 infrequency and inconsistency subscales, the DQS and the 6 indicators used in their analyses. As seen in the top half of Table 5, the 3-class model (supported over 1 or 2-class models by Bayesian Information Criterion) conceptually replicated the findings of Meade and Craig (2012), revealing: (1) a large attentive class (Class 1; 95.1%), (2) a generally inattentive class (Class 2; 4.0%) marked by elevations on a range of inattention indices including ARS and DQS scores, high multivariate distances, low even–odd consistency and low correlations between psychometric synonym pairs, and (3) a smaller class of patterned responders (Class 3; 0.9%) made up of participants who tended to select the same answer repeatedly for an entire block of questions (also showing elevations on the ARS infrequency subscale). This pattern of results also emphasizes one difference between the infrequency and inconsistency subscales of the ARS—the ARS infrequency subscale (like the long string index) is likely to pick up on patterned responding whereas the ARS inconsistency subscale would not. For instance, an individual choosing the same response for every

Table 4
Criterion validity of inattention scales: effect sizes between attentive and inattentive respondents on behavioral and self-report measures of inattentiveness.

	Measures of inattention						Non-compliance with study tasks			Self-reported inattention			
	Mistakes on directed Qs	Multivariate distances	Long string index	Psychometric synonyms	Psychometric antonyms	Even–odd consistency	Mistakes on pronouns	Seconds of video skipped	Time spent on survey	Inattentive	Patterned	Rushed	Skipping instructions
Study 1													
ARS-33 combined													
M for attentive *P*'s	.14	52.74	4.06	.76	.44	.78	1.65	13.51	29.43	1.87	1.30	2.19	2.56
M for inattentive *P*'s	2.43	62.30	9.85	.43	.11	.50	9.23	49.71	22.72	2.94	2.58	2.86	3.32
Cohen's *d*	2.94**	.47*	1.49**	1.02**	.77**	.90**	1.86**	1.12**	.53**	1.33**	1.71**	.66**	.56**
Cohen's *d* effect sizes for additional indices													
ARS-33: 11-item infreq.	4.83**	.42	1.65**	1.26**	1.05**	1.09**	2.48**	1.66**	.59*	1.66**	2.22**	.97**	.82**
ARS-33: 22-item incon.	1.32**	.84*	1.04*	.87*	.63*	.75*	1.08**	.54†	.63**	.98*	.93**	.34	.40
ARS-18 combined	2.36**	.48*	.98**	.84**	.52*	.72**	1.60**	1.12**	.40*	.80**	1.33**	.50**	.46**
ARS-18: 6-item infreq.	4.52**	.49	1.39*	1.14**	.76*	.94**	2.33**	1.43**	.58*	1.44**	2.02**	.89**	.84**
ARS-18: 12-item incon.	.58	.88*	.28	.40	.34	.58*	.79*	.62†	.32	.00	.31	.11	.07
DQS: 7 Directed Qs	–	.65	1.64*	1.04**	1.06**	.78**	2.74**	1.54**	.42	1.75**	2.48**	1.11**	.92**
Study 3													
ARS-33 (7.4% identified as inattentive)													
M for attentive *P*'s	.18	53.15	3.81	.74	.32	.76	1.73	29.67	28.78	1.91	1.32	2.26	2.66
M for inattentive *P*'s	1.79	65.89	8.82	.52	.12	.52	6.56	63.79	25.63	2.67	2.40	2.62	3.20
Cohen's *d*	2.14**	.62**	1.55**	.75**	.38*	.67**	1.20**	.76**	.24	.96**	1.64**	.37*	.42**
Cohen's *d* effect sizes for additional indices													
ARS-33: 11-item infreq.	3.19**	.29	2.13**	1.04**	.67**	.91**	1.68**	.56*	.47*	1.37**	2.34**	.66**	.69**
ARS-33: 22-item incon.	1.56**	1.07**	.76†	.57*	.15	.52†	.57†	.41	−.08	.65*	1.22**	.26	.28
ARS-18 combined	2.17**	.74**	1.58**	.94**	.48**	.70**	.89**	.85**	.40*	.89**	1.57**	.30	.40*
ARS-18: 6-item infreq.	4.14**	.96*	2.15*	1.27**	.82**	1.08**	1.79**	1.08**	.72**	1.66**	3.17**	.78*	.69*
ARS-18: 12-item incon.	1.10*	.75*	.73	.71*	.32	.43	.22	.25	.12	.38	.83†	.11	.28
DQS: 7 Directed Qs	–	.35	2.73**	1.32**	.81**	1.18**	2.54**	1.15**	1.01**	2.14**	4.09**	1.30**	1.24**
IMC	.62**	.01	.42**	.18*	.14†	.37**	.46**	.72**	.04	.64**	.59**	.36**	.60**

Note: All analyses were tested with *t*-tests assuming unequal variances.
† $p < .10$.
* $p < .05$.
** $p < .01$.

Table 5
Latent profile analysis results across Studies 1, 2 and 3.

Variable	Study 1			Study 2			Study 3		
	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned
Class size	641 (95.1%)	27 (4.0%)	6 (0.9%)	279 (94.3%)	15 (5.1%)	2 (0.7%)	740 (97.1%)	22 (2.2%)	5 (0.7%)
Means									
ARS-33 incon.	5.10	**<u>8.99</u>**	4.00	4.88	**<u>9.74</u>**	0.00	5.06	**<u>9.45</u>**	4.80
ARS-33 infreq.	2.61	**<u>17.88</u>**	**<u>16.83</u>**	2.97	**<u>16.56</u>**	**<u>23.00</u>**	2.64	**<u>19.55</u>**	**<u>17.80</u>**
PAI incon.	–	–	–	4.11	4.77	4.00	–	–	–
PAI infreq.	–	–	–	3.89	11.61	**<u>16.00</u>**	–	–	–
IMC	–	–	–	–	–	–	0.25	**<u>0.71</u>**	**<u>0.80</u>**
Long string	4.13	5.19	**<u>42.00</u>**	4.15	6.18	**<u>38.00</u>**	3.89	6.74	**<u>38.60</u>**
DQS directed Qs	0.12	**<u>4.44</u>**	3.33	–	–	–	0.18	**<u>4.29</u>**	**<u>4.20</u>**
MV distances	52.96	**<u>74.79</u>**	16.83	53.36	57.05	48.19	53.75	**<u>71.92</u>**	36.64
Time spent	29.24	21.14	27.67	–	–	–	28.86	18.43	**<u>15.60</u>**
Even–odd	.77	**<u>.45</u>**	.87	.75	**<u>.07</u>**	**<u>.00</u>**	.75	**<u>.17</u>**	.79
PM synonyms	.75	**<u>.29</u>**	.93	.74	**<u>.35</u>**	1.00	.73	**<u>.19</u>**	.90
PM antonyms	.44	.03	**<u>−.52</u>**	.46	.24	**<u>−.50</u>**	.32	.03	**<u>−.78</u>**
Ability of specific indices to identify latent classes									
Inattention index	Proportion of latent class identified as inattentive								
	Study 1			Study 2			Study 3		
	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned	Class 1 Attentive	Class 2 Inattentive	Class 3 Patterned
ARS-33 joint	**.04**	**.82**	**1.00**	**.02**	**.87**	**1.00**	**.05**	**.88**	**1.00**
ARS-33 incon.	.03	.26	.33	.01	.40	.00	.03	.41	.20
ARS-33 infreq.	.01	.78	.83	.004	.80	1.00	.03	.88	1.00
ARS-18 joint	**.04**	**.78**	**.67**	**.03**	**.80**	**1.00**	**.04**	**.77**	**.80**
ARS-18 incon.	.03	.15	.00	.02	.33	.00	.03	.29	.20
ARS-18 infreq.	.01	.78	.67	.01	.60	1.00	.01	.77	.60
PAI joint	–	–	–	**.05**	**.20**	**1.00**	–	–	–
PAI incon.	–	–	–	.04	.07	.00	–	–	–
PAI infreq.	–	–	–	.004	.20	1.00	–	–	–
IMC	–	–	–	–	–	–	.25	.71	.80
Long string	.05	.19	1.00	.07	.20	1.00	.04	.18	1.00
DQS directed Qs	.002	1.00	.83	–	–	–	.01	.94	1.00
MV distances	.03	.15	.00	.04	.00	.00	.05	.06	.00
Time spent	.06	.52	.33	–	–	–	.06	.65	.40
Even–odd	.06	.33	.00	.09	.47	.50	.07	.47	.00
PM synonyms	.03	.26	.00	.03	.20	.00	.03	.24	.00
PM antonyms	.02	.11	.50	.03	.07	.50	.06	.12	.60

Note: In the top half of the table, bolded and underlined means reflect the highest levels of inattention/non-compliance significantly different from the remaining means as assessed by Tukey post hoc tests. The bottom half of the table presents the proportion of each inattentive class identified by each index using the same cut-scores described in Table 1.

question may have an elevated infrequency score (as he or she would provide extremely unlikely responses on half of the infrequency items) even though his or her responses to inconsistency item-pairs would appear highly consistent. Similarly, the DQS demonstrated higher sensitivity for detecting patterned responders than the ARS inconsistency subscale as those responders would be less likely to skip questions or alter their pattern to fill in the requested answers. Thus, both the ARS infrequency subscale and the DQS demonstrated elevations in both latent groups of inattentive respondents whereas the ARS inconsistency subscale showed elevations just for the larger group of inattentive respondents.

As seen in the bottom half of Table 5, we then examined the ability of each index of inattention to accurately identify individuals in the two latent inattentive classes using previously validated cut-scores. When used jointly, the ARS subscales were highly effective at identifying the respondents in those latent groups (identifying 80% and 100% of the respondents in classes 2 and 3 respectively) and the DQS was also quite effective (identifying 100% and 83% respectively). Thus, the ARS and DQS scales would seem to provide researchers practical and easily implemented methods of identifying both of these forms of inattention. The remaining indices identified no more than about half of the latent inattentive respondents across these 2 classes.

2.6.6. Inattentive responding and the quality of self-report data

To evaluate the quality of data among attentive and inattentive participants, we examined the internal consistency of well-validated self-report scales. As seen in Table 6, all of the scales demonstrated high levels of internal consistency in the respondents identified as sufficiently attentive by: the latent profile analysis, the ARS-33, the ARS-18, the 11-item ARS-33 infrequency subscale, the 6-item ARS-18 infrequency subscale, and the 7-item DQS. However, these scales demonstrated significantly lower levels of internal consistency (tested using AlphaTest; Lautenschlager & Meade, 2008) among participants identified as inattentive by any of these indices. This indicated that inattentive respondents (even those identified by as few as 6 items) provided markedly poor quality self-report data, riddled with enough error variance to make the items of even high quality scales appear to be notably less correlated with one another.

2.6.7. Inattentive responding and the quality of correlational analyses

To extend these results further, we examined multivariate regression analyses in both attentive and inattentive participants, thereby examining how inattention might affect the substantive analyses being addressed within a study. The analyses attempted to replicate previous work showing that the personality

Table 6
Internal consistency for participants flagged as attentive and inattentive using the ARS and the PAI.

Measure	Study 1			Study 2		
	Attentive α	Inattentive α	$\chi^2(1)$	Attentive α	Inattentive α	$\chi^2(1)$
Attentive vs. inattentive latent classes						
Extraversion	.87	.39	40.88***	.87	.44	16.05***
Agreeableness	.80	.61	6.10*	.77	.18	13.66***
Conscientiousness	.83	.40	24.58***	.85	.19	26.68***
Neuroticism	.87	.51	27.57***	.88	.00[a]	41.99***
Openness	.80	.53	9.65**	.79	.60	2.98 †
Attentive vs. inattentive on the ARS-33						
Extraversion	.87	.60	33.01***	.86	.78	1.67
Agreeableness	.81	.51	19.65***	.77	.56	3.37 †
Conscientiousness	.83	.64	13.18***	.86	.25	28.15***
Neuroticism	.87	.64	25.55***	.88	.47	19.27***
Openness	.80	.61	9.86**	.80	.72	0.82
Attentive vs. inattentive on the ARS-18						
Extraversion	.87	.61	31.65***	.87	.65	9.05**
Agreeableness	.81	.58	13.95***	.77	.59	3.20 †
Conscientiousness	.83	.72	5.10*	.85	.65	7.19**
Neuroticism	.87	.68	19.23***	.88	.61	13.22***
Openness	.80	.68	4.77*	.79	.78	0.01
Attentive vs. inattentive on the ARS-33: 11-item infrequency subscale						
Extraversion	.87	.30	56.69***	.87	.57	9.06**
Agreeableness	.80	.56	9.78**	.77	.31	8.96**
Conscientiousness	.83	.49	21.35***	.85	.31	18.26***
Neuroticism	.87	.40	43.16***	.88	.00	33.45***
Openness	.80	.48	13.69***	.80	.61	2.51
Attentive vs. inattentive on the ARS-18: 6-item infrequency subscale						
Extraversion	.87	.44	36.92***	.87	.46	12.63***
Agreeableness	.80	.18	35.03***	.77	.50	3.84*
Conscientiousness	.83	.54	14.85***	.85	.61	5.42*
Neuroticism	.87	.65	14.40***	.88	.00	30.73***
Openness	.80	.58	7.24**	.80	.61	2.45
Attentive vs. inattentive on the 7 directed questions of the DQS						
Extraversion	.87	.37	42.68***	–	–	–
Agreeableness	.80	.60	6.88**	–	–	–
Conscientiousness	.83	.42	23.39***	–	–	–
Neuroticism	.87	.58	21.03***	–	–	–
Openness	.80	.53	9.39**	–	–	–
Attentive vs. inattentive on the PAI						
Extraversion	–	–	–	.86	.80	0.79
Agreeableness	–	–	–	.76	.82	0.47
Conscientiousness	–	–	–	.85	.70	3.60†
Neuroticism	–	–	–	.88	.78	2.33
Openness	–	–	–	.80	.73	0.57

Note: Internal consistency of the scales was assessed with unstandardized Chronbach's alpha coefficients.
[a] Responses were extremely heterogeneous in this group, resulting in a negative alpha. This negative value was truncated to 0.
† $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

dimensions of neuroticism, extraversion, and conscientiousness are fairly robust predictors of self-esteem (Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001). In the participants classified as attentive by either latent profile analyses, the ARS-33, the ARS-18, or the DQS, the adjusted R^2 statistics suggested that the personality variables collectively accounted for 44–45% of the variance in self-esteem (all corresponding *F*s significant at $p < .001$) and the regression coefficients replicated previous work showing significant associations between each of the three personality traits and self-esteem (all significant at $p < .001$). However, the personality variables failed to predict significant variance in self-esteem among participants included in the two inattentive latent classes ($F(3,28) = 0.90$, $p = .453$) or marked as inattentive using the DQS ($F(3,28) = 1.43$, $p = .26$). Similarly, among participants marked as inattentive by either the ARS-33 or the ARS-18, personality accounted for only 29% and 20% of the variance in self-esteem (respectively), and only neuroticism remained a significant predictor ($p < .05$). These results continued to suggest that the inattentive respondents provided data of markedly poor quality, so riddled with error variance that meaningful regression results were obscured in those subjects.

3. Study 2

Taken as a set, the results of Study 1 suggested that the inattentive respondents (identified by the multivariate latent profile analyses, the ARS, or the DQS) demonstrated low levels of effort and compliance with study tasks and seemed to provide poor quality self-report data, even to the point of obscuring regression results found in the attentive respondents. Those results not only replicated the multivariate work of Meade and Craig (2012), but also helped to quantify the impact of inattention on analyses making use of self-report data. In fact, those results suggest that the inattention being identified represents an extreme degree of non-compliance. This form of inattention does not reflect participants providing an accurate answer quickly with little cognitive elaboration; rather, such responding seems to contribute primarily error variance to analyses. The results further suggested that indicators like the ARS might offer a method for quickly identifying the latent inattentive classes uncovered by Meade and Craig, offering a simple method of eliminating both of these sources of error variance. To link this work with the literature examining validity in large-scale clinical assessment batteries (e.g., the PAI; Morey, 1991), Study 2 examined a broad set of indicators of inattention in direct comparison to the infrequency and inconsistency subscales of the PAI.

3.1. Study 2 method

3.1.1. Participants and procedure

A total of 296 undergraduate students participated in Study 2 in exchange for course credit in psychology classes. Participants were predominantly female (71%) and Caucasian (72%), 17% Asian, and 4% African American. The mean age was 19.9 years (*SD* = 1.4). Participants were recruited from an undergraduate psychology participant pool. The study was presented online and took roughly 15 min to complete. Extra credit for psychology courses served as the primary recruitment incentive.

3.1.2. Measures

3.1.2.1. Inattentive responding items. Participants completed the 33-item Attentive Responding Scale (ARS) as developed in Study 1. As in Study 1, the ARS items were split evenly between both halves of the survey.

3.1.2.2. Personality Assessment Inventory validity scales. The study included the 8-item infrequency and the 12-item inconsistency subscales of the PAI (Morey, 1991). As noted previously (in Section 1.3.1), the PAI is a large clinical inventory that includes measures of inattention that employ a strategy similar to that of ARS. The 20-item PAI is shorter than the 33-item ARS, and therefore might potentially offer lower precision in assessing extreme inattention. In addition, whereas the ARS was developed from a large item pool tested across multiple large-scale samples (see Sections 2.2 and 2.6.4.1) as a primary area of focus, the PAI validity scales were developed as secondary components of a large clinical assessment battery. Despite these differences, the PAI infrequency and inconsistency subscales represent validated measures of inattention from the existing literature that can provide benchmarks for the performance of the ARS subscales. To ensure that the PAI items obtained similar levels of variance to that of the ARS, the PAI items were rated on the same 5-point scale (1 = *not at all true* to 5 = *very true*) as the ARS items and were scored in an identical manner to the ARS subscales.

3.1.2.3. Individual difference measures. The same scales from Study 1 were again used to assess individual differences in personality and self-esteem. The scales continued to demonstrate adequate levels of internal consistency in this sample (see Table 6).

3.1.2.4. Other indicators of inattention and non-compliance. Participants completed many of the same indicators of inattention and non-compliance used in Study 1, including: multivariate distances based on responses to the 54 items of the BFI and RSES, psychometric synonyms and antonyms based on the 5 most positively and 5 most negatively correlated item pairs in the BFI and RSES, even–odd consistency based on split-half correlations across the BFI and RSES subscales, and a long string index based on responses to the 44 items of the BFI.

3.2. Study 2 results and discussion

3.2.1. Scale level analyses

The ROC analyses in this sample continued to support ARS cut-scores identical to those established and validated in Study 1. For the PAI, ROCs suggested cut-scores of 14.5 for the infrequency scale (3.0 *SD*s above the mean) and 8.5 for the inconsistency scale (1.9 *SD*s above the mean). Using these cut-scores, the PAI scales identified approximately the same proportion of inattentive respondents as did the ARS (see Table 3). When these cut-scores were used to discriminate between actual and random data, the ARS-18, ARS-33, and PAI scales achieved very high accuracy (97–100%) at identifying randomly generated data, yielding the expected proportions of inattentive participants (7–9%).

3.2.2. Latent profile analyses of inattention indices

As shown in Table 5, using a slightly different set of inattention indices, we once again replicated the two latent inattentive classes of Meade and Craig (2012), revealing a generally inattentive class (5.1%) and a second, smaller class characterized by patterned responding (0.7%). As shown in the bottom half of Table 5, the ARS continued to effectively identify these latent classes (identifying 88% of these latent inattentive respondents) whereas the PAI and the remaining indices of inattention were less effective (identifying no more than 47% of those individuals). As the PAI infrequency and inconsistency subscales represent validated inattention indices from the existing literature, these results indicate that the ARS and its subscales exceed the quality of screening afforded by these benchmarks (likely due to the extra validation efforts invested into the development of the ARS).

3.2.3. Inattentive responding and the quality of self-report data

To evaluate the quality of data provided by inattentive participants, we once again examined the internal consistency of well-validated self-report scales. As seen in Table 6, all of the scales demonstrated high levels of internal consistency in the respondents identified as sufficiently attentive by the latent profile analysis, the ARS-33 or ARS-18 total scales, the ARS-33 or ARS-18 infrequency subscales, or the PAI. However, the self-report scales demonstrated significantly or marginally lower levels of internal consistency in participants identified as excessively inattentive by the latent class analyses or by the ARS scales/subscales. In contrast, only 1 of the 6 scales was (marginally) less internally consistent among the PAI inattentive respondents. Thus, in comparison to that benchmark, the latent profiles and the ARS seem to have been particularly effective at identifying individuals providing lower quality data.

3.2.4. Inattentive responding and the quality of correlational analyses

Extending these results, we examined how inattentive respondents might affect the substantive analyses of a study by running regression analyses replicating links between neuroticism, extraversion, conscientiousness and self-esteem (Robins et al., 2001). In the participants classified as attentive either by latent profile analyses, the ARS-33, the ARS-18, or the PAI, the adjusted R^2 statistics demonstrated that the personality variables collectively accounted for 53–54% of the variance in self-esteem (all *F*s significant at $p < .001$). In contrast, the amount of variance accounted for by the predictors was lower among inattentive respondents identified by latent profile analysis (42%), the ARS-33 (39%), the ARS-18 (7%), or the PAI (47%). These results again suggest that inattentive responding was associated with markedly higher levels of error variance, obscuring meaningful results. The results also continue to suggest that the ARS performs at least as well as the PAI in identifying problematic respondents.

4. Study 3

4.1. Effects of inattention on statistical power

The results of Study 2 replicated the findings of Study 1, underscoring the potential effects of inattention on correlational analyses. Study 3 sought to extend this work by examining the effects of inattention on experimental manipulations and the associated power to detect significant effects for those manipulations. Experiments often include relatively subtle manipulations involving alternate forms of instructions, primes, pictures or other stimuli. Inattentive participants may not be influenced by such manipulations either because they do not attend to what is being manipulated, or because their responses are inaccurate. Thus, removing them from a dataset could increase statistical power (e.g., Oppenheimer et al., 2009). Toward this end, Study 3 adopted the criteria used by Oppenheimer et al. (2009) by including two experimental manipulations adapted from Thaler (1985) involving subtle differences in wording embedded within instructional paragraphs. Although both of these tasks involved text-based manipulations, they differ in that one task is fairly short and uses a Likert response-scale as the outcome measure (requiring minimal effort for participants), whereas the second task involves the manipulation of a single phrase within a longer paragraph and requires participants to type in an open-ended response (requiring notably more effort on the part of participants). We hypothesized that highly inattentive participants would not be influenced by these manipulations. We also examined gender differences in the personality trait of openness to experience, hypothesizing that such gender differences would be obscured among inattentive respondents. We further hypothesized that excluding inattentive respondents would yield measurable increases in the power to detect these effects in a series of resampling analyses. By examining multiple effects, Study 3 sought to determine how well the potential power boosts afforded by screening for inattention might generalize across various types of analyses.

4.2. Instructional manipulation check (IMC)

Study 3 also added the IMC to the battery of inattention indices being examined, a tool developed by Oppenheimer et al. (2009) to identify participants who do not sufficiently attend to instructions. As described earlier, the IMC identifies participants inattentive to text-based manipulations, potentially yielding clearer effects of text-based manipulations when used to screen for inattention. However, the IMC identifies a remarkably high proportion of respondents as inattentive, potentially requiring experimenters to exclude as many as 35–48% of their original samples (e.g., Oppenheimer et al., 2009; Simmons & Nelson, 2006), limiting its practical utility as a measure of inattention. Given the much smaller estimates of inattentive responding associated with the latent profile analyses and the ARS (5–9%), we hypothesized that other indices like the ARS might offer a more efficient method of screening for inattention when compared to the IMC as a benchmark, yielding comparable improvements to those obtained with the IMC through the exclusion of a much smaller proportion of participants.

4.3. Study 3 method

4.3.1. Participants and procedure

A total of 767 individuals completed the online survey. Participants were predominantly female (70%) and Caucasian (74%), 15% Asian, and 6% African American. Participants reported an average of 13.9 years of education (SD = 2.1) and the median household income range was \$40,000–49,999. The mean age was 28.16 years (SD = 11.51). Participants were recruited from Mechanical Turk (60%) and from a psychology participant pool (40%) with either small monetary incentives (25 cents) or extra credit (respectively). The survey took roughly 20 min to complete. All participants were also entered into a lottery for \$50.

4.3.2. Measures

4.3.2.1. Indicators of inattention and non-compliance. Participants completed the same indicators of inattention and non-compliance used in Study 1, including: the ARS-33, the 7 directed questions of the DQS, multivariate distances, psychometric synonyms and antonyms, even–odd consistency, a long string index, the pronoun identification task, time spent watching a 2.5 min long video clip (assigned to a random half of the sample), and time spent on the survey.

4.3.2.2. Self-report measures. Participants completed the same items assessing self-reported responding styles used in Study 1, asking about their careless responding, patterned responding, rushed responding, and skipping of instructions. Participants also completed the BFI.

4.3.2.3. Instructional manipulation check. Participants completed the instructional manipulation check (IMC; Oppenheimer et al., 2009) to assess inattention in reading instructions. The IMC consists of what appears at face value to be a simple question about engagement in various physical activities (skiing, soccer, etc.) which is preceded by a lengthy (7-sentence) paragraph of instructions. Embedded in this paragraph is text instructing participants to ignore the question and its corresponding response choices, ignore the continue button at the bottom of the page and instead click on the title at the top of the page to advance to the next page of the survey. Only participants reading the entire instructional paragraph would notice these instructions. The survey recorded whether or not participants successfully followed the instructions embedded at the end of the lengthy paragraph.

4.3.2.4. Experimental manipulations. Participants completed the same two experimental manipulation tasks used by Oppenheimer et al. (2009) to validate the usefulness of the IMC, both adapted from classic studies (Thaler, 1985). The first task examined a sunk cost effect by asking participants how likely they would be to attend a football game during a freezing cold day (Oppenheimer et al., 2009):

> Imagine that your favorite football team is playing an important game. You have a ticket to the game that you [have paid handsomely for] [have received for free from a friend]. However, on the day of the game, it happens to be freezing cold. What do you do?

Participants were randomly assigned to one of the two versions of this paragraph (stating that they had either received the ticket for free or paid handsomely for it), and were then asked how likely they would be to attend the game on a 9-point scale ranging from *1 (definitely stay at home)* to *9 (definitely go to the game)*. The second task asked participants how much they would be willing to pay to purchase a can of soda:

> You are on the beach on a hot day. For the last hour you have been thinking about how much you would enjoy an ice cold can of soda. Your companion needs to make a phone call and offers to bring back a soda from the only nearby place where drinks are sold, which happens to be a [run-down grocery store] [fancy resort]. Your companion asks how much you are willing to pay for the soda and will only buy it if it is below the price you state. How much are you willing to pay?

Participants were randomly assigned to complete one of the two versions of the task (where the purchase is from either a grocery store or a fancy resort) and were then asked to provide an exact amount for how much they would be willing to pay for the soda (in US dollars and cents). We truncated these values to a maximum of \$10 (the 97.5th percentile) prior to running analyses.[5]

4.4. Study 3 results and discussion

4.4.1. Contrasting the ARS with the IMC

4.4.1.1. Scale level analyses. As seen in Table 3, the ARS-33 and ARS-18 subscales demonstrated high levels of accuracy (99–100%) at identifying randomly generated data, while identifying 7.4% and 5.9% of participants as inattentive, respectively. Consistent with previous work (e.g., Oppenheimer et al., 2009), the IMC identified a much larger number of participants (26%) as problematic, highlighting a key difference between the IMC and ARS. As suggested by the results of Study 1, skipping instruction sets is a fairly common phenomenon as 84% of respondents reported doing it at least some of the time and 19% of respondents reported doing it at least half of the time. Thus, the IMC's exclusive focus on skipping instructions identifies a large portion of participants, potentially eliminating respondents who skipped instructions but might still have attended carefully to the individual items of the survey.

4.4.1.2. Links with other indices of inattention. As shown in Table 4 and Fig. 1, contrasts between attentive and inattentive respondents as assessed by the ARS-33, the ARS-18, the ARS-18 or ARS-33 subscales, and the 7-item DQS yielded notably larger effect sizes on the remaining indices of compliance and attention than similar contrasts using the IMC, exceeding the performance of that benchmark. Analyses directly comparing the magnitude of these effects across the ARS-33, ARS-18, and IMC (see Rosenthal & Rubin, 1982; Fig. 1) suggested that the comparison effects were significantly larger for analyses using the ARS-33 or ARS-18 than for analyses using the IMC for 7 of the 13 indices. Thus, although the IMC exhibited statistically significant effects on 11 of the 13 inattention/non-compliance indices, those contrast effects were generally moderate in magnitude and smaller than those obtained with the ARS-33 and ARS-18. The ARS-33 and ARS-18 yielded comparable effect sizes for all analyses, with no significant differences between them. Taken as a set, these results continued to support the utility of the ARS and the DQS in assessing a broad range of inattention with relatively low associated costs (excluding 7% of a sample) whereas the IMC might be more specific in the form of inattention being identified with correspondingly higher associated costs (excluding ⩾27% of a sample).

4.4.2. Latent profile analyses of inattention indices

As shown in Table 5, even after adding the IMC to the set of inattention indices being examined, we once again replicated the two latent inattentive classes of Meade and Craig (2012), revealing a generally inattentive class and a second, smaller class characterized by patterned responding. The ARS subscales and the DQS continued to effectively identify the people in these latent classes whereas the remaining indices of inattention were less effective at identifying the two types of inattentive responding. Although the IMC was nearly as effective as the ARS-18 at identifying people in the inattentive latent classes, it came with the associated cost of excluding 25% of the attentive class, highlighting a central drawback to using the IMC to screen data (as opposed to an intervention as suggested by its developers).

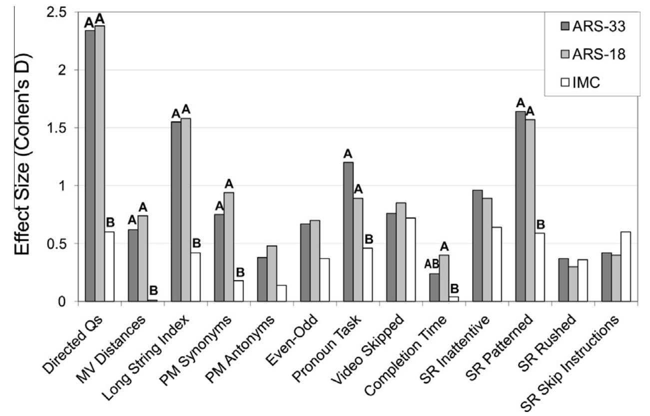

Fig. 1. Effect sizes representing mean differences between attentive and inattentive respondents as identified by each indicator. Bars with different letters differ significantly in the magnitude of the effect size for that dependent variable.

4.4.3. Inattentive responding and text-based manipulations

To evaluate the impact of inattention on text-based experimental manipulations, we ran separate analyses with attentive and inattentive respondents attempting to replicate the results of the sports ticket manipulation and the soda-purchase manipulation from Thaler (1985).

4.4.3.1. Sports ticket manipulation. The sports ticket manipulation yielded significant effects in the ARS-33 and ARS-18 attentive respondents (ARS-33: $M_{paid} = 7.00$, $M_{free} = 6.19$, $t(686) = 5.89$, $p < .001$, Cohen's $d = .45$; ARS-18: $M_{paid} = 6.98$, $M_{free} = 6.19$, $t(699) = 5.79$, $p < .001$, Cohen's $d = .44$), whereas the manipulation failed to yield significant effects in the inattentive respondents (ARS-33: $M_{paid} = 6.19$, $M_{free} = 6.42$, $t(55) = -0.44$, $p = .66$, Cohen's $d = -.12$; ARS-18: $M_{paid} = 6.25$, $M_{free} = 6.46$, $t(42) = -.36$, $p = .72$, Cohen's $d = -.11$). Similarly, while the manipulation yielded significant effects in the DQS attentive respondents ($M_{paid} = 6.95$, $M_{free} = 6.21$, $t(720) = 5.40$, $p < .001$, Cohen's $d = .40$), the manipulation failed to yield significant effects in the DQS inattentive respondents ($M_{paid} = 6.80$, $M_{free} = 6.08$, $t(21) = 1.13$, $p = .27$, Cohen's $d = .50$). These results were consistent with our hypotheses, suggesting that the respondents identified as inattentive by the ARS or DQS were generally not influenced by this text manipulation. In contrast to the findings with the ARS and DQS, this manipulation yielded significant effects both in respondents passing the IMC ($M_{paid} = 6.97$, $M_{free} = 6.20$, $t(540) = 4.85$, $p < .001$, Cohen's $d = .42$) and in respondents failing the IMC ($M_{paid} = 6.85$, $M_{free} = 6.20$, $t(196) = 2.64$, $p = .009$, Cohen's $d = .38$). Thus, the ARS and DQS seemed to outperform this benchmark on this manipulation, more effectively and efficiently identifying individuals who might not have fully attended to the instructions.

4.4.3.2. Soda manipulation. After truncating extreme scores as previously described, the soda manipulation yielded significant effects in the ARS-33 and ARS-18 attentive respondents (ARS-33:

[5] Given the open-ended format of the response in the soda manipulation, 7 respondents provided such extreme (and nonsensical) answers that they exceeded the 97.5th percentile of responses (\$10). To ensure that the benefits gained by screening out inattentive respondents were not simply due to eliminating these outliers (a common data cleaning step), we truncated values to within the top 97.5% of the distribution. In fact, the most extreme univariate outlier on this measure (who offered to pay \$1,000 for a soda purchased from a run-down grocery store) was identified as excessively inattentive based on nearly every indicator of inattention that we examined (both ARS subscales, multivariate distances, directed questions, time spent completing the survey, and the IMC), with the exception of the long string index. Without truncating these scores, our results are much stronger, such that using measures of inattention to clean the data had a more robust and impressive effect on results. However, we chose to present this more conservative analysis to see if screening out highly inattentive respondents would affect results and increase statistical power even after screening for obviously problematic responses.

M_{resort} = \$3.28, $M_{grocery}$ = \$2.36, $t(707) = 6.96$, $p < .001$, Cohen's $d = .52$; ARS-18: M_{resort} = \$3.26, $M_{grocery}$ = \$2.36, $t(719) = 6.91$, $p < .001$, Cohen's $d = .52$), whereas the manipulation failed to yield significant effects in the inattentive respondents (ARS-33: M_{resort} = \$3.32, $M_{grocery}$ = \$3.05, $t(54) = 0.37$, $p = .72$, Cohen's $d = .10$; ARS-18: M_{resort} = \$3.69, $M_{grocery}$ = \$3.17, $t(42) = .55$, $p = .58$, Cohen's $d = .18$), once again suggesting that the respondents identified as inattentive by the ARS were generally not influenced by this text manipulation. Similarly, while the manipulation yielded significant effects in the DQS attentive respondents ($M_{resort} = 3.27$, $M_{grocery} = 2.40$, $t(739) = 6.76$, $p < .001$, Cohen's $d = .50$), the manipulation failed to yield significant effects in the DQS inattentive respondents ($M_{resort} = 3.00$, $M_{grocery} = 2.92$, $t(22) = 0.07$, $p = .94$, Cohen's $d = .03$). Consistent with these results, the soda manipulation yielded significant effects in the respondents passing the IMC (M_{resort} = \$3.29, $M_{grocery}$ = \$2.18, $t(540) = 7.49$, $p < .001$, Cohen's $d = .65$), but failed to yield significant effects in the respondents failing the IMC (M_{resort} = \$3.31, $M_{grocery}$ = \$2.97, $t(196) = 1.16$, $p = .25$, Cohen's $d = .17$). Thus, on the text manipulation making use of a longer paragraph (and therefore more directly comparable to the IMC itself), the IMC demonstrated particularly strong efficacy. Taken as a set, these results supported our hypothesis that inattentive participants might not be affected by text-based manipulations, potentially adding error variance to the effects and thereby obscuring meaningful results.

4.4.4. Inattentive responding and statistical power

4.4.4.1. Resampling analysis strategy. To build on these manipulation results, we used a series of resampling analyses to quantify the impact of removing inattentive respondents on statistical power across a range of sample sizes ($N = 60$ to $N = 200$) with moderate levels of power for detecting modest effects (for the two experimental manipulations and gender differences on the trait of openness), making them more sensitive to the presence of error variance due to inattentive responding. For each of these sample sizes, we created a total of 2000 resamples (subsets of participants selected randomly from the main sample with replacement). Within each resample, we then tested the effects: (1) in all of the participants in that resample, (2) after removing ARS-33 inattentive participants, (3) after removing ARS-18 inattentive respondents, (4) after removing the DQS inattentive participants, and (5) after removing respondents failing the IMC. Building on the work of Meade and Craig (2012), we also tested effects after removing respondents deemed inattentive based on: the latent inattentive classes, time spent on the survey, each of the 5 statistical indices of inattention, and each of the ARS subscales. We recorded the number of times across the 2000 resamples that the manipulation yielded a statistically significant effect in the predicted direction (at $p < .05$) under each of those conditions after removing participants exceeding an optimized cut-score[6] on each indicator. These resamples therefore provided estimates of the power that could be gained (or lost) in detecting the manipulation effects when using various indicators to screen out inattentive participants across a range of sample sizes. In effect, these analyses allowed us to directly examine if the benefits of screening with each indicator (i.e., the reduced error variance) indeed outweighed the costs (i.e., the decreased power due to smaller sample size) associated with such a strategy by evaluating the net benefit in terms of power.

4.4.4.2. Power gains on the sports ticket manipulation using the ARS, DQS and IMC. As seen in Table 7, which presents average power gains aggregated across the various sample sizes, screening out inattentive respondents with the ARS-33 and ARS-18 increased statistical power by approximately 5.0% and 3.4% (respectively) for the sports ticket manipulation. As seen in Fig. 2, these estimated improvements in power (aggregated across all three effects examined) were consistent across the range of sample sizes tested, suggesting that use of the ARS could provide additional power in both smaller and larger samples. Examining the ARS infrequency and inconsistency subscales separately, the resampling results suggested that use of the inconsistency scales to screen for inattention might have been more effective at boosting power on the sports ticket manipulation than use of the infrequency scales. In contrast to the findings with the ARS, the resampling results suggested that removing participants identified as excessively inattentive on the DQS resulted in a very slight decrease in statistical power (by approximately 1.6%). Similarly, removing participants failing the IMC decreased statistical power on this task (by approximately 10.5%). Thus, on the sports ticket manipulation, the ARS (but not the DQS or IMC) reliably increased power, yielding consistent benefits for detecting the manipulation effect across a range of sample sizes. These results start to suggest that longer scales might show greater precision in screening out excessively inattentive respondents, consistent with a larger body of measurement work on self-report scales assessing other constructs (e.g., Funk & Rogge, 2007).

4.4.4.3. Power gains on the soda manipulation using the ARS, DQS and IMC. As seen in Table 7, screening out inattentive respondents with the ARS-33, ARS-18, and DQS increased statistical power (by approximately 5.3%, 4.8%, and 2.8% respectively) for the soda manipulation. Once again, these results suggest that the use of more items to screen for inattention might lead to greater increases in precision in that process. Examining the ARS subscales separately, the resampling analyses suggested that the infrequency subscales were particularly effective at screening for problematic inattention on this task. This manipulation involved subjects reading a lengthy paragraph, noticing a small phrase within that paragraph, and then actively typing in a dollar amount (rather than just clicking a radio button, as with the sports ticket manipulation). In contrast to the findings with the sports ticket manipulation, the resampling results suggested that removing participants failing the IMC led to an even larger boost to statistical power on this task (by approximately 9.2%). Thus, on the text manipulation making use of a longer paragraph (and therefore more directly comparable to the IMC itself), the IMC demonstrated a marked boost to statistical power.

4.4.4.4. Power gains for gender differences in openness using the ARS, DQS and IMC. The prior power analyses focused on experimental manipulations. To diversify our operationalization of effects that would be of interest to researchers and thereby examine how the power benefits of screening for inattention might generalize beyond text-based manipulations, we conducted an additional set of resampling analyses to obtain similar power gain estimates for detecting a gender difference on the personality trait of openness to experience. Past research using the BFI has found that men in the United States tend to score *slightly* higher than women on openness to experience as measured by the BFI (e.g., Schmitt,

[6] To establish cut-scores for the continuous indices of inattention in these resampling analyses, we tested various cut-scores which identified between 3% and 10% of respondents as inattentive, and report the cut-scores that yielded the greatest average gain in power. This yielded final cut-scores of < 16 min for time spent completing the survey, > 2 for number of directed questions missed on the DQS, < $-.65$ for psychometric antonyms, > 7 for the long string index, < $-.03$ for psychometric synonyms, < .30 for even-odd consistency, and a significant χ^2 (p < .001) for multivariate distances. Interestingly, the cut-scores for the last 4 indicators listed converged with cut-scores used in the previous literature (e.g. Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012). Although we tried the cut-score of < $-.03$ suggested for psychometric antonyms (e.g., Huang et al., 2012; Johnson, 2005), that cut-score excluded 25% of the sample in Study 3 and led to estimated drops of 4-9% in power to detect the sport and resort manipulation effects (respectively) within the resampling analyses.

Table 7
Summary of resampling analyses (aggregated across n = 60–200) for all inattention indicators.

Inattention indices (cut-score used)	Power gain (aggregated across sample sizes) (%)						Average % excluded
	Sports task	Soda task	Gender difference in openness			Average power gain	
			Study 1	Study 2	Study 3		
ARS-33 (identified by at least one subscale)	5.0	5.3	2.8	14.6	6.2	6.8	7.5
11 Infrequency items of the ARS-33 (>11.5)	1.1	6.2	3.0	8.0	4.1	4.5	5.2
22 Inconsistency items of the ARS-33 (>10.5)	5.5	1.0	0.1	8.9	3.4	3.8	3.7
ARS-18 (identified by at least one subscale)	3.4	4.8	0.4	12.2	4.5	5.1	6.6
6 Infrequency items of the ARS-18 (>7.5)	1.1	6.6	1.6	8.5	2.3	4.0	3.8
12 Inconsistency items of the ARS-18 (>6.5)	2.5	0.0	−1.4	3.4	3.2	1.5	3.5
DQS 7-item scale (3 or more mistakes)	−1.6	2.8	1.7	–	3.1	1.5	3.4
3 Directed questions (2 or more mistakes)	−1.6	1.9	2.4	–	4.2	1.7	3.4
1 Directed question (1 mistake)	−1.7	3.2	2.1	–	3.2	1.7	3.6
Additional indices of inattention							
Latent Classes 2 and 3 (membership in either)	−0.2	5.1	1.5	7.5	3.4	3.5	3.7
Time spent completing survey (<16 min)	2.5	2.9	0.3	–	2.3	2.0	6.8
IMC (fails)	−10.5	9.2	–	–	7.0	1.9	26.3
Long string index (7 or more identical answers in a row)	−5.3	2.0	−0.1	3.3	4.8	0.9	5.5
Psychometric synonyms (<−.03)	−1.7	−3.4	1.9	2.9	2.9	0.5	3.4
Multivariate outliers ($\chi^2(54)$'s > 91.87, $p < .001$)	1.3	−2.1	−0.6	2.9	0.5	0.4	4.2
Psychometric antonyms (<−.65)	1.2	−0.4	−1.2	−0.8	−1.9	−0.6	5.3
Even–odd consistency (<.30)	−5.3	−2.7	3.7	−2.3	0.6	−1.2	8.2
PAI (identified by at least one subscale)	–	–	–	1.3	–	1.3	6.1
8 Infrequency items of the PAI (>14.5)	–	–	–	0.3	–	0.3	2.0
12 Inconsistency items of the PAI (>8.5)	–	–	–	0.5	–	0.5	4.4

Note: For each additional index of inattention, a range of cut scores (identifying 3–10% of subjects as inattentive) was tested to find the cut score giving the optimal boost in power. The power gains presented represent the increased (or decreased) fraction of resamples showing significant results after cleaning with each index of inattention (subtracting the percentage of resamples that were significant without any cleaning).

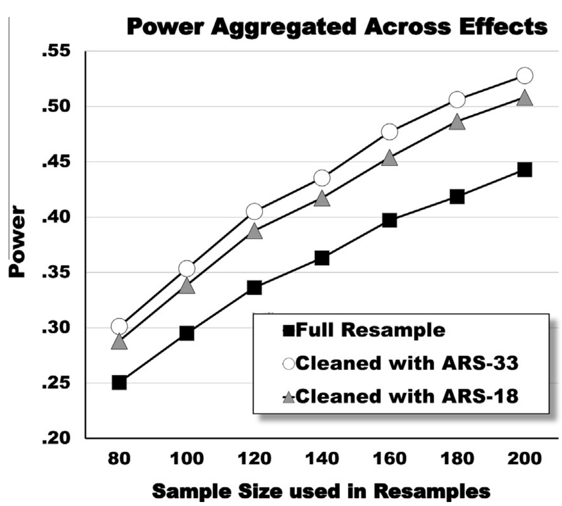

Fig. 2. Influence of screening for inattention on power to detect manipulation effects.

Realo, Voracek, & Allik, 2008). We chose to focus on this subtle gender difference as its modest effect size would yield lower power across the sample sizes tested in the resampling analyses, increasing sensitivity to detecting the benefits of screening for inattention. We replicated this gender difference in the current sample, finding that men (M = 3.81, SD = 0.66) scored slightly higher on openness to experience than did women (M = 3.69, SD = 0.68), $t(745) = 2.14$, $p = .03$, yielding a small effect size (Cohen's d = 0.17). As before, we conducted resampling analyses with 2000 resamples in each of 8 sample sizes (ranging from N = 60 to N = 200), measuring the number of times this analysis yielded a significant gender difference (at $p < .05$) within each resample in the full resample and then after removing participants flagged as inattentive by each indicator. We were also able to conduct these same resampling analyses in Studies 1 and 2 to evaluate their consistency across samples.

As seen in Table 7, screening out inattentive respondents with the ARS-33 and ARS-18 subscales increased statistical power to detect this gender difference in openness to experience across all three studies. Similarly, screening out inattentive respondents with the 7-item DQS consistently increased statistical power to detect this gender difference. In fact, even a 3-item version of the DQS or the use of single DQS items (presented as an average of the gains achieved using each of the 7 items separately in the row of Table 7 labeled "1 Directed Question") yielded slightly more modest power gains. Although the use of just 1–3 DQS items led to slight power *losses* for the sports task, these results begin to suggest that even the use of just a few items for screening inattention might improve data quality. Removing participants failing the IMC also boosted power in the Study 3 resampling analyses (the only study to contain the IMC), suggesting that even with its markedly higher rate of exclusion it was beneficial in helping to uncover this subtle gender difference.

4.4.4.5. Power gains from using other inattention indices. As seen in Table 7, the remaining indices resulted in smaller power gains (e.g., time spent on survey) or gains in power on some analyses but not others (e.g., latent classes, directed questions, psychometric antonyms, multivariate outliers, long-string index, IMC, psychometric synonyms and even–odd consistency) when used to screen out excessively inattentive respondents. Notably, whereas psychometric synonyms decreased power on both experimental manipulations, it increased power for detecting the gender difference in openness across all three samples. These results suggest that screening for inattentive responding using psychometric synonyms might improve power primarily for analyses using the scales on which the psychometric synonyms index is calculated. Taken as a whole, these results suggest that the ARS (and to a slightly lesser extent the DQS) scales were particularly effective at improving power across all of these effects.

4.4.4.6. Power gains from combinations of inattention indices. We ran additional exploratory analyses on power gains with the sports and soda experimental manipulations to examine various combinations of scales. Using other indices in combination with the ARS-33 did not lead to greater gains in power (relative to simply using the ARS-33 alone), with one exception—cleaning the data by removing respondents identified as inattentive based on either time spent on the survey or the ARS-33 (roughly 12–13% of the samples) led to greater gains in power across both manipulations, with average power gains of 7% for the sports ticket manipulation and 5% for the soda manipulation. Adding time spent on survey with either of the ARS-33 subscales individually also yielded slightly larger power gains (compared to the ARS-33 alone) across the two manipulations. We examined more than 20 additional combinations of the indicators of inattention (e.g., time spent on the survey used in conjunction with the DQS). In all of the remaining analyses, these combinations either did not increase power after averaging across the two manipulations, or the power gains were smaller than would be obtained by simply using one of the indicators by itself. In several cases, the combinations decreased power considerably more than the indicators used individually. To give a specific example of one such combination, using time spent on the study, the even-odd measure, and multivariate distances to screen out inattentive respondents (a specific suggestion offered by Meade and Craig (2012)) *decreased* power by an average of 8.6% in the sports manipulation task and by 1.3% in the soda manipulation task across the sample sizes tested in these resampling analyses. These decreases in power are likely due to the combination identifying a slightly larger proportion (over 15% of participants) as inattentive. Overall, these secondary analyses examining combinations of indicators suggested that such combinations tended to be less effective than using the individual indicators separately.

However, when taken as a set, the results of these resampling analyses with individual indicators are consistent with the internal consistency and regression analyses of Studies 1 and 2, suggesting that inattentive respondents provide poor quality data that could obscure meaningful results. Furthermore, the resampling analyses suggested that the ARS (as well as the ARS used in conjunction with time spent on the survey) served as a particularly effective tool to remove highly inattentive respondents, yielding as much as a 7 percentage point gain in power.

4.4.4.7. Power gains from Single DQS items. Finally, to determine the minimum number of items that could effectively screen for inattention and provide boosts to statistical power, we ran additional exploratory analyses on power gains when screening with individual DQS items. The first 3 rows of Table 8 present results for the ARS-33, the 7-item DQS and a 3-item version of the DQS (using three items presented roughly at the beginning, middle and end of the survey) from Table 7 to provide a point of comparison. As Study 2 did not contain the DQS, the analyses in Table 8 present resampling analyses in Studies 1 and 3 only. As seen in the lower portion of the table, several of the individual items of the DQS performed as well as the full 7-item DQS when used individually to screen for inattention. In fact, the very first item of the DQS (presented on page 3 of the survey) was particularly effective at boosting power for these effects. Although these items still show slight power *losses* on the sports ticket manipulation and do not offer power gains as strong as those estimated for the longer ARS scales, these results do indicate that researchers might be able to obtain improvements in the power for their analyses by screening for inattention with as little as a single item.

5. Study 4

In the previous studies, the various indices of inattention consistently identified approximately 3–9% of the sample as providing highly inattentive responses. Other estimates of inattention rates in the existing literature have varied widely (e.g., 3.5% in Johnson, 2005; 10–12% in Meade & Craig, 2012; 35–46% in Oppenheimer et al., 2009). In part, this wide range of estimates could be due to the diverse and often inconsistent ways that inattention has been conceptualized and measured across studies. However, the proportion of inattentive subjects is likely to vary considerably

Table 8
Item-level resampling power analyses in the DQS.

Scale/item text	Study 3 locations in survey				Power gains when used to screen (%)				
					Study 3			Study 1	Average power yield
	Survey page	Distance in pages from sports task	Distance in pages from soda task	Distance in pages from BFI	Sports ticket task	Soda task	Openness gender diff.	Openness gender diff.	
Scales									
ARS-33	4, 12	–	–	–	5.0	5.3	6.2	2.8	4.8
DQS full 7 items	See below	–	–	–	−1.6	2.8	3.1	1.7	1.5
DQS 3-items	3, 9, 20	–	–	–	−1.6	1.9	4.2	2.4	1.7
DQS items									
I read instructions carefully. To show that you are reading these instructions, please leave this question blank.	3	−11	−2	−6	−2.3	3.2	6.3	3.0	2.6
Please skip this question	4	−10	−1	−5	−1.6	5.1	3.4	2.8	2.4
This is a control question. Leave this question blank.	9	−5	4	0	−0.7	3.0	3.5	3.3	2.3
Please skip this question	9	−5	4	0	−1.1	2.1	3.6	2.0	1.7
This is a control question. Mark “Mostly True” and move on.	12	−2	7	3	−2.9	3.0	1.2	−0.8	0.1
This is an extra line. Leave this question blank.	19	5	14	10	−1.5	3.0	1.9	2.5	1.5
This is a control question. Mark “Rarely” and move on.	20	6	15	11	−2.0	3.0	2.9	2.2	1.5

Note: BFI = Big Five Inventory. ARS = Attentive Responding Scale. DQS = Directed Questions Scale. As with Table 7, these resampling analyses involved 2000 separate resamples for each of 8 sample sizes ranging from $n = 60$ to $n = 200$. The power gains presented represent the increased (or decreased) fraction of resamples showing significant results (averaged across all of the sample sizes tested) after cleaning with each index of inattention (subtracting the percentage of resamples that were significant without any cleaning).

across studies, depending on the type of sample (e.g., community volunteer samples vs. undergraduate participant pool samples), incentives (none, extra credit, monetary incentives, individualized feedback), characteristics of participants (motivation, conscientiousness, distraction, interest in the study topic), and characteristics of the study itself (topic, length, variety and novelty of tasks involved, level of supervision). While rates of inattention may be negligible in a brief study on a novel and highly interesting topic (e.g., a 5–10 min survey on casual sex), inattention may be much higher in a lengthy (e.g., 30–40 min) survey on a more mundane topic. The fact that the current studies were relatively brief and included a variety of tasks may help to explain why the current estimated rates of inattention fall on the lower end of other published estimates. Furthermore, including multiple measures of inattention and self-reported questions about typical responding behavior may have encouraged greater attentiveness by drawing participants' attention to their own responding behavior.

Although the current studies still offer compelling evidence to suggest that inattentive respondents provide poor-quality data, the low rates of inattention in those studies might have served to underestimate the effects of inattention on correlational and experimental results. Thus, in the final study, we wanted to simulate a broader range of inattentive responding (e.g., as low as 1% and as high as 25% of the sample), allowing us to more precisely examine two central issues in a manner that might generalize to a broader range of study designs. First, these simulated data allowed us to estimate the effect of varying proportions of inattentive responding on the power to detect experimental effects (and on the resulting effect sizes), which to our knowledge has not been explored in previous research. Second, the simulation analyses allowed us to estimate the gains in power and effect size that may be obtained by screening for inattentive responding under varying rates of inattention. We expected that increased levels of simulated inattention would markedly decrease statistical power and yield smaller estimates of experimental effect sizes. We further hypothesized that screening for inattention would mitigate some of this loss of power and yield more accurate and robust effect size estimates, especially in samples with higher rates of simulated inattentive responding.

5.1. Study 4 method

5.1.1. Simulation-resampling strategy

5.1.1.1. Simulating inattention. Drawing upon the original data from Study 3, we started with 466 robustly attentive respondents who were identified as sufficiently attentive on the 5 indices used in the current study (see below). In addition to these 466 robustly attentive responses, we also created a large pool of simulated inattentive responses by creating 10 copies of the original dataset and then adding random error to degrade each response using a strategy similar to Meade and Craig (2012). In these simulated inattentive responses, we randomly selected 75% of the items *in each row of data* to be degraded by replacing those items with randomly generated values (within each item's range) based on a uniform distribution.[7] This data degradation was done across 10 separate copies of the dataset to ensure that any results obtained from the subsequent simulation-resampling analyses would be less likely to be influenced by any unusual error variance created within a single degraded dataset or individual. This strategy resulted in a final dataset of 466 robustly attentive responses and 4660 simulated inattentive responses that we drew upon in the subsequent simulation-resampling analyses.

5.1.1.2. Resampling analyses with controlled proportions of inattention. After creating the pools of attentive and simulated inattentive responses, we then performed resampling analyses highly similar to those presented in Study 3 across a range of sample sizes ($N = 100$, 120, and 140). However, in contrast to Study 3, we programmed a simulation macro to create each resample by randomly selecting (with replacement) the appropriate number of cases from among the 466 attentive and 4660 inattentive cases to yield various proportions of inattentive responses within each simulated resample (0%, 1%, 5%, 10%, 15%, 20%, and 25%). We created a total of 2000 resamples for each simulation condition. Within each resample, we tested the experimental manipulation effects: (1) in all participants in the resample and (2) after removing participants identified as inattentive (separately for each indicator of inattention). As in Study 3, we recorded the number of times across the 2000 resamples in each simulation condition that the manipulation yielded a statistically significant effect in the predicted direction (at $p < .05$) as an estimate of statistical power.

5.1.2. Experimental manipulations

We analyzed the experimental manipulations used in Study 3 (from Thaler, 1985). Specifically, we examined a sunk cost effect in which participants answered how likely they would be to attend a football game on a cold day if they had paid for the tickets, or if the tickets were free. In the second experimental manipulation, we asked participants how much they would pay for a soda for a friend from either a run-down grocery store or a resort. As in Study 3, scores for the second task were truncated to a maximum of $10. When this item was selected to be randomized in the simulated random data, we generated a random value between $0 and $10.

5.1.3. Indices of inattention examined

Because our data simulation strategy involved adding random responses to a subset of items, we focused only on those indicators of inattention computed directly from responses to Likert-scale type items. Therefore, we used the following indicators of inattention in these simulation analyses: the ARS-33 and ARS-18 subscales, Psychometric Antonyms, Psychometric Synonyms, and the Even–Odd index,[8] using the cut-scores optimized in Studies 1–3. Given that the randomized degradation used to simulate inattention was incompatible with the underlying approach of the DQS (which asks participants to skip several items), the DQS was not evaluated in these simulation analyses.

5.2. Study 4 results

5.2.1. Identifying simulated inattention

The resampling analyses gave nearly identical results across the three sample sizes tested, and so aggregate results are presented. The ARS-33 inconsistency and infrequency subscales identified 94% and 88% of simulated inattentive responses, respectively. The shorter ARS-18 inconsistency and infrequency subscales identified 80% and 69% of these responses, respectively. When the

[7] We chose to degrade 75% of the items in each row based on results from studies 1 and 3, in which participants making up the inattentive latent classes missed roughly 78-79% of the directed questions (an average of 5.45 and 5.54 mistakes out of 7 questions).

[8] We did not calculate multivariate outliers in these analyses, as these analyses included 84,000 resamples (3 sample sizes x 7 levels of inattention x 2 experimental manipulations x 2,000 resamples each), and it would have been computationally taxing to calculate multivariate distances within each resample. We were unable to include the IMC, the long string index, directed questions or time spent completing the survey in these analyses as their formats (i.e., clicking on a heading in the IMC, selecting the same answer for a block of questions, leaving some questions blank when instructed, a quantity of time) would have required different simulated degradation strategies than those used on the self-report responses in the current study.

Table 9
Average power gain in simulation analyses (aggregated across n = 100–140) for inattention indicators.

Inattention indices (cut-score used)	Sports ticket task							Soda task						
	Proportion of simulated inattentive subjects							Proportion of simulated inattentive subjects						
	0%	1%	5%	10%	15%	20%	25%	0%	1%	5%	10%	15%	20%	25%
Power in the uncleaned resamples	.737	.719	.669	.604	.533	.487	.412	.803	.776	.712	.656	.571	.509	.437
Power gain after screening out inattentive data														
ARS-33 (identified by at least one subscale)	–	.011	.042	.086	.126	.163	.184	–	.022	.063	.106	.158	.191	.235
11 Infrequency items of the ARS-33 (>11.5)	–	.009	.036	.076	.101	.136	.158	–	.016	.051	.087	.119	.149	.171
22 Inconsistency items of the ARS-33 (>10.5)	–	.011	.039	.081	.115	.148	.166	–	.021	.064	.099	.148	.180	.218
ARS-18 (identified by at least one subscale)	–	.012	.040	.083	.117	.153	.178	–	.021	.061	.101	.148	.185	.217
6 Infrequency items of the ARS-18 (>7.5)	–	.009	.024	.056	.082	.098	.111	–	.012	.048	.073	.097	.120	.137
12 Inconsistency items of the ARS-18 (>6.5)	–	.010	.034	.072	.098	.121	.148	–	.020	.052	.082	.112	.146	.165
Additional indices of inattention														
Even–odd consistency (<.30)	–	.005	.014	.033	.043	.048	.061	–	.007	.028	.047	.066	.074	.086
Psychometric synonyms (<−.03)	–	.003	.012	.033	.042	.046	.058	–	.008	.026	.047	.062	.071	.081
Psychometric antonyms (<−.65)	–	.003	.016	.032	.045	.057	.059	–	.010	.028	.043	.056	.075	.086

Note: The analyses presented here use the same resampling method to that presented in Tables 7 and 8 except that in these analyses differing levels of inattention are simulated in the datasets during that resampling process. This allowed us to examine how higher levels of inattention impact power, effect sizes, and the power gains obtained from screening for inattention with a variety of indices.

infrequency and inconsistency subscales were used jointly, the ARS-33 identified 99.5% of these responses and the ARS-18 identified 94%. These results continue to suggest higher levels of precision from the use of longer scales to screen for excessive inattention. The other indicators of inattention identified a smaller proportion of simulated inattentive responses (50% for Even–Odd Consistency, 47% for Psychometric Antonyms, and 41% for Psychometric Synonyms).

5.2.2. Effects of inattention on power and effect sizes

As shown in Table 9, the simulation analyses demonstrated that the power to detect a significant effect on both the sports ticket and soda tasks decreased as the proportion of inattentive respondents increased. With no simulated inattentive respondents, the sports ticket and soda tasks were significant an average of 74% and 80% of the time (respectively) across the 3 sample sizes tested. With just 5% of responses simulated to be inattentive, power dropped to 67% and 71% for these tasks, respectively. Thus, even fairly low levels of simulated inattention (well within previously published rates of inattention and therefore likely to be anticipated in a majority of published research) can lead to notable losses of statistical power. With higher levels of simulated inattention (i.e., 25% of respondents) the effects of inattention were more pronounced as power decreased to 41% and 44%, respectively – nearly cutting the ability to detect these meaningful effects in half. These results were consistent with our hypothesis that higher levels of inattention would have increasingly detrimental effects on power. These simulation analyses also revealed that increasing levels of inattention were associated with decreases in the corresponding effect size estimates in each of the resamples. As seen in Fig. 3, the simulation analyses suggested that a rate of 25% inattentive respondents was associated with effect size estimates for the differences between the two experimental groups dropping from .49 to .33 on the Sports Ticket task and from .53 to .34 on the Soda task. Thus, consistent with our hypothesis, increased levels of simulated inattention undermined the ability to accurately estimate effect sizes, yielding biased effect size estimates.

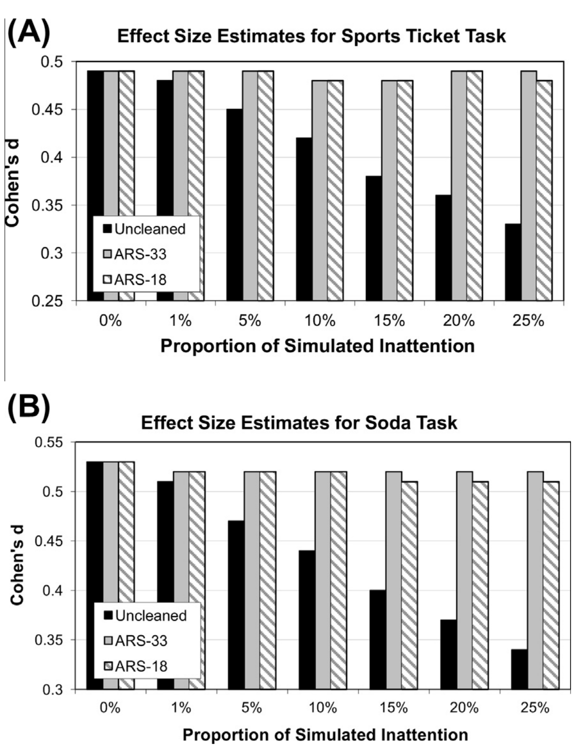

Fig. 3. Influence of screening for inattention on effect size estimates.

5.2.3. Effects of screening out inattention on power and effect sizes

The bottom portion of Table 9 shows that removing responses identified as inattentive using either the ARS-18 or the ARS-33 increased power across all proportions of simulated inattention – helping to mitigate the losses in power associated with higher levels of inattention. When a modest proportion of the data was simulated to be inattentive (i.e., 5%, corresponding roughly to that observed in Studies 1–3), the power improved between .04 to .06, consistent with the power gains estimated from the resampling analyses in Study 3. When a larger proportion of the sample was simulated to be inattentive, the power gains were more pronounced (e.g., as large as .18 for the sports ticket task and .24 for the soda task). Cleaning the data using the other indicators of inattention also consistently increased power in these simulations, although the gains in power (of as much as .09) were smaller than those afforded by the ARS-33 or ARS-18. Thus, consistent with our hypothesis, these results suggested that it might be more advantageous to screen for inattention in studies likely to have higher levels of inattentive responding. As hypothesized, screening for inattention also mitigated the negative effects of inattention on estimated effect sizes for group differences between the experimental conditions. As seen in Fig. 3, using either the ARS-33 or the ARS-18 to screen out inattentive respondents yielded more accurate and stable estimates of effect size across all of the simulated proportions of inattention examined.

6. General discussion

Not all participants read each and every item on a scale carefully before providing a response. Although inattention is likely to add error variance to studies, it has only recently begun to receive more direct methodological attention within the literature (e.g., Meade & Craig, 2012). Furthermore, despite recent work (e.g., Oppenheimer et al., 2009), its full impact on data quality and statistical power have yet to be assessed. Across 4 studies, we examined the effects of inattention on basic aspects of psychological research (e.g., compliance with study tasks, self-report data quality, correlational analyses, experimental manipulations, gender differences, and statistical power). Building on the multivariate analyses of Meade and Craig (2012), the present studies developed and validated the ARS-33, ARS-18, and DQS to provide researchers highly efficient and effective methods of identifying inattentive respondents. Analyses across all 3 samples suggested that subjects identified as inattentive by either latent profile analyses, the ARS or the DQS were markedly less compliant with study tasks and provided poorer quality self-report data (i.e., lower levels of internal consistency on well-validated scales), obscuring regression results and undermining experimental manipulations. Resampling analyses further suggested that using the ARS and time spent on the survey to remove highly inattentive participants provided 5–7 percentage point gains in power, outperforming the other indicators tested. Even the use of a single DQS item led to 2–3 percentage point gains in power. Resampling analyses based on differing proportions of simulated inattentive data extended these findings by revealing that higher proportions of simulated inattention markedly reduced both power (dropping as much as 36 percentage points) and the resulting effect size estimates (e.g., dropping from .53 to .34 on the Soda task) for the two experimental manipulations examined. The simulation-resampling analyses further suggested that the losses in power and effect size caused by simulated inattentive responding were notably mitigated by screening out inattentive responses, increasing power by as much as 24 percentage points. We would argue that the diverse methodologies examined across these 4 studies (e.g., correlational analyses, experimental manipulations, gender differences, and common study tasks) are directly relevant to a very broad range of study designs. Taken as a set, the results of these studies build on previous work by quantifying the impact of inattention on research and demonstrating the utility of screening for inattention across a large majority of the research being conducted in our field.

6.1. Implications

6.1.1. Inattention exists

Despite the robust prevalence of research relying on self-reported data, relatively little attention has been given in the published literature to levels of inattention on self-report scales. Consistent with recent work (Johnson, 2005; Meade & Craig, 2012), the current findings suggest that roughly 3–9% of respondents exhibited problematic levels of inattention across the studies presented. These estimates converge with estimates using similar indicators of inattention. For example, 3.5% of respondents in a sample of over 20,000 participants completing an Internet survey with over 300 items were found to have repeatedly selected the same response option many times in a row (Johnson, 2005).

6.1.2. Inattention is distinct

The current results suggest that inattention is a distinct construct from social desirability, demonstrating weak negative associations with impression management and self-deception. This result was consistent with the conceptualization of social desirability, as it requires sustained attention to effectively present oneself favorably across an entire survey with varying item content. The correlations in Table 2 present a preliminary profile of inattentive responders as people who tend to be less agreeable, less conscientious, and less open to experience, as well as less internally motivated to participate in research. Although this suggests that the use of recruitment methods pulling for internal (rather than external) motivation might yield higher quality data, it should be noted that even these correlations were modest to weak in magnitude. Thus, inattention would seem to be fairly independent of personality traits, self-esteem, social desirability and motivations behind participating in research, requiring separate methods of assessment/screening. Future research should examine a broader set of individual differences (e.g., personal relevance of the study, attention capacity) and methodological factors (e.g., inclusion criteria, study format, study length, incentives, item redundancy, recruitment sources) to fully characterize the proximal correlates of inattention.

6.1.3. Inattention can be measured

The results across our 4 studies support the utility of several indicators at identifying highly inattentive participants, demonstrating strong criterion validity with behavioral markers of inattention and identifying a set of respondents contributing largely error variance to substantive analyses. These results stand in contrast to the findings of Piedmont and colleagues (2000), who questioned the utility of validity scales. Although they presented their findings largely as a blanket condemnation of all validity scales, we believe those results might simply have highlighted the need to develop more effective measures of inattention. The results presented here offer compelling evidence to suggest that the ARS, the DQS, and a few other indices were effective at identifying inattentive respondents. The results presented also offered support for the multivariate latent classes of inattentive responding identified by Meade and Craig (2012), and demonstrated that the ARS and the DQS did fairly thorough jobs of capturing this multivariate phenomenon in easily implemented scales.

6.1.4. Inattention can affect data quality and analyses

Using methods common to both correlational survey research as well as experimental research, our results revealed that meaningful findings present within the attentive respondents were not present among the inattentive respondents. Specifically, internally consistent scales lacked internal consistency, regression analyses failed to replicate previous findings and experimental manipulations embedded in paragraphs failed to produce significant effects within inattentive respondents. These results are consistent with a growing body of findings (e.g., Johnson, 2005; Meade & Craig, 2012; Oppenheimer et al., 2009; Woods, 2006), suggesting that inattention can obscure both experimental manipulations and correlational results. Taken as a set, these results suggest that high levels of inattention could result in largely meaningless data, primarily adding error variance (i.e., noise) to analyses. Furthermore, inattentive participants spent less time and effort on tasks commonly used as experimental manipulations (e.g., viewing video clips, reading paragraphs, engaging in a supraliminal priming task), again suggesting that the inclusion of such respondents would serve to undermine the effectiveness of manipulations. Simulation analyses further suggested that the drops in power and estimated effect sizes were intensified with increasing levels of inattention, highlighting the negative effects of inattention.

6.1.5. Screening out inattentive subjects can improve power

The resampling analyses suggested that using the ARS to screen for inattention yielded consistently higher power (averaged across effects) for detecting the experimental and group difference effects examined. Thus, removing this source of error variance could generally augment researchers' ability to detect meaningful results in self-report data. Several other indices (including the DQS) also yielded gains in power on one or more of the effects tested, suggesting that they also identified respondents with data sufficiently problematic to obscure results. Furthermore, even a single directed question from the DQS improved power by 2–3%. Although weaker than the effects for the longer ARS-33 (which improved power by nearly 7%, on average), these results suggest that data quality may be improved by adding just a single item to a study. However, the gains for those indices tended to be smaller in magnitude and/or less consistent across the effects examined. As the ARS subscales were among the few indices that effectively identified both latent inattentive classes and consistently improved power on both tasks, these results begin to suggest that the ARS might represent a broadband measure of inattention – effectively identifying a majority of the respondents with problematic data.

6.2. Recommendations

6.2.1. Researchers should screen for inattention

Given the demonstrated impact of inattention on correlational/experimental analyses, we would recommend that researchers routinely screen their datasets for inattentive respondents. This could be as simple as excluding participants who completed the study in less than half of the average time (an index that would need to be adjusted/optimized in each new study). In the samples presented, the optimized cut-scores for time spent on the survey roughly corresponded to half of the 5% trimmed mean completion time (excluding clear outliers to the distribution). Given the superior performance of the ARS in the current studies, we would recommend its use whenever possible (see Supplementary appendix).[9] The use of these scales would be of greatest value in studies with modest effect sizes or small sample sizes (making power a critical factor to the success of the study), particularly in studies requiring researchers to invest large numbers of person-hours into each subject. Thus, if a study involves lengthy lab sessions, extensive interventions, or multiple waves of longitudinal follow-up assessment, then the use of the ARS in a baseline assessment could not only help to improve data quality and boost power, but it could also save critical resources on that project. Nonetheless, we acknowledge that shorter scales will be desirable in many research contexts. Toward this end, our results also supported the utility of the 11-item or 6-item infrequency subscales of the ARS-33 and ARS-18 (respectively), as these scales demonstrated convergent validity and practical utility, consistently improving power when used to screen datasets for excessive inattention.

6.2.2. Even a single item can be effective

The power analyses reported in Study 3 also supported the use of a single item from the DQS to screen for excessive inattention as several of the individual items of the DQS led to increases in power on most of the effects examined. It is worth noting that the three items yielding the largest power gains were presented in the first half of the survey and all three items specifically asked respondents not to respond to them. In fact, the item showing the highest power gains (the first DQS item) was also the most subtle item – embedding the request to skip the question *after* what seemed to be a normal item stem (''I read instructions carefully. To show that you are reading these instructions, please leave this question blank.''). Although it remains unclear which of these factors (placement in the first half of the survey, asking subjects to skip the question, embedding that request after a seemingly typical item stem) might have led to the improved efficacy at screening out problematic respondents, these results give future researchers some tentative guidelines for using single items as inattention screens in studies. In fact, the finding that single items can be effective screening tools opens up a number of additional applications for the current work. Whereas it would not be practical to include even just a 6- or 7-item inattention scale in a daily diary packet, it would be quite possible to include a single directed question (i.e., the 1st item of the DQS) in daily diary surveys – potentially placing the item at different points in the survey on different days. To ensure that the item remains a novel test of inattention, researchers could even change the item stem (''I read instructions carefully'') on different days of the diary study to more closely match the content of the surrounding questions in which it is embedded that day.

6.2.3. Researchers should be careful about combining inattention indices

Based on the secondary analyses in Study 3, we would also recommend that researchers use indicators of inattention judiciously. In our analyses, combining time spent on the survey with the ARS-33 or either of its subscales alone led to larger gains in power, supporting such combinations. However, our analyses further suggested that other combinations of inattention indices did not lead to larger gains in power than just using a single index (and, in some cases led to decreased power). Thus, using combinations of multiple indicators (as suggested by Meade and Craig (2012)) could prove counterproductive. At a minimum, the results presented suggest that if researchers were going to use a combination of multiple indices, they would need to carefully optimize and validate the corresponding cut-scores used for each index so that the multivariate combination would successfully enhance (and not reduce) power. In fact, if researchers wanted to take a multivariate approach to screening for inattention (without adding additional items to a study by using statistical post hoc indicators), the results presented would support using latent profile analyses to identify the most excessively inattentive respondents. However, this would require researchers to: (1) calculate the multiple statistical indices of inattention employed in the current studies, (2) tailor each index to the specific items in each study, and (3) run latent profile analyses to identify excessively inattentive respondents. By comparison, the ARS combined with time spent on the study (or even just the first item of the DQS) offers a much simpler and potentially more effective alternative.

6.2.4. Instructions should be used cautiously

Study 1 found that 84% of participants report occasionally skipping instructions, and 19% report skipping instructions in surveys more than half the time. Participants may be even more likely to skip lengthy instructions that appear unnecessary, as evidenced by the proportion of participants who fail the IMC (35–46% in Oppenheimer et al., 2009). This suggests that researchers should be judicious in their use of text instructions and text-based manipulations (keeping them as succinct and engaging as possible), with the understanding that a sizable portion of participants will simply skip lengthy instructions, potentially adding error variance to the associated measures.

6.3. Limitations and future directions

Several limitations of the current studies are worth noting. First, although some participants were recruited from participant pools

[9] The supplementary appendix (available online) lists all of the ARS items along with scoring information and syntax.

and completed the study in the lab, a majority of participants were recruited online and completed the study online. As a result, it is possible that some results may not generalize to research conducted with paper surveys or studies conducted in person. Future studies should examine inattention across a broader range of methods to further examine study design factors that might moderate the current findings. Second, the samples were predominantly Caucasian. Although we have no theoretical reason to suspect that the effects of inattention on research would fundamentally differ across racial and ethnic groups or other demographic variables, we did find that inattention and non-compliance were slightly higher in non-Caucasian respondents, suggesting possible mean differences in rates of inattention across demographic groups. Future studies should examine inattention in more diverse samples to ensure that the effects of inattention remain relatively consistent across demographic groups. Third, our measures and analyses focused on a specific type of problematic responding (i.e., the type of careless or inattentive response that might result from answering questions randomly). As a result, our strategy of degrading data in Study 4 by randomizing a subset of items in order to simulate inattentive responding did not model all of the forms of careless and biased responding that are likely to appear in research studies. Nonetheless, these simulation results serve as a first step in understanding how various levels of inattention influence data quality and statistical power. These simulation analyses also helped address a fourth limitation—that including multiple measures of inattention and responding behavior may have encouraged greater attentiveness. As a result, our results may underestimate the prevalence of inattentive responding and the benefits of screening out extremely inattentive respondents. Fifth, it is possible that the sensitivity of the post hoc indicators might have been limited in the current samples as they were based on 54 items rather than 300 + items (e.g., Johnson, 2005). We would argue that our results are relevant to the large majority of published research. However, post hoc indicators might demonstrate greater benefits as screening tools in longer surveys as they offer a much larger item pool on which to draw. Future research might compare the effects of these various indices in studies using larger item pools and with different experimental and correlational analyses. Sixth, our analyses did not systematically examine the effects of inattention at various points in a study. Our resampling analyses on the items of the DQS suggested that identifying inattention near the start of a study might have helped to identify some of the most problematic respondents. However, future research will need to more precisely examine the timing of inattention within a study to determine its full impact on data quality. Finally, researchers may worry that removing participants identified as highly inattentive may decrease external validity. Indeed, Study 1 suggests that highly inattentive participants tend to have slightly lower self-esteem, agreeableness, conscientiousness, and openness to experience. However, our results strongly suggest that these participants provide little more than error variance. Given the small proportion of participants identified as inattentive, we feel that the risk to external validity is low, and is far outweighed by the potential benefits of such screening.

6.4. Conclusion

The results of the present studies call into question the widespread neglect of inattention in research using self-report measures. Our results indicated that 3–9% of participants were highly inattentive, failing to follow instructions, completing self-report scales in a seemingly haphazard and inconsistent manner, and failing to demonstrate both correlational and experimental effects. Results suggested that screening out this source of error variance could markedly enhance power and provide more accurate estimates of effect size – mitigating the widespread negative effects of inattention on research.

Acknowledgments

We thank Soonhee Lee and Elizabeth Baker-Davidson, Maria Saavedra-Finger, Amy Rodrigues, Christine Walsh, Amanda Shaw, and Silvia Marin for helping collect preliminary data on the item pool. We also thank the participants who completed our studies and Harry Reis for his comments on earlier versions of the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jrp.2013.09.008.

References

Bruehl, S., Lofland, K. R., Sherman, J. J., & Carlson, C. R. (1998). The variable responding scale for detection of random responding in the Multidimensional Pain Inventory. *Psychological Assessment, 10*, 3–9.

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. http://dx.doi.org/10.1037/1040-3590.4.3.340.

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

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., et al. (2002). Increasing response rates to postal questionnaires: Systematic review. *British Medical Journal, 324*, 1–9.

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology, 21*, 572–583.

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). "I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science, 10*, 321.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion, 9*, 87–108.

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*, 99–144.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 54*. Berkeley: University of California, Berkeley, Institute of Personality and Social Research.

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

Lautenschlager, G. J., & Meade, A. W. (2008). AlphaTest: A Windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 32*, 502–503.

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

Morey, L. (1991). *Personality Assessment Inventory*. Odessa, FL: Psychological Assessment Resources.

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.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

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

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

Pinsoneault, T. B. (2005). Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory – Adolescent protocols. *Psychological Assessment, 17*(4), 476–480.

Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality correlates of self-esteem. *Journal of Research in Personality, 35*, 463–482.

Rosenberg, M. (1965). *Society and adolescent self-image*. Princeton, NJ: Princeton University Press.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500–504.
Rottenberg, J., Ray, R., & Gross, J. J. (2007). Emotion elicitation using films. *The Handbook of Emotion Elicitation and Assessment*, 9–28.
Saavedra, M. C., Chapman, K. E., & Rogge, R. D. (2010). Clarifying links between attachment and relationship quality: Hostile conflict and mindfulness as moderators. *Journal of Family Psychology, 24*, 380–390.
Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in big 5 personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*, 168–182.
Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135*, 409–428.
Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science, 4*, 199–214.
Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 186–191. http://dx.doi.org/10.1007/s10862-005-9004-7.