General Article

Advances in Methods and
Practices in Psychological Science
January-March 2024, Vol. 7, No. 1,
pp. 1–19
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459241231581
www.psychologicalscience.org/AMPPS

Careless Responding: Why Many Findings Are Spurious or Spuriously Inflated

Morgan D. Stosic[1], Brett A. Murphy[2], Fred Duong[3], Amber A. Fultz[4], Summer E. Harvey[5], and Frank Bernieri[4]

[1]Department of Psychology, University of Maine, Orono, Maine; [2]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [3]Department of Psychology, University of Toronto, Toronto, Ontario, Canada; [4]School of Psychological Science, Oregon State University, Corvallis, Oregon; and [5]Department of Psychology, Northeastern University, Boston, Massachusetts

Abstract

Contrary to long-standing conventional wisdom, failing to exclude data from carelessly responding participants on questionnaires or behavioral tasks will frequently result in false-positive or spuriously inflated findings. Despite prior publications demonstrating this disturbing statistical confound, it continues to be widely underappreciated by most psychologists, including highly experienced journal editors. In this article, we aim to comprehensively explain and demonstrate the severity and widespread prevalence of careless responding's (CR) inflationary effects in psychological research. We first describe when and why one can expect to observe the inflationary effect of unremoved CR data in a manner accessible to early graduate or advanced undergraduate students. To this end, we provide an online simulator tool and instructional videos for use in classrooms. We then illustrate realistic magnitudes of the severity of unremoved CR data by presenting novel reanalyses of data sets from three high-profile articles: We found that many of their published effects would have been meaningfully, sometimes dramatically, inflated if they had not rigorously screened out CR data. To demonstrate the frequency with which researchers fail to adequately screen for CR, we then conduct a systematic review of CR screening procedures in studies using paid online samples (e.g., MTurk) published across two prominent psychological-science journals. These findings suggest that most researchers either did not conduct any kind of CR screening or conducted only bare minimal screening. To help researchers avoid publishing spuriously inflated findings, we summarize best practices to help mitigate the threats of CR data.

Keywords

careless responding, insufficient effort responding, inattentive responding, random responding, data screening

Received 6/12/23; Revision accepted 1/18/24

Dr. Diligence and his team are investigating whether peak-pandemic individual adherence to COVID-19 safety guidelines (masking, not drinking bleach, etc.) was positively associated with religious tolerance. Using self-report measures of both constructs, they run a survey study online. After analyzing the data from all 230 participants, Dr. Diligence is extremely pleased to find the predicted positive association between self-reported COVID-19 safety adherence and religious tolerance ($r = .28$, $p < .001$).[1] A publishable finding seems to be at hand!

Yet after further examination of the data, Dr. Diligence and fellow researchers notice that 30 participants completed the survey in an impossibly fast time, producing response patterns that appear completely random. Viewing these data as merely noise, they assume the relationship between religious tolerance and COVID-19 policy

Corresponding Author:
Brett A. Murphy, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
Email: brettmur@email.unc.edu

adherence will be even stronger once these data are removed. Yet in a shock to Dr. Diligence's team, when those 30 respondents are excluded, the association between the two variables disappears completely ($r = .02$, $p = .79$). They frantically check to see if they have made some huge error. How could adding random noise data to one's analyses create an illusory positive finding?

The conventional wisdom has long been that participants who carelessly respond to questionnaires and other psychological measures merely add random noise to a data set and that such error variance will, at worst, attenuate observed associations between variables.[2] This conventional wisdom reduces the perceived threat of poor data quality because (a) researchers are often more worried about the threat of false positives than false negatives and (b) researchers may be misled into thinking they can compensate for the presence of some careless responding data simply by using larger sample sizes.

Yet numerous researchers have demonstrated that this conventional wisdom is wrong: Data from careless respondents will often create systematic covariance that spuriously inflates associations between individual items (e.g., inflating scale reliabilities; Carden et al., 2019; Wise & DeMars, 2009) and between different multi-item measures, such as self-report questionnaires and cognitive ability tests (Credé, 2010; Holden et al., 2019; Holtzman & Donnellan, 2017; Huang et al., 2015; Wood et al., 2017). In fact, the presence of careless respondents can inflate any kind of statistical estimate based on covariance, including factor correlations, factor loadings, logistic regression coefficients, and so on (King et al., 2018). As we explain, this inflationary phenomenon is extremely common, not rare. The presence of careless responding (CR)[3] participants in the field's data sets has likely generated a very large number of false positive and spuriously inflated results in published literature, especially in an era of unproctored online studies with anonymous paid participants. Beyond false-positive or inflated results in individual studies, this inflation can have further deleterious effects on everything from meta-analytic estimates to calculations of statistical power (e.g., if prior effect sizes are overestimated, then statistical power will be overestimated for subsequent studies that screen out CR).

Even many of the most experienced psychological scientists underappreciate this major threat to the field's validity and replicability. In a preregistered study, we distributed a brief survey to the editorial boards of 10 different psychology journals that frequently publish studies with anonymous online samples (e.g., *Psychological Science*, *Clinical Psychological Science*). In all, 227 board members (median lifetime number of manuscripts as editor or peer reviewer = 150, interquartile range [IQR] = 50–350) gave us their perspectives on the risks of CR data (for more detailed information, see Appendix A). Although most editors (66%) believe that CR will very frequently increase risks of Type 2 errors (dilute effect sizes), only a modest few (19%) believe that it will very frequently increase risks of Type 1 errors (inflate effect sizes).[4] More than half (57%) indicated that they have never asked authors to report screening and excluding of CR; of editors who have done so, only a minority indicated that they always (9%) or often (28%) view such reporting as necessary for an article to be accepted for publication. This leniency is likely due to most editors being unaware that CR poses a Type 1 error risk. Most editors (76%) readily acknowledged that they would benefit from training or educational resources to better understand CR.

Given that several research teams have previously demonstrated the inflationary risks of CR, why has this critical issue failed to enter the general awareness of the psychological-research community? To some extent, it may be that the highly technical writing and/or specialized journal identities of some of these past demonstrations have been limiting factors in promoting this issue to broader audiences. The wide variety of terms used to denote CR (e.g., "indiscriminate responding," "lazy responding," "rapid guessing," "inattentive responding," "insufficient effort responding") may also be a factor leading to "jangle fallacies" (the fallacy of erroneously assuming two constructs are different because they have different names; Kelley, 1927).[5] The largest obstacle, however, may simply be that the threat of Type 2 error posed by CR is an easy intuition to grasp, whereas the threat of Type 1 error caused by CR can feel counterintuitive or even paradoxical. For instance, even general review articles on best practices in addressing CR allot only a few sentences to discussing the risk of inflation (e.g., Ward & Meade, 2023), do not mention inflation risks at all (e.g., Malamis & Howley, 2022), or cite the relevant works in ways that may inadvertently lead a reader to think that attenuation is the only major concern (e.g., see Arthur et al., 2021, p. 110). Even for individuals who are theoretically aware of the inflationary possibilities of CR, the counterintuitiveness of the phenomenon may bias one into assuming it is a rare phenomenon and not something researchers should worry much about.

In this article, we aim to bring greater awareness of CR's inflation risks into general research knowledge and practice in psychology by demonstrating that the inflationary risk of CR is prevalent, severe, and frequently unaddressed by researchers using samples with a high CR risk (e.g., paid online samples of anonymous participants). First, we provide a less technical explanation of the confounding effects of CR and present new introductory educational resources suitable for use in classrooms and lab meetings. Second, to concretely demonstrate the prevalence and severity of these effects in real data, we reanalyzed real data (i.e., not simulated data) from three recent articles published in the *Journal of Personality and Social Psychology* [*JPSP*]; we show that if the authors of these three articles had not rigorously screened for CR, their results would have often been meaningfully (sometimes dramatically) inflated. Third, to illustrate the potential pervasiveness of spurious or inflated findings in current psychological-science literature, we conducted a systematic review of CR screening practices in articles using paid online samples published across two flagship journals (*JPSP* and *Psychological Science* [*Psych Science*]). Finally, to help mitigate CR data's contamination of future published research, we summarize recommendations for identifying, excluding, and reporting CR.

What Is CR?

There are many kinds of problematic participant responses that can affect the validity of research findings, but here we focus on only one: CR, which is any responding that is not attentive to the item content of a survey or test (see Maniaci & Rogge, 2014).[6] CR encompasses random responding, lazily overly consistent responding (e.g., repeatedly giving the same response for long strings of items regardless of item content), and most other response styles that stem from a participant not paying attention to the content of a survey or test. In online samples exclusively, CR also includes data generated from software applications (i.e., bots) that are programmed to respond to survey questions automatically. Some CR produces data that are systematically invariant, such as when a respondent gives the same answer over and over to each survey item; other CR produces data that are less systematic, even approximately random (for an investigation of the analytical differences between different types of CR, see DeSimone et al., 2018). Less systematic (i.e., more random) CR data are thought to be more prevalent (for a discussion, see Ward & Meade, 2023), especially given that many careless responders are likely motivated to not be identified as such.

CR is often, sometimes implicitly and sometimes explicitly, discussed as a trait-like individual difference; a person who is careless in one study will tend to also be careless in other studies. For example, Bowling et al. (2016) observed that respondents who tended to be careless across multiple studies were rated by their acquaintances as being lower in conscientiousness, agreeableness, extroversion, and emotional stability. On the other hand, CR will also often be merely a situation-specific methodological nuisance; even generally careful respondents will sometimes become careless, especially if they are in a distraction-prone environment or if the research study is boring, fatiguing, overly long, and so on (see discussion in Ward & Meade, 2023).

Although CR can be found in any kind of research participant sample, paid online samples (e.g., MTurk, Prolific, Dynata) are perhaps the most susceptible to high rates of CR given that (a) hourly wage rate incentivizes speedy completion and reducing fatigue (i.e., reducing cognitive effort), (b) researchers cannot verify all their participants to be humans (i.e., not bots) or fluent readers of the study's language (e.g., participants in foreign countries using server farms to pose as English-speaking American participants), and (c) researchers cannot assist participants to ensure they are not confused. Indeed, when rigorous screening methods are employed, published studies often identify between 15% and 50% of respondents as CR in paid online samples (e.g., Balzarini et al., 2021; Krems et al., 2021; Lassetter et al., 2021). Although there appears to be dramatically wide variability in rates of CR across various paid online sample venues (see Eyal et al., 2021; Litman et al., 2021), even the best-performing premium prescreening options will inevitably end up allowing some CR to occur.

How can CR inflate effect sizes?

When a careless responder invariably gives the same answer over and over to the items in a test (i.e., "straight-lining"), it is easy to intuit how such data would often naturally inflate internal reliabilities and associations between measures (at least when all items are keyed in the same direction[7] (for a discussion, see King et al., 2018). However, what is less intuitive is that even fully random data provided by careless respondents will often inflate such associations (Credé, 2010). Our goal here is to present a concise, less technical introduction to this CR statistical confound than those that have been described previously in excellent detail (e.g., Huang & DeSimone, 2021; Huang et al., 2015) in hopes of connecting with a wider readership.

For CR as a variable[8] to be a confounding factor, it must causally effect scores for both the independent variable (*X*) and the dependent variable (*Y*). In other words, careless responders must produce group means on both the *X* and *Y* variables that are different from the group means produced by careful responders. Thus, to gauge the risk of one's data producing effect sizes inflated by CR, one must answer two questions: What are the means of *X* and *Y* among careless respondents, and what are the means of *X* and *Y* among careful respondents?

For respondents who provide exclusively random responses on a psychological scale or test, their mean scores, on average, can be expected to be very close to the midpoint of a rating scale or at chance-level performance on a test. For respondents who provide only partially random data (e.g., individuals who respond carelessly to some but not all items), their mean scores

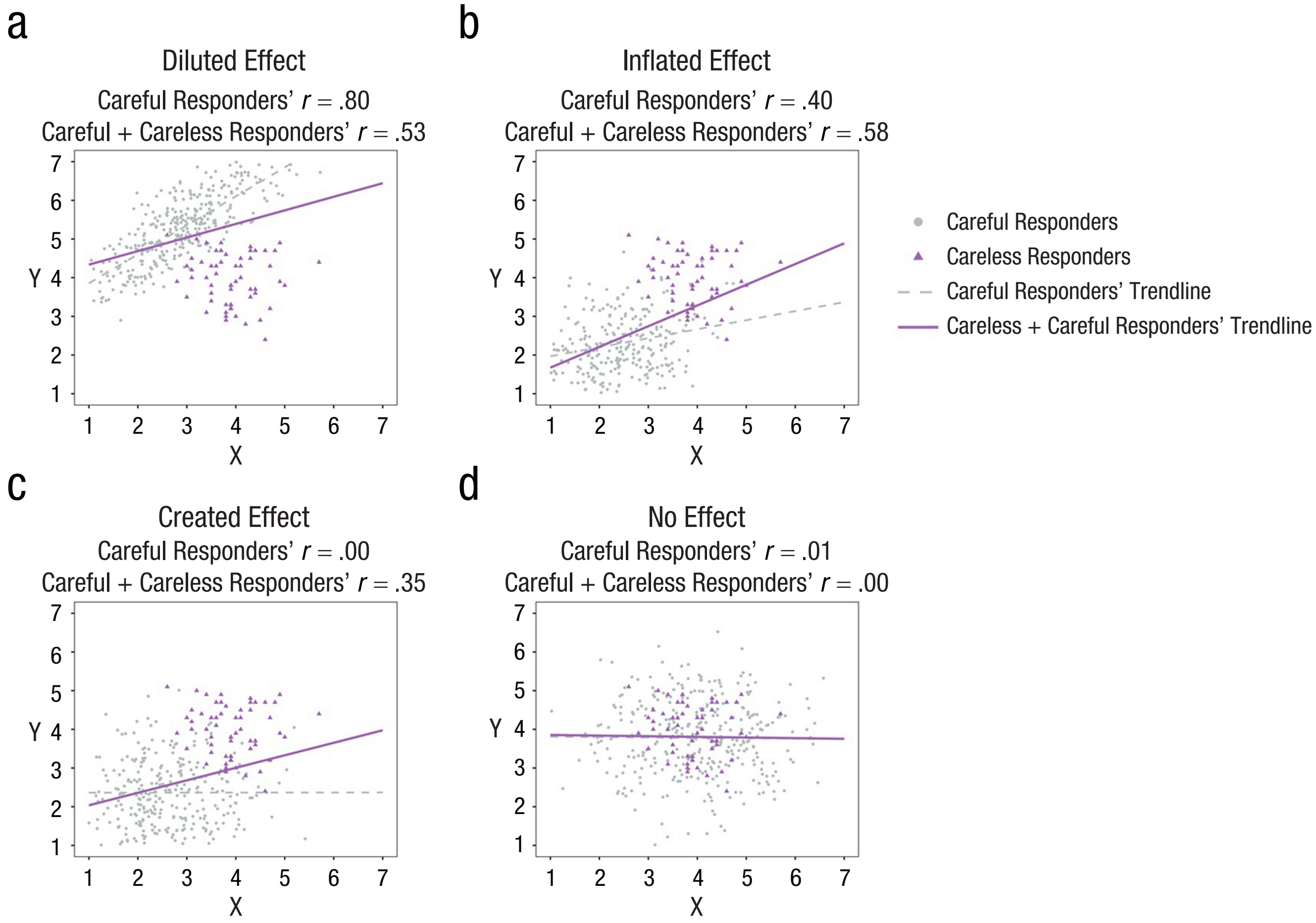
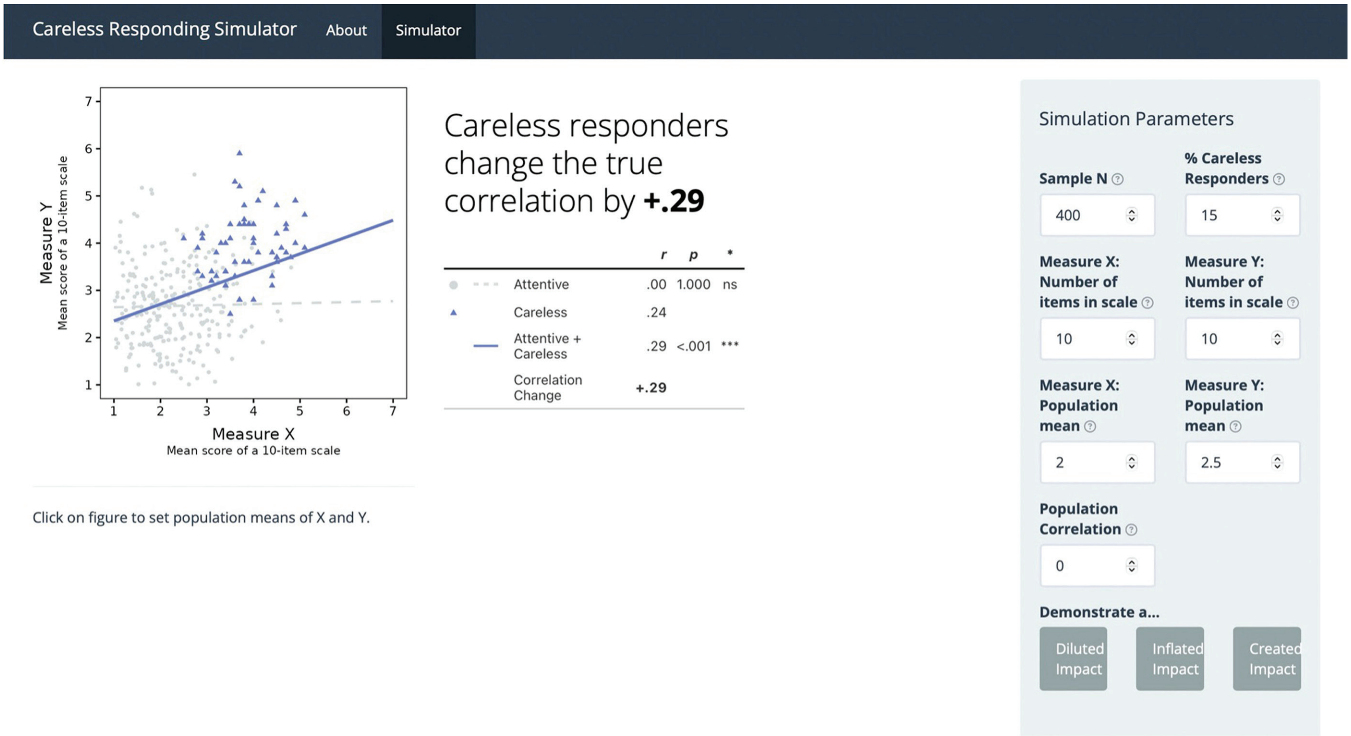


Fig. 1. Example of different effects failing to remove careless responders has on the relationship between *X* and *Y*.

should still be closer to the midpoint of a scale or chance-level performance than will fully careful respondents (Huang et al., 2015). Empirical studies that examined CR (Huang & DeSimone, 2021; Huang et al., 2015; Litman et al., 2021) have generally corroborated that this tendency naturally emerges in real studies.

Given that one can conceptualize the group means of careless responders as roughly constant across studies (i.e., near to a scale's midpoint or chance level), the effect that CR has on one's data will depend on the group means of one's own variables of interest, *X* and *Y*, among careful respondents. If the mean scores of careful participants on *X* and *Y* do not measurably differ from the mean scores of careless participants on *X* and *Y* (e.g., all scores are approximately central to each scale's midpoint), CR will typically not be correlated with either of the variables. If the mean scores of careful participants differ from careless participants for only *X* or only *Y*, then CR will typically be correlated with only one of these variables. In either of these cases, the addition of CR data can simply be understood as adding only random error variance to the association between *X* and *Y* and will thus dilute the true underlying relationship between them. In other words, when CR is not correlated with both variables of interest, failing to remove CR participants from one's data set will increase a researcher's probability of commiting a Type II error (Fig. 1a)—unless the relationship between two variables is already a true null, in which it would remain so (Fig. 1d).

However, if mean scores of carefully responding participants are measurably different than the scores of careless respondents for both *X* and *Y* (i.e., diverging from scale midpoint or chance level for both variables), then the addition of CR data will tend to generate systematic covariance between the variables, putting one at risk for committing a Type I error by either inflating a true underlying relationship in one's data (Fig. 1b) or creating an entirely spurious relationship in which there is a true null (Fig. 1c). Moreover, as the mean score of the careful participants diverges further from chance level on either variable, such as is especially the case with low/high base-rate variables (e.g., ingesting bleach to prevent COVID-19; cf. Litman et al., 2020) or less difficult skills tests (for investigation of CR in low-stakes assessments, see Rios et al., 2017), the degree to which the relationship between *X* and *Y* is inflated should


Careless Responding Simulator
About
Simulator
Measure Y
Mean score of a 10-item scale
Measure X
Mean score of a 10-item scale
1
2
3
4
5
6
7
Diluted Impact
Inflated Impact
Created Impact

Careless responders change the true correlation by **+.29**

		r	p	*
● - - -	Attentive	.00	1.000	ns
▲	Careless	.24		
—	Attentive + Careless	.29	<.001	***
	Correlation Change	+.29		

Click on figure to set population means of X and Y.

Simulation Parameters

Sample N: 400

% Careless Responders: 15

Measure X: Number of items in scale: 10

Measure Y: Number of items in scale: 10

Measure X: Population mean: 2

Measure Y: Population mean: 2.5

Population Correlation: 0

Demonstrate a...

Fig. 2. Screenshot of the Careless Responding Simulator, a Shiny app to demonstrate how the introduction of careless responding can distort true effects.

increase (for a nuanced discussion, see Credé, 2010; Huang et al., 2015).

It is typically the case that mean scores on a self-report scale diverge from the midpoint of a scale among careful responders (Credé, 2010). For example, most self-report measures of psychopathology and of clinical-symptom counts are positively skewed (see review in Field & Wilcox, 2017). In addition, most ability measures have population norms above chance performance. Thus, rather than being a rare phenomenon, CR inflation should, in many cases, be presumed to be more likely than CR attenuation.

Educational tools to demonstrate and explain CR

The past decades of improvement in psychological-science methods have shown that videos and visual tools can be very helpful to spreading awareness of important analytic issues. To this end, we developed the Careless Responding Simulator to visually and dynamically demonstrate the inflating/deflating effects of CR on bivariate associations.[9] Figure 2 shows a screenshot of the app that researchers can use to define different parameters for theoretical careful and careless responders, along with other parameters, to show how a correlation among careful responders changes with the addition of careless responders. There are also simulation presets that demonstrate a diluted effect, an inflated effect, and a created effect. It can be accessed online at https://fuhred.shinyapps.io/CarelessRespondingSimulator/.[10] Moreover, we have also created (a) a brief video introducing the CR inflation effect and (b) a brief video explaining how to use the Careless Responding Simulator (accessible at: https://www.youtube.com/watch?v=niTPWqr6fsE, https://www.youtube.com/watch?v=uUrgRFbiTks). These tools are not intended to help researchers estimate the specific effects of CR in their own data sets. Instead, we hope these educational aids will help readers better understand and promote the CR issues discussed in this article.

Case Studies of Three Recently Published Articles

In the previous section, we aimed to describe how unremoved CR participants can inflate effect sizes. The next step is to demonstrate this inflationary effect in real data. Whereas our explanations and simulations above assume CR that produces fully random responding, CR by real participants could potentially take on a wide variety of patterns (e.g., alternating between "1" and "3" on a response scale or producing some other systematic pattern that is not merely straightlining). Moreover, the patterns produced by CR may be influenced by specifics of individual study designs in unpredictable ways (e.g., some study designs may lead CR to manifest as severe central tendency bias, whereas others may tend to lead to CR that manifests as extreme responding bias). Thus, it is important to investigate the Type 1 error risks of CR using real participant data.

Although prior demonstrations of the inflationary effects of CR have primarily relied at least partly on simulated data, we do not use any simulated data here and instead focus on three data sets from studies recently published in *JPSP* in which the authors (a) carefully screened for CR and (b) subsequently excluded a substantial proportion of their sample. We conducted novel reanalyses of these data sets to assess whether these researchers' results would have been inflated had they not taken care to remove CR participants. In addition, whereas prior studies have used CR screening methods decided on by the CR article authors, we instead adopt the screening methods used by the respective study authors, which substantially differed from one another. In sum, in our reanalyses, we aimed to reduce potential bias that might result from our own perspectives on the nature of CR and how to best screen for it.

Before our systematic reviews of *JPSP* and *Psych Science* (described later in this article), we identified five potential studies to investigate; our aim was to sample from multiple types of variables (e.g., self-reported attitudes, low base-rate self-reported behavior, and effortful behavioral tasks), multiple paid online sample sources (e.g., not just MTurk), and authorship teams that used different methods from one another to screen for CR data. We were able to obtain the necessary data to conduct reanalyses for three of these studies.[11]

Study 1: self-report scales with relatively normal distributions

Kachanoff et al. (2020) provided an example of how CR might meaningfully inflate associations between self-report scales even if the mean responses of careful and careless participants deviate only modestly from one another. Kachanoff et al.'s goal was to investigate whether collective autonomy restriction, "a feeling that other groups seek to control and restrict how their own ingroup articulates and expresses its sociocultural identity" (p. 601), motivates groups to improve their power position in a social hierarchy.

In Kachanoff et al.'s (2020) Study 1, they administered a series of six self-report scales to a sample of 412 Black American participants recruited online via the research panel firm Dynata. They used three nearly identical attention-check items in Study 1 to screen for CR (e.g., "If you're paying attention, please select 3"). They excluded 101 participants (24.5% of the sample) who failed at least one of these items.

First, we examined whether there were mean differences on the six self-report variables between participants Kachanoff et al. (2020) retained in their study analyses (i.e., careful participants) and participants they excluded for failing an attention check (i.e., careless participants; Table 1). There were medium to large significant differences ($p < .001$) between the careful and careless groups for four of the six main variables assessed by Kachanoff et al. Thus, we expected that failing to exclude CR participants would inflate the correlations observed between these four variables.

To test this, we ran the correlations between the study variables for the whole sample (including careless participants) and compared them with the published correlations (with careless participants removed) in Kachanoff et al. (2020). As shown in Table 1, for correlated variables that both displayed mean differences between careless and careful groups, every effect originally reported by Kachanoff et al. was inflated to varying degrees when data from careless participants were included (median effect size r increased by .07, or proportion of explained variance by $R^2 = .05$).

The CR inflation observed in our reanalysis of the data provided by Kachanoff et al. (2020) was modest but meaningful. Even at this level of inflation, including CR data can cause Type 1 errors; distort moderation, mediation, and other analyses; and inflate effect-size estimates when included in meta-analyses.

Study 2: low base-rate self-reported behaviors

Compared with the distortion demonstrated in the Kachanoff et al. (2020) data set, the CR inflationary dynamic will be more pernicious with variables whose true values are more extremely tilted away from the midpoint of a scale. This is especially the case for low and high base-rate variables (e.g., see demonstrations in Chandler et al., 2020). Low/high base-rate phenomena could range from mental-health diagnoses (e.g., schizophrenia) or other characteristics (e.g., identifying as cisgender) to certain behaviors (e.g., physically confronting Black Lives Matters protesters; Bartusevičius et al., 2021) or other specific outcomes.

Costello et al.'s (2022) work on the correlates of left-wing authoritarianism (LWA), a construct that describes authoritarianism in service of left-wing outcomes, provides a valuable window into how CR can be particularly inflationary when working with low base-rate phenomenon. We focused our reanalysis efforts on Costello et al.'s Study 6, in which the authors investigated whether LWA predicted self-reported participation in protest violence during summer 2020 and other acts of self-reported political violence in the last 5 years

Table 1. Comparing Correlations Without Careless Responding (n = 311) and With Careless Responding (n = 412): Kachanoff et al. (2020) Study 1

	Careful-responder group *M* (*SD*)	Careless-responder group *M* (*SD*)	Cohen's *d*	1 Without CR/with CR	2 Without CR/with CR	3 Without CR/with CR	4 Without CR/ with CR	5 Without CR/with CR	6 Without CR/with CR
1. Collective action support	5.66 (1.06)	4.52 (1.06)	1.08*	—	–.24*/**–.35***	.45*/.54*	.41*/.48*	.16/.06	.28*/.22*
2. System justification	3.26 (1.11)	3.94 (0.83)	0.69*	ΔR^2 = .06	—	–.26*/**–.31***	–.24*/**–.26***	–.09/–.01	.10/.13
3. Illegitimacy of group position	5.63 (1.36)	4.77 (1.32)	0.64*	ΔR^2 = .09	ΔR^2 = .03	—	.31*/.37*	.13/.07	.09/.10
4. Collective autonomy restriction	5.23 (1.30)	4.47 (1.17)	0.61*	ΔR^2 = .06	ΔR^2 = .01	ΔR^2 = .04	—	.29*/.23*	.23*/.24*
5. Support of Black Power Movement	42.52 (34.58)	54.35 (32.15)	0.35	ΔR^2 = –.02	ΔR^2 = –.01	ΔR^2 = –.01	ΔR^2 = –.03	—	.16/.15
6. Group identification	5.51 (1.25)	5.37 (1.40)	0.11	ΔR^2 = –.03	ΔR^2 = .01	ΔR^2 = .00	ΔR^2 = .00	ΔR^2 = .00	—

Note: Correlations with CR participants included in analyses are in bold. ΔR^2 = change after adding CR data; CR = careless responding.
*p < .001.

Table 2. Comparing Left-Wing Authoritarianism Predicting Binary Variables in Binary Logistic Regression Without Careless Responding (n = 805) and With Careless Responding (n = 934), Controlling for Symbolic Ideology: Costello et al. (2021), Study 6

	Careful-responder group *M* (*SD*)	Careless-responder group *M* (*SD*)	Cohen's *d*	LWA Exp(B) (without careless responders)	LWA Exp(B) (with careless responders)	ΔR^2 adding CR
LWA[a]	–.07 (.96)	.38 (1.03)	0.45*	—	—	—
2020 protest-violence behavior	.02 (.15)	.28 (.45)	0.78*	2.96*	6.02*	.29
2020 protest-violence support	.08 (.27)	.29 (.45)	0.57*	5.52*	7.77*	.04
5-year political-violence behavior	.01 (.08)	.27 (.45)	0.80*	5.89*	11.92*	.39
5-year political-violence desire	.09 (.28)	.27 (.44)	0.49*	2.58*	4.13*	.07

Note: Exp(B)s with careless participants included in analyses are in bold. LWA = Left-Wing Authoritarianism Scale; CR = careless responding; ΔR^2 = change after adding careless-participant data.
[a]Converted to z score.
*p < .001.

because these behaviors have low base rates among the general population.

Costello et al. (2022), in Study 6, recruited 1,000 participants from Prolific, a paid online data-sourcing platform. In addition to two basic attention-check items (e.g., "Balls are round," agree/disagree), participants were also screened for CR by being instructed to "write a sentence that you think has probably never been said before (e.g., 'the red, disingenuous marmoset galloped over the Atlantic Ocean while wearing sunglasses')" (Costello, personal communication). If a participant left the answer blank, wrote gibberish, failed to produce an actual sentence, or wrote a sentence that was not at all unusual, that participant was excluded.[12] On the basis of these three CR screening items, Costello et al. identified 16.9% of the sample as careless and excluded them from analyses, leaving a final sample of 834 participants.

We first identified whether (a) the LWA scale and (b) the low base-rate variables had significantly different means for the careless- and careful-participant groups. Scores on the LWA scale and each of the four binary political-violence variables all demonstrated medium to large mean differences between the careful and careless participants (p < .001). For example, whereas only 2.3% of the careful participants reported engaging in violence during summer 2020 protests, 28.2% of the excluded careless participants reported engaging in protest violence in summer 2020. Thus, we expected LWA's relationship with all four of these low base-rate variables to be inflated with the addition of the careless-participant data.

We examined the associations between LWA and the binary political-violence variables with and without careless participants by conducting a series of binary logistic regressions controlling for symbolic ideology (the same regression procedure reported by Costello et al., 2022). As shown in Table 2, adding the careless participants inflated associations between every single one of these variables yet was particularly dramatic for the two lowest base-rate behavior variables. For example, without careless participants, the odds of having taken part in violence during protests in summer 2020 increased by a factor of 2.96 for each standard-deviation increase in LWA. When careless participants were included, the odds increased by a factor of 6.02 for each standard-deviation increase in LWA. In sum, the careful CR screening conducted by Costello et al. (2022) was critical to the quality of their findings. This case study raises serious concerns about similar online studies working with low/high base-rate variables that do not screen for CR.

Study 3: behavioral tasks

Although the case study of Costello et al. (2020) demonstrates the notably dangerous effects of CR with highly skewed or low base-rate self-report data, CR has the potential to undermine almost anything that a study participant could be asked to do. For instance, virtually all behavioral tasks require at least some degree of effort and/or attention, whether one is labeling the emotional content of faces, completing logic puzzles, answering trivia questions, or reading prompts for experimental manipulations (for similar concerns, see Huang & DeSimone, 2021).

The behavioral task studies of Sanchez and Dunning (2021) illustrated CR's potential to inflate associations between different behavioral tasks and between behavioral tasks and self-report questionnaires. These researchers reported a range of correlates of the behavioral construct of "jumping to conclusions" (JTC), defined as "collecting only a few pieces of evidence before reaching a decision" (p. 790).

We focused our reanalysis on Study 1B of Sanchez and Dunning (2021) because this was the only study in their article in which they screened for CR. A total of

Table 3. Comparing Associations Without Careless Responding (N = 298) and With Careless Responding (N = 346): Sanchez and Dunning (2021), Study 1B

		Careful-responder group M (SD)	Careless-Responder Group M (SD)	Cohen's d	JTC rs (without careless responders)	JTC rs (with careless responders)	ΔR^2 adding careless responders
	JTC[a]	3.55 (1.97)	1.88 (1.57)	0.94*	—	—	—
Self-report	Paranormal beliefs	2.13 (0.87)	3.39 (1.04)	1.31*	.30*	.45*	.11
	Aggregate oddball beliefs	2.09 (0.84)	3.18 (1.11)	1.11*	.32*	.47*	.12
	Conspiracy beliefs	2.06 (0.91)	3.08 (1.23)	0.94*	.27*	.44*	.12
	Medical myth beliefs	2.07 (0.99)	3.07 (1.21)	0.90*	.29*	.44*	.11
	Schizotypy	0.36 (0.24)	0.51 (0.23)	0.64*	.08	.19*	.03
	Need for cognition	90.01 (32.15)	80.02 (21.78)	0.37	–.14*	–.19*	.02
	Need for closure	175.46 (25.45)	176.04 (17.41)	0.03	–.05	–.02	.00
Reasoning performance	Cognitive ability	39.96 (7.65)	24.65 (14.72)	1.31*	–.43*	–.57*	.14
	Denominator neglect	2.67 (1.73)	4.16 (1.98)	0.80*	.30*	.41*	.08
	Critical reasoning task	3.17 (1.70)	1.73 (1.93)	0.79*	–.43*	–.52*	.09
	Belief bias	3.39 (2.72)	4.23 (1.75)	0.37	.29*	.32*	.02
Civics test performance	Accuracy	15.52 (3.04)	10.13 (5.50)	1.05*	–.36*	–.51*	.13
	Overconfidence	8.00 (14.65)	31.32 (28.81)	1.02*	.33*	.48*	.12
	Discrimination	13.40 (13.03)	6.21 (10.11)	0.62*	–.16*	–.21*	.02
	Confidence	85.76 (11.12)	82.09 (13.35)	0.30	–.05	–.09	.01

Note: Correlations with careless participants included in analyses are in bold. JTC = jumping to conclusions; ΔR^2 = change after adding careless-participant data; CR = careless responding.
[a]Lower number = higher JTC.
*p < .001.

346 participants were recruited from MTurk. To measure JTC, participants watched a character fishing from a lake containing either (a) 80% red fish and 20% gray fish or (b) 80% gray fish and 20% red fish. One fish was caught at a time, and after each fish catch, participants could ask to watch another fish be caught. When a participant felt ready, they could stop watching fish catches and make their guess as to whether the character was fishing out of the majority-red lake or majority-gray lake. Each participant's JTC score was a function of how many fish catches they asked to see before making their decision.

CR is an obvious possible confound for JTC because failing to ask to see more fish catches could be a result of lack of effort or a desire to complete the study as quickly as possible (for particularly germane demonstrations of this problem, see Zorowitz et al., 2021). Thus, to evaluate whether their results might be driven by lack of effort/attention by participants, Sanchez and Dunning (2021) employed a pair of attention checks that were unobtrusive (i.e., participants were not likely to realize that they were being attention checked because the items seamlessly blended with the rest of the survey).[13] Sanchez and Dunning excluded 13.9% of their sample for failing at least one of these attention-check items, leaving a final sample of 298 participants.

Would Sanchez and Dunning's (2021) Study 1B results have been meaningfully different if they had failed to exclude careless participants? To address this, we first examined which variables had significantly different means for the participants they retained (careful participants) and participants they excluded (careless participants). For the JTC behavioral task, there was a major difference between careful and careless participants; the careless participants on average asked to see only 1.88 fish being pulled from the lake before deciding, whereas the careful participants on average asked to see 3.56 fish being pulled from the lake before deciding. Of the remaining 15 variables described in their Table 1, 11 demonstrated mean differences between the careful and careless participants (p < .001). Thus, for the 11 variables that demonstrated significant mean differences between the careful and careless groups, we expected the strength of their correlation with JTC to be inflated when careless responders were included in analyses.

We ran correlations between all the variables including careless participants and compared these correlations with those reported in Table 1 of Sanchez and Dunning (2021)—in which careless participants had been excluded. As shown in our Table 3, JTC's associations with the 11 variables we had identified were all inflated when careless participants were included in the analyses. For JTC's associations with the self-report variables from this set, on average, the presence of careless participants increased the proportion of explained variance by $R^2 = .12$ (median increase in $r = .15$). For associations with the behavioral-performance measures (i.e., reasoning and civics performance) from this set, including careless participants increased the proportion of explained variance by $R^2 = .09$, on average (median increase in $r = .09$). For the four variables that did not significantly differ in their means between the careful- and careless-participant groups, we still noticed a slight increase in effect size (median increase in $r = .04$) when careless participants were included in analyses, perhaps because these variables differ just enough (even if not significantly) in their means between the careful and careless groups.[14]

Recent Data-Screening Practices in Paid Online Samples in Two Flagship Journals

As the three case examples presented above illustrate, failing to screen for and exclude CR participants will often spuriously create or inflate findings. Indeed, 28 of the 34 effect sizes (82%) that we examined became stronger, not weaker, when careless participants were included in the analyses. This begs the following question: How frequently and rigorously are researchers screening for and removing CR participants from their data sets, especially in samples collected from paid online platforms? Are potentially vast numbers of findings published in contemporary psychological science spuriously created or inflated because of lack of CR screening procedures?

Researchers in various fields (e.g., marketing research; Arndt et al., 2022) have investigated the prevalence of CR screening by systematically reviewing specific journals or published articles on a specific topic. Within psychology, such reviews have thus far (to our knowledge) been focused only on clinical psychology. For example, King et al. (2018) reviewed every study published in 2016 in 14 journals focused on addictions research; they observed that only 11 out of more than 2,079 studies (< 0.01%) reported any kind of screening for CR when collecting data online. Sharpe et al. (2023) conducted a similar review of recent studies in three other clinical-psychology journals; they observed that 14 out of 20 (70%) studies engaged in data-quality screening when collecting data from MTurk or other online recruitment panels. Jones et al. (2022) systematically reviewed alcohol-research studies employing online recruitment panels published from 2011 to 2021; 51 out of 96 (53.13%) identified articles reported screening for CR. A main limitation of all three of these clinical-psychology reviews is that none of them directly contacted authors to inquire about their screening practices; given that most journals do not currently have requirements for reporting CR data exclusions, some authors may have screened for CR but not explicitly reported it in the main text of their articles. Thus, a more valid estimate requires directly contacting study authors when the presence or absence of CR screening cannot be determined from the article text.

To improve on these prior reviews and extend them to the broader field of psychological science, we conducted a systematic review of CR-screening practices in two flagship journals widely read by psychologists outside of clinical research. Specifically, we examined articles published in *JPSP* and *Psych Science* over a 1-year period from May 2020 through April 2021. The review of *Psych Science*, which came after the review of the *JPSP*, was recommended by journal editors and was preregistered (https://osf.io/u5jgr) to replicate the methods used in the earlier *JPSP* review.

We focused on studies published in these two journals that used paid online samples. In addition to their heightened risk of CR, one of the main advantages of paid online samples is that they allow researchers to attain larger sample sizes rapidly and cheaply. As sample size increases, however, so too does the false-positive rate generated by CR inflation; in large samples, even a small CR-caused distortion can spuriously make a null effect appear to be statistically significant (Zorowitz et al., 2021; also cf. "paradox of big data"; Meng, 2018).

For full details regarding these two systematic reviews, see the Supplemental Material available online. In sum, we reviewed every article in these two journals during this 1-year period and identified studies using paid online samples. We then coded (a) whether CR screening was conducted for the study and if so, (b) whether multiple methods for detecting CR were used and (c) how many CR respondents were subsequently excluded. Every article was examined by at least two coders, and consensus on final codes was reached through group discussion if there were discrepancies between coders. In many cases, some of these codes could not be conclusively established based solely on the information in the published articles and accompanying supplemental material; in such cases, we directly contacted author teams to request the missing information.

When coding for whether a study engaged in CR screening, we erred on the side of overcrediting researchers: A study qualified as having screened for CR if any method that might be considered a form of CR screening was employed[15] even if the researcher may not have been intending to screen for CR (e.g., excluding participants who did not adequately follow instructions). Thus, our findings should be considered a conservative estimate of the problem of CR-screening neglect.

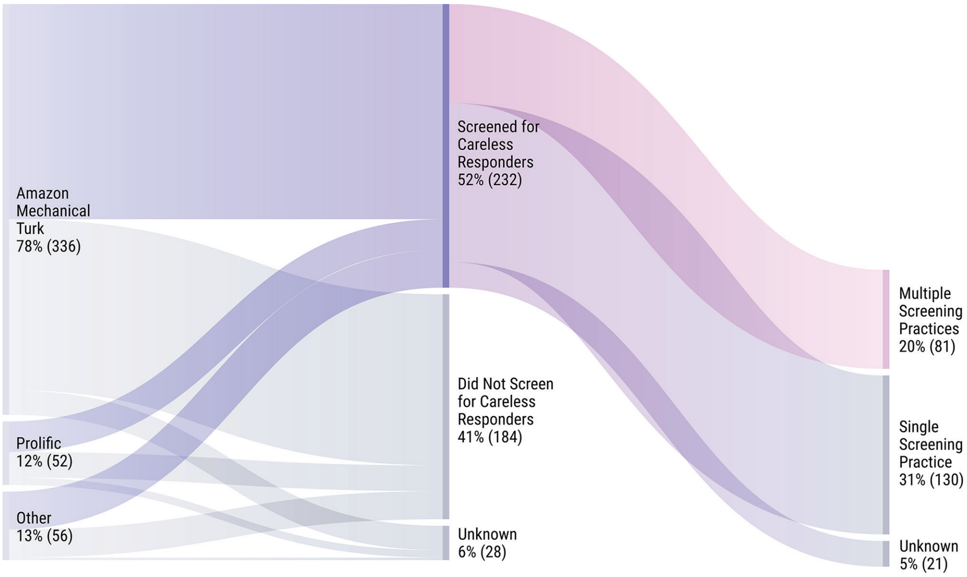


Fig. 3. Alluvial plot of screening practices in *Journal of Personality and Social Psychology* [*JPSP*] and *Psychological Science* [*Psych Science*] over 1 year, for studies using paid online samples.

An alluvial plot summary of our review findings is provided in Figure 3. Of the 273 articles spanning 613 studies published in these two journals in our specified time frame, 104 articles spanning 444 studies analyzed data from paid online samples (48.2% of all articles published in *JPSP* and 27.9% of all articles published in *Psych Science* across 12 issues). Across the 444 studies using paid online samples, 336 came from MTurk, 52 came from Prolific, and the remaining came from 56 other specified sources.

Of these 444 studies, only 217 (48.9% total; 43.0% in *JPSP* and 62.4% in *Psych Science*) explicitly reported in text (or in online supplemental material) that the authors had conducted some kind of CR screening. Including the 49 author teams, representing 164 studies, who responded to our direct inquiries regarding whether they had screened for CR data (10 authorship teams, representing 31 studies, did not respond), 232 (52.3%) studies screened for CR in total (*JPSP*: total = 149; *Psych Science*: total = 83).

Of the 232 studies that conducted CR screening, regardless of whether it was reported in text or to us in response to our email, 81 (34.9% total; 39.5% in *JPSP*, 26.5% in *Psych Science*) indicated that they employed a combination of different screening methods. For example, it was somewhat frequent for studies to use (a) one or more basic attention-check questions (e.g., "If you are paying attention, please select number 9") alongside (b) screening for unreasonably short study completion times. It remains possible, however, that some studies used more than one type of screening, yet the authors did not explicitly say so.

Of the 232 studies that conducted at least some kind of CR screening, 207 studies (89.2%) reported the number of participants that were excluded. Only about one sixth (15.1% in total; 14.1% from *JPSP* and 16.9% from *Psych Science*) of the studies that screened for CR excluded 15% or more of their participants. Instead, authors generally excluded only minimal numbers of participants based on CR screening (*JPSP: Mdn* = 7%, IQR = 9%; *Psych Science: Mdn* = 6%, IQR = 10%). Studies that reported using multiple types of CR-screening methods tended to exclude larger proportions of their samples (with multitype screening, *JPSP: Mdn* = 9.5%, *IQR* = 8.0%; *Psych Science: Mdn* = 7.5%, IQR = 9.75%; without multitype screening, *JPSP: Mdn* = 5.0%, IQR = 8.5%; *Psych Science: Mdn* = 5.5%, IQR = 9.75%).

In sum, only about half of online studies published in *JPSP* during this 1-year period appear to have screened for CR, and even among those, many likely did so in only weakly effective ways. Although studies published in *Psych Science* were slightly more likely to conduct some kind of CR screening, they were slightly less likely than studies published in *JPSP* to screen in rigorous ways (i.e., using more than one screening method). In sum, many of the studies in these two journals may be at risk of having reported spurious or inflated results.

General Discussion and Recommendations

Our overarching aim in this article was to explain and demonstrate the severity and frequency of Type 1 error risks that result from the "insidious confound" (Huang et al., 2015) created by CR research participants. Counter to long-standing conventional wisdom, the presence of partly or fully random responding will very often spuriously inflate associations between variables (e.g., Credé, 2010). Because of the counterintuitiveness of this phenomenon and perhaps also the technical nature and specialized journal identities of past articles on this topic, this serious confound has continued to be widely underappreciated.

In the three case examples we presented, the original authors took careful steps to identify and exclude substantial amounts of CR data. Our reanalyses of the data in these studies revealed just how right the authors were to conduct CR screening; if they had not done so, the majority of the effects they would have reported would have been spuriously created or inflated, some modestly and some dramatically. Beyond demonstrating that CR will often inflate effect sizes more frequently than diluting them, the studies we reanalyzed further illustrate how CR can spuriously inflate effect sizes in almost any type of research that requires effortful cognitive processing, from self-reporting attitudes to engaging in behavioral tasks. Yet our reviews of screening practices in two prominent journals show that researchers are commonly failing to rigorously screen for and remove CR participants from their data sets. Moreover, our survey study of journal editorial boards further confirmed that many of them do not adequately appreciate the Type 1 error risks posed by CR data, which likely explains why many of them do not place much weight on CR screening in reviews of articles. Thus, there are likely a very large number of false-positive and spuriously inflated results continuing to be published, especially in an era of unproctored online studies with anonymous paid participants.

Addressing CR in research

A number of excellent reviews are available to help researchers think through the various strategies to identify, remove, and report CR in their articles (e.g., Arthur et al., 2021; Curran, 2016; Goldammer et al., 2020; Hong et al., 2020).[16] In particular, Ward and Meade (2023) recently presented a helpful *Annual Review* of CR data and screening; we strongly encourage readers to consult it. Here, we synthesize a few of the major recommendations from the literature and note a few places in which we diverge from points made by Ward and Meade.

Identifying CR. Scholars have used a wide range of techniques for identifying CR (e.g., Curran, 2016; Hong et al., 2020). Examples of these screening techniques are instructed items (e.g., "Please select STRONGLY AGREE"), bogus items ("I have never eaten food"), consistency between psychometric antonyms (e.g., "I love my life" and "I do not love my life"), speed of page completion (considered superior to speed of total survey completion), recall checks (e.g., "What was the name of the person in the vignette you just read?"), outlier detection methods, and many other techniques. A full range of different methods for screening for CR were helpfully explained by Ward and Meade (2023; for a review of screening methods in the context of bot detection specifically, see Storozuk et al. 2020). The R package *careless* (Yentes & Wilhelm, 2021) can implement a variety of these methods.

These different CR-screening methods appear to have both strengths and weaknesses such that "there does not seem to exist a universally effective . . . detection method" (Hong et al., 2020, p. 313). Indeed, given that CR takes different forms, such as random responding or lazily overly consistent responding (i.e., straightlining), no single screening method will be excellent at detecting all the different manifestations of CR. Moreover, the merits of any approach also depend on the characteristics of a particular research study, such as the other tasks and measures within it. For example, a screening method that would be unobtrusive for one kind of study might be awkward and intrusive for another kind of study in such a way that it causes participant reactance or willful noncompliance (Silber et al., 2022). Thus, rigorous CR screening requires researchers to develop a diverse tool kit of varying CR-identification techniques that can be applied appropriately to one's specific study designs.

Although we refrain from recommending a catchall CR-identification method to researchers, our central recommendation echoes that presented in other CR reviews (e.g., Chmielewski & Kucker, 2020; Hong et al., 2020): "The use of any of these techniques should not be applied in a vacuum void of other techniques. . . . The strongest use of these methods is to use them in concert" (Curran, 2016, p. 6). Ward and Meade (2023) provided various suggested combinations of different screening methods, some of which must be built into a study (i.e., a priori screening) and some of which can be used with archival data (i.e., post hoc screening). Optimally, both a priori and post hoc methods should be used in tandem. In their first table, Ward and Meade arranged their a priori and post hoc CR-identification suggestions into three levels of screening rigor: minimal, moderate, and extensive. Note that only a small percentage of the online studies we reviewed in two flagship journals meet even the minimal screening rigor elaborated by Ward and Meade.

Excluding CR. Because carelessness is inherently a continuous, rather than categorical, phenomenon, researchers must next determine a proper threshold for excluding respondents from analyses. For instance, should even a single failed attention check be sufficient for exclusion, or should a series of failed attention checks be necessary? As the stringency of CR screening increases, some meaningfully careful participants will tend to be inadvertently excluded (Kim et al., 2018); as stringency of CR screening decreases, data quality will tend to suffer.[17] Below, we first address a few prominent misconceptions in CR exclusion considerations and highlight what we believe to be valid considerations for being more or less stringent in CR exclusion criteria.

The first misconception is the belief that CR data are less risky and should be less stringently excluded "if sample sizes are smaller and the analyses are somewhat more robust (e.g., correlations)" (Ward & Meade, 2023, p. 588). The results of our various case studies, however, show how even correlational analyses are at risk for spuriously inflated results if CR is left unremoved. As detailed by King et al. (2018), this risk is present for essentially any analysis that deals with covariances between items/variables. In addition, whereas larger sample sizes do amplify the potential for CR data to generate false-positive findings under tests of statistical significance (Zorowitz et al., 2021), small samples are also at risk of inflated effect sizes because of CR (for a similar discussion, see Sharpe et al., 2023). We therefore advise that sample size and analytic method should not be used as justification for lessened screening stringency.

Many researchers also initially dismiss the need to stringently exclude CR participants in their samples by pointing to strong internal consistencies for their measures. This is a natural intuition, one likely even held by some psychometricians; for instance, in Ward and Meade's (2023) review, CR is depicted as only tending to reduce the internal reliability of measures (also see Arthur et al., 2021, p. 115), leaving readers to potentially infer that high reliabilities in their measures signal less of a need to stringently exclude CR participants. Yet the presence of even substantial amounts of CR will often not decrease the internal consistency of measures; in fact, CR data will sometimes inflate internal consistency, especially when all test items are keyed in the same direction (for a demonstration, see online simulator app in Carden et al., 2019).[18] As with any other statistical estimate based on covariance matrices, CR can add systematic covariance between items within a scale. "High Cronbach's alpha is no indicator of data quality" (Hong & Cheng, 2019, p. 622) and thus should not be used as a rationale for less stringent CR exclusion criteria.

What we believe researchers should consider when determining a proper threshold for excluding respondents from analyses are research sample characteristics and study design. Although CR can occur in practically any research context (e.g., undergraduate study pool samples, surveys voluntarily completed by journal editorial-board members), Ward and Meade (2023) noted that CR is likely to be especially prevalent in studies administered online; studies that are long,[19] repetitive, or uninteresting to participants; and studies in which participants face little or no consequences for responding carelessly. With Ward and Meade, we advise that researchers should be moderate to extensive in their CR exclusion efforts when studies have any of these characteristics and rely only on more minimal exclusion criteria when one's study design is unlikely to invite many careless responders (e.g., proctored in-person studies with intrinsically motivated participants).

Above all, we make an important recommendation in accordance with the findings of the present article: Researchers should assess the expected mean scores of the variables in their study and adjust their screening stringency accordingly. If any of the variables of interest have expected mean scores that differ, even modestly, from the midpoint of the variable range (or chance level on a test), then CR exclusion scrutiny and stringency should be heightened. When working with highly skewed or low/high base-rate self-report variables or with behavioral tasks that are inherently sensitive to participant effort, CR screening and exclusion should be extremely rigorous to avoid spuriously inflated effects.

Finally, whatever level of screening stringency a researcher may adopt, best practices in open science recommend running and transparently presenting analyses both with and without identified CR participants (e.g., a "multiverse approach"; Del Giudice & Gangestad, 2021; Steegen et al., 2016). In fact, such analyses can provide additional information to evaluate the necessary stringency: If the results substantially differ, that should raise additional concerns about the underlying data and prompt consideration of further increasing the stringency.

Registering and reporting CR. We encourage researchers to establish both their identification and exclusion criteria before conducting their research and to preregister these decisions. Then, just as researchers should be transparent about other study qualities that may affect the validity of their findings in articles (e.g., whether hypotheses were determined a priori, whether their measures were reliable), the identification and exclusion of CR should be described in the main text of journal articles. Although extensive specifics may be reasonably reserved for supplemental documents, we suggest that at least the following topics be addressed in the main text: screening methods employed, whether the exclusion criteria were determined before data collection, how much data were subsequently excluded, and whether any deviations from one's preregistration were made if a screening plan was preregistered (see similar but slightly more intensive recommendations by Chmielewski & Kucker, 2020). Not only does this allow journal editors and readers to better evaluate the credibility of one's reported effect sizes, but it also assists others with testing the reproducibility of findings.

Conclusion

We aimed to bring the issue of CR's inflationary effects on observed associations to the forefront of the minds of researchers in an easily accessible manner. Given that our systematic reviews revealed many or most high-profile research studies with paid online samples failed to adequately screen for CR, it is prudent to presume that many published findings are spuriously inflated. It is our hope that, especially in the current era of reliance on paid online samples, researchers will view CR screening as critical to safeguarding the credibility of their psychological research.

Appendix A

Survey of psychology journal editors' perspectives on careless responding

This preregistered survey was conducted to gauge the level of awareness of the Type 1 error risks posed by careless responding. Rather than survey the general psychological-researcher population, we focused our study on highly experienced researchers who serve a gatekeeping role in maintaining the validity of published research findings: editorial board members of prominent psychological-science journals. Although this approach likely overestimates the extent to which the field understands the effects of careless responding on research data (because these editorial-board members are among the most highly experienced and trusted members of the research community), it allowed us the advantage of illuminating how careless responding is perceived by individuals handling the journal review process.

Method

Participants

Editors from the following journals were contacted individually by email and were asked to take part in a 3- to 5-min survey examining perceptions of screening and excluding low-quality data from research participants: *Psychological Science; Journal of Personality and Social Psychology; Emotion; Journal of Experimental Psychology: General; Journal of Psychopathology and Clinical Science; Personality Disorders: Theory, Research, and Treatment; Clinical Psychological Science; Social Psychological and Personality Science; Personality and Social Psychology Bulletin*; and *Journal of Personality*. These 10 journals were selected because in addition to being highly credible, they often publish studies using paid online samples. Of the 1,047 editors contacted, 227 editors responded to the survey. Of these respondents, three (1.3%) described their domain of psychological science as biological, 39 (17.2%) described their domain of psychological science as clinical, 23 (10.1%) described their domain of psychological science as cognitive, 11 (4.8%) described their domain of psychological science as developmental, two (0.9%) described their domain of psychological science as educational, two (0.9%) described their domain of psychological science as industrial-organizational, 27 (11.9%) described their domain of psychological science as personality, seven (3.1%) described their domain of psychological science as quantitative, 96 (42.3%) described their domain of psychological science as social, 12 (5.3%) described their domain of psychological science as "something else," and five (2.2%) did not respond. The median lifetime number of manuscripts for which the editors had served as a journal reviewer or action editor was 150 (interquartile range = 50–350). There was no compensation associated with this survey and all materials were approved by University of North Carolina at Chapel Hill Institutional Review Board.

Procedure and materials. Editors were provided with the following definition of careless responding:

> "Careless responding" is used to refer to a participant response style that is not attentive to the item content of a survey or test. Sometimes this kind of responding is produced by people who are distracted, confused, or simply rushing through a study so quickly that they are not adequately reading the items. In online contexts, this can also occur when software "bots" are completing surveys. Although careless responding is possible in any kind of study, it is especially likely to occur in anonymous online samples, particularly when the online study is long, repetitive, boring, or fatiguing.

Directly following this statement, editors were asked a series of self-report questions regarding their understanding of the effects of careless responding on data and their experiences with careless responding as editors during the review process.

All data and research materials are available at https://osf.io/k5m7d/?view_only=abb2f8a12200487582d0a1d0dd93393b. The preregistration is available at https://osf.io/9up8j. The survey materials deviate from the preregistration to a small degree: We slightly changed the wording of the items before running the study but accidentally failed to update the language in the preregistration before

Appendix Table A1. Coded Free Responses to the Question "What Impact Might Careless Responders Have on Researchers' Analyses if Not Removed From a Dataset (if Any)?"

Response code	Frequency (%)	Example response
Mention Type I error only	0 (0.0%)	n/a
Mention Type II error only[a]	45 (21.3%)	"They can increase false negatives by introducing error."
Mention both errors	30 (14.2%)	"An analysis will have more noise in the data which might lead to false negative or false positives."
Mention Type I error but make incorrect assumptions regarding when Type I errors will occur	25 (11.8%)	"If careless responders respond in a random fashion (or, at least, a fashion not responsive to content presented), their data may weaken the presence of true patterns in the data. If they respond in a systematic fashion (always choosing a first response presented for example), I suppose that their presence could potentially introduce patterns that aren't real (among non-careless responders)."
Mentioned neither type of error	111 (52.6%)	"It can add some noise to the data."

Note: Sixteen participants did not respond.
[a]Responses were considered as mentioning Type II error if any of the following effects were mentioned: decreasing statistical power, decreasing reliability, and increasing standard error or standard deviation.

its posting. In addition, multiple of the items in the survey were "filler" questions that we did not intend to analyze; we analyzed only the items mentioned in the preregistration.

The quantitative findings are presented on page 2 of the manuscript. In addition, we also preregistered our intention to code responses to an open-ended question (see Appendix Table A1) for whether respondents indicated that careless responding would cause (a) Type 1 errors, (b) Type 2 errors, (c) both Type 1 and Type 2 errors, or (d) did not clearly indicate either type of error. In the process of coding these free responses, we concluded that we should have also included an additional category, (e) when respondents indicate the potential for Type 1 errors but either think this applies only to straightlining or other systematic patterns of careless responding, think this applies only to outliers, or think this is a concern only in small sample sizes. We are not able to know how many responses coded as Category c might actually be due to the kinds of misunderstandings captured in Category e. In addition, many of the free responses in Category d probably could reflect Category b, but we did not count them as Category b if the respondent used only vague terms, such as "noise" and "bias" and "unreliable data." In sum, these free-response codes should be understood as only tentatively suggestive evidence, not dispositive evidence. M. D. Stosic and B. A. Murphy independently coded all the free responses and then discussed any discrepancies between them to arrive at final codes. For the results of this coding procedure, see Appendix Table A1.

Transparency

Action Editor: Pamela Davis-Kean
Editor: David A. Sbarra
Author Contributions

M. D. Stosic and B. A. Murphy contributed equally to this article and share first authorship.

Morgan D. Stosic: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Brett A. Murphy: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Fred Duong: Conceptualization; Investigation; Methodology; Software; Visualization; Writing – review & editing.

Amber A. Fultz: Conceptualization; Writing – review & editing.

Summer E. Harvey: Conceptualization; Data curation; Writing – review & editing.

Frank Bernieri: Conceptualization; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Morgan D. Stosic https://orcid.org/0000-0001-7693-5521
Brett A. Murphy https://orcid.org/0000-0002-2619-9199
Fred Duong https://orcid.org/0000-0003-4825-2744

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459241231581

Notes

[1] The data generated for the "careful" respondents in this fictitious scenario come from 200 participants' real responses to measures in an existing data set. The data for the 30 "careless" respondents were created using a random number generator. Thus, although the labels of these scales were changed for demonstration purposes, the data presented reflect the impact that introducing a small sample of random responses can have on the relationship between two truly unrelated constructs. For the example data set, see https://osf.io/k5m7d/.

[2] For a representative example of this conventional wisdom in action, see Episode 99 (December 7, 2022) of the popular psychological-science podcast *Two Psychologists Four Beers* (Inbar et al., 2022). The episode was focused on data quality in online samples. The hosts agreed that poor data quality in online samples is probably not a major issue in published research because "in published studies, that [poor data quality] is not very common, because if the data quality is low, that adds noise and so you don't get a significant result" and "I don't see how bad data quality could produce a false positive . . . it seems like it would just produce false negatives."

[3] "Careless" has pejorative connotations that may not always be applicable, such as if participants are trying to be careful but are temporarily distracted through no fault of their own or if participants are doing their best but the study design is so fatiguing or boring that some participants understandably struggle to pay full attention. We thus prefer the term "inattentive responding," but we nonetheless follow Ward and Meade (2023) in using the term "careless responding" for the purposes of limiting confusion and redundant construct proliferation.

[4] Even among this 19% of editors, their free-response explanations indicate that many think the risk of Type 1 errors can be ameliorated by using larger sample sizes or seem to think that only systematic patterns of carelessness (e.g., straightlining) can lead to Type 1 errors. As we explain elsewhere in this article, these ideas inaccurately decrease the perception of CR's tendency to cause Type 1 errors.

[5] For example, although Credé (2010) deserves credit and attention for being perhaps the first to systematically explain and demonstrate the inflationary effects of CR, his article is often not cited in literature reviews on the topic, even in articles focused on CR's inflationary dynamic (e.g., Huang et al., 2015). We suspect this may be because Credé used the term "random responding," which may have hampered the tendency for subsequent scholars to realize he was also describing "careless," "inattentive," or "indiscriminate" responding.

[6] For the purposes of this article, CR does not encompass other problematic response styles, such as willingly fraudulent responding that is often prevalent in paid online samples (for a discussion, see Chandler et al., 2020). CR also does not encompass what some have referred to as "weak satisficing" responding, in which participants are attentive enough to item content to be able to quickly offer a modestly reasonable response but do not "maximize" their cognitive effort and memory searching to respond to the intricacies of items and response options (Krosnick, 1999). Obviously, these response styles (and others, e.g., extreme responding, malingering, and acquiescence) introduce sources of variance that can contaminate results in different ways (e.g., see investigation of different contamination effects in Zijlstra et al., 2011).

[7] If items are not all keyed in the same direction, inflation could still occur if the careless respondents systematically select the neutral option/middle option of the rating scale (e.g., "neither true or false," "neither agree nor disagree"). Given that the middle option of a rating scale is often a response that allows one to avoid exerting full cognitive effort (e.g., Douven, 2018), it may be that CR tends to be disproportionately related to "the error of central tendency" (Guilford, 1954). To our knowledge, however, no extant research has empirically investigated whether and if so, under what conditions this potential link between CR and central tendency bias operates in real data.

[8] Although CR exists on a continuum (low carelessness to high carelessness), our approach in this article simplifies it by treating CR in terms of categorial groups (e.g., careful responders who passed all attention checks and careless responders who failed at least one attention check).

[9] Holtzman and Donnellan (2017) provided an Excel spreadsheet tool (available at https://nickholtzman.com/publications/) that can also be used to simulate the impact of CR on the association between two variables (e.g., a measure of psychopathy and a measure of narcissism); although this tool appears to be excellent for technical research, it requires substantial effort to understand and use.

[10] The app was implemented using the R package *shiny*. It is also available via a Git repository hosted on GitHub at https://github.com/fuhred/CarelessRespondingSimulator. Users can download the app and run it locally on their computer in RStudio.

[11] One other set of authors indicated that data for the excluded CR participants in their study were not retained after data collection. Another set of authors did not get back to us to provide their study data, and we decided not to press them. Of course, we must admit that we would have put more effort into collecting and reanalyzing data from additional studies if our three initial study reanalyses had ended up not being as helpfully illustrative as they ended up being. We also note that two of the articles we reanalyzed were published by colleagues of Brett Murphy and were initially identified partly because of his prior familiarity with those articles.

[12] This kind of open-ended response task served two aims quite well. First, even a sophisticated bot program will be unable to pass this task, which is important given that some research has indicated that bots can be programmed to overcome some kinds of attention checks (Pei et al., 2020). Second, it requires more mental effort than most typical attention-check items; thus, it has heightened capacity to identify insufficient effort by participants. We also note that Costello et al. (2022) presented the instructions as an image of text so (a) respondents could not copy and paste the example provided in the

instructions and slightly modify it and (b) it would be more difficult for bots to parse.

13. Participants were presented with a historical event (e.g., the stock market crash that triggered the Great Depression) and were asked to guess the date of the event via (a) a best-guess estimate, (b) a lower-bound year estimate, and (c) an upper-bound year estimate. Participants were excluded if they left an answer box blank, typed gibberish in a box, or gave answers that indicated a failure to follow the instructions.

14. Although these four variables did not meet our $p < .001$ cutoff, three out of the four demonstrated mean differences between the included (non-CR) and excluded (CR) participants at $p < .05$.

15. Examples of this included screening out participants based on unreasonable completion time, failed attention checks, nonsensical free responses, inadequate literacy comprehension, failing to follow instructions, presence of univariate or multivariate outliers, anomalous response patterns, and straightlining.

16. Optimally, researchers can first think through ways to prevent CR before participants even engage in the behavior. Some evidence supports that setting clear expectations for careful responding and providing warnings of the consequences of CR prevents some CR behavior (Huang et al., 2012; Ward & Meade, 2018; Ward & Pond, 2015). However, these prevention techniques rely on the hope that respondents are careful enough to read through study instructions to begin with and, of course, has no bearing on bot applications.

17. Malamis and Howley (2022) helpfully analogized these CR screening dilemmas to the differing "burden of proof" standards used in legal courts: For some issues, the legal system finds it appropriate to merely use a standard of "preponderance of the evidence," whereas for other issues, a much stricter "beyond a reasonable doubt" standard is used. The risks of punishing innocent parties versus the risks of letting culpable parties escape justice are balanced differently depending on the characteristics of the legal case. So, too, should CR screening stringency be tailored to characteristics of the research project.

18. For example, in Costello et al.'s (2022) Study 6, reviewed earlier, the internal consistency of the LWA scale improved slightly when the CR data were added (without CR: $\alpha = .91$, mean inter-item-correlation [MIC] = .21; with *CR*: $\alpha = .93$, MIC = .25).

19. Participants' carelessness increases as they work through long surveys (e.g., Bowling et al., 2021).

References

Arndt, A. D., Ford, J. B., Babin, B. J., & Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing*, *39*(1), 117–133. https://doi.org/10.1016/j.ijresmar.2021.05.001

Arthur, W., Jr., Hagen, E., & George, F., Jr. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Balzarini, R. N., Muise, A., Dobson, K., Kohut, T., Raposo, S., & Campbell, L. (2021). The detriments of unmet sexual ideals and buffering effect of sexual communal strength. *Journal of Personality and Social Psychology*, *120*(6), 1521–1550. https://doi.org/10.1037/pspi0000323

Bartusevičius, H., Bor, A., Jørgensen, F., & Petersen, M. B. (2021). The psychological burden of the COVID-19 pandemic is associated with antisystemic attitudes and political violence. *Psychological Science*, *32*(9), 1391–1403. https://doi.org/10.1177/09567976211031847

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229.

Carden, S. W., Camper, T. R., & Holtzman, N. S. (2019). Cronbach's alpha under insufficient effort responding: An analytic approach. *Stats*, *2*(1), 1–14. https://doi.org/10.3390/stats2010001

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49–55. https://doi.org/10.1037/abn0000479

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Costello, T. H., Bowes, S. M., Stevens, S. T., Waldman, I. D., Tasimi, A., & Lilienfeld, S. O. (2022). Clarifying the structure and nature of left-wing authoritarianism. *Journal of Personality and Social Psychology*, *122*(1), 135–170. https://doi.org/10.1037/pspp0000341

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

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

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920954925

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, *25*, 1203–1211.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*, 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38.

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), Article 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill.

Holden, R. R., Marjanovic, Z., & Troister, T. (2019). Indiscriminate responding can increase effect sizes for clinical phenomena in nonclinical populations: A cautionary note. *Journal of Psychoeducational Assessment*, *37*(4), 464–472. https://doi.org/10.1177/0734282918758809

Holtzman, N. S., & Donnellan, M. B. (2017). A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. *Personality and Individual Differences*, *114*, 187–192. https://doi.org/10.1016/j.paid.2017.04.013

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312–345. https://doi.org/10.1177/0013164419865316

Hong, M. R., & Cheng, Y. (2019). Clarifying the effect of test speededness. *Applied Psychological Measurement*, *43*(8), 611–623. https://doi.org/10.1177/0146621618817783

Huang, J. L., Curran, P., Keeney, J., Poposki, E., & DeShon, R. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., & DeSimone, J. A. (2021). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*, *36*, 807–828. https://doi.org/10.1007/s10869-020-09707-2

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Inbar, Y., Tullett, A., & Inzlicht, M. (Hosts). (2022, December 7). Is MTurk too good to be true? (No. 99) [Audio podcast episode]. In *Two psychologists four beers*. https://www.fourbeers.com

Jones, A., Earnest, J., Adam, M., Clarke, R., Yates, J., & Pennington, C. R. (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. *Experimental and Clinical Psychopharmacology*, *30*(4), 381–399. https://doi.org/10.1037/pha0000546

Kachanoff, F. J., Kteily, N. S., Khullar, T. H., Park, H. J., & Taylor, D. M. (2020). Determining our destiny: Do restrictions to collective autonomy fuel collective action? *Journal of Personality and Social Psychology*, *119*(3), 600–632. https://doi.org/10.1037/pspi0000217

Kelley, E. L. (1927). *Interpretation of educational measurements*. World Press.

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior Research Methods*, *50*(5), 1960–1970. https://doi.org/10.3758/s13428-017-0964-9

King, K. M., Kim, D. S., & McCabe, C. J. (2018). Random responses inflate statistical estimates in heavily skewed addictions data. *Drug and Alcohol Dependence*, *183*, 102–110. https://doi.org/10.1016/j.drugalcdep.2017.10.033

Krems, J. A., Williams, K. E. G., Aktipis, A., & Kenrick, D. T. (2021). Friendship jealousy: One tool for maintaining friendships in the face of third-party threats? *Journal of Personality and Social Psychology*, *120*(4), 977–1012. https://doi.org/10.1037/pspi0000311

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Lassetter, B., Hehman, E., & Neel, R. (2021). The relevance appraisal matrix: Evaluating others' relevance. *Journal of Personality and Social Psychology*, *121*(4), 842–864. https://doi.org/10.1037/pspi0000359

Litman, L., Moss, A., Rosenzweig, C., & Robinson, J. (2021). *Reply to MTurk, prolific or panels? Choosing the right audience for online research*. SSRN. https://doi.org/10.2139/ssrn.3775075

Litman, L., Rosen, Z., Ronsezweig, C., Weinberger, S. L., Moss, A. J., & Robinson, J. (2020). *Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data*. MedRXiv. https://doi.org/10.1101/2020.12.11.20246694

Malamis, P., & Howley, M. J. (2022). Anticipating careless responders in survey design and analysis. *Applied Clinical Trials*, *31*(11). https://www.appliedclinicaltrialsonline.com/view/anticipating-careless-responders-in-survey-design-and-analysis

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

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Pei, W., Mayer, A., Tu, K., & Yue, C. (2020). Attention please: Your attention check questions in survey studies can be automatically answered. In *Proceedings of the Web Conference 2020* (pp. 1182–1193). Association for Computing Machinery.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Sanchez, C., & Dunning, D. (2021). Jumping to conclusions: Implications for reasoning errors, false belief, knowledge corruption, and impeded learning. *Journal of Personality and Social Psychology*, *120*(3), 789–815. https://doi.org/10.1037/pspp0000375

Sharpe, B. M., Lynam, D. R., & Miller, J. D. (2023). Check and report: The state of data validity detection in personality disorder science. *Personality Disorders: Theory, Research, and Treatment*, *14*(4), 408–418. https://doi.org/10.1037/per0000601

Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, *34*(4), 346–360. https://doi.org/10.1177/1525822X221115830

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472–481. https://doi.org/10.20982/tqmp.16.5.p472

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231–263. https://doi.org/10.1111/apps.12118

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

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

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "reliability of speeded number-right multiple-choice tests". *Applied Psychological Measurement*, *33*(6), 488–490. https://doi.org/10.1177/0146621607304655

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Yentes, R. D., & Wilhelm, F. (2021). *Careless: Procedures for computing indices of careless responding* [R Package]. https://cran.r-project.org/web/packages/careless/index.html

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, *36*(2), 186–212. https://doi.org/10.3102/1076998610366263

Zorowitz, S., Niv, Y., & Bennett, D. (2021). *Inattentive responding can induce spurious associations between task behavior and symptom measures*. PsyArXiv. https://doi.org/10.31234/osf.io/rynhk