Controlling for careless responding requires causal justification

Taym Alsalti, University of Leipzig, Germany. ORCID 0000-0002-1767-1367[1];

Jamie Cummins, University of Bern, Switzerland. ORCID 0000-0002-9729-4900;

&

Ruben C. Arslan, University of Leipzig, Germany. ORCID 0000-0002-6670-5658.

Note that this is a work in progress and a preprint version of a 100% CI blogpost, uploaded here for archiving purposes.

Abstract

Careless responding is defined as a pattern of responding to survey items that reflects a latent construct distinct from—and disruptive of—the measurement of the primary construct of interest. In the social science literature, these response patterns are sometimes labeled "insufficient effort," "unserious," or "bogus, among other terms."The issue of such responses has been gaining in relevance as research data are increasingly collected through anonymous online surveys. Standard practice calls for identifying careless responders and excluding them from the data. The common reasoning behind this approach is that such responses can bias "descriptive" statistics (e.g., sample means) and effect size estimates / measures of associations (e.g., correlations and standardised mean difference) as well as increase measurement error. When they confound our causal effect of interest, deleting or adjusting for careless responses is justified. However, using directed acyclic graphs (DAGs), we show that different careless responding patterns can plausibly take on the role of a mediator, a collider, or both simultaneously, among others. Given this, we argue that, in contrast to previous calls, there cannot be a general rule about how to deal with careless responses and that greater attention to their data-generating mechanism from a causal perspective is necessary.

[1] Corresponding author: taymalsalti@gmail.com

What is careless responding?

"Careless" is often used to describe specific patterns of invalid responding, such as straightlining responses on a survey page (e.g., choosing "strongly agree" 10 times in a row) or choosing responses "randomly". The causal principles we discuss here definitely apply to such a specific understanding of careless responding (CR), but not only. We thus use CR to refer to any pattern of responding that reflects a latent construct which is distinct from, and interferes with the measurement of, the construct of interest[2]. CR is caused by person (e.g., personality), study (e.g., boringness), and context (e.g., what music the participant is listening to while taking the study). So we also mean response patterns here that people don't usually think of when they read "careless responding", such as acquiescence bias, central tendency bias, socially desirable responding, and intentionally fraudulent responses.

Measures to ensure data fidelity predate the public world wide web (e.g., the MMPI validity scales) and are often needed in any human data collection endeavour. However, nowadays people most often complain about CR in the context of anonymous online surveys, where a few major blunders recently generated some buzz, e.g., that one time YouGov/The Economist overestimated the proportion of young people yay-saying the statement "The Holocaust is a myth" by ~17 percentage points (20% by YouGov vs. 3% by Pew). Concerns about data quality issues in online surveys are probably justified, and things seem to be getting worse for surveys that include open-ended questions which are economical to answer using LLM responses.

How to deal with it?

Well, first we have to measure it. There are many different ways to do that and we refer the reader to Goldammer et al., 2020 and Ward & Meade, 2023 for reviews and guidelines, but standard practice involves using such CR identification measures to form binary decision rules based on which the offenders are list-wise excluded. And it is with this meet-'n-yeet approach as a standard way of dealing with CR that we take issue[3] — with few exceptions (e.g., Alvarez & Li, 2023 caution that excluding careless respondents can compromise sample representativeness), the prevailing guideline is that once identified, careless responses should be purged.

The common reasoning behind this approach is that such responses can bias "descriptive" statistics (e.g., sample means) and effect size estimates / measures of associations (e.g., correlations and standardised mean difference) as well as increase measurement error. CR *can* definitely do these things if we don't account for it, but it can also bias our sample estimates if we *do* account for it. In the following, we show under which conditions this can happen using directed acyclic graphs (DAGs)[4]. We illustrate and provide some examples for a few basic, simple X → Y constellations (see e.g., Rohrer, 2018 or Wysocki et al., 2022, but the same principles apply in the case of more complex causal structures or when no causal effects are targeted at all (e.g., when estimating the population distributions of certain variables or the associations between others).

---

[2] Many *many* other terms are used in the literature, with unclear overlaps in their definitions e.g., inattentive, insufficient effort, lazy, random, disengaged, nonserious, mischievous, bogus, untruthful, fraudulent, etc. The question of how much these meaningfully overlap is in itself interesting, but not the focus here.

[3] The dichotomisation of a continuous phenomenon is a problem in and of itself, but we're picking our fights here. Although we do use the more general "control for" here to include alternatives such as weighting using metric measures of CR (see Ulitzsch et al., 2024, for an example) or including them as covariates in a regression analysis.

[4] Well, not 100% kosher DAGs, to be upfront, since they incorporate bidirectional associations (dashed lines), but you can think of these non-kosher elements as annotations of the DAGs proper.

CR as a confounder

One constellation where controlling for CR is unambiguously necessary is when it causes both our predictor (X) and outcome (Y). The result is a "spurious" or a "spuriously inflated" association between X and Y (Stosic et al. 2024). In DAG, this pattern is traditionally depicted as a "fork":

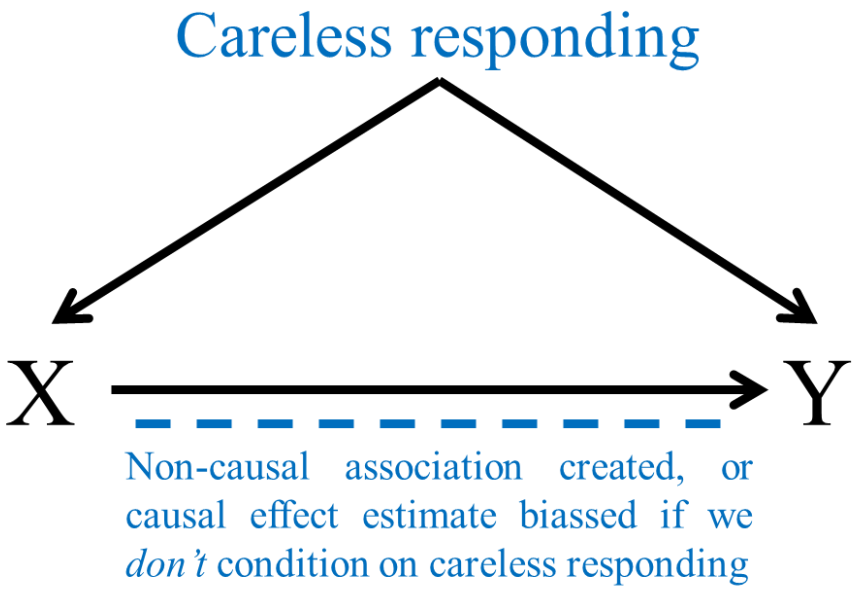

Figure 1: CR as a confounder

One way we might be dealing with this scenario is when the distributions of serious X and Y responses are skewed (such as when they show floor or ceiling effects), while careless responses randomly or systematically cluster around the two scales' middle points[5]. An example is if our X is self-reported frequency of watching InfoWars and Y is the self-reported frequency of bleach ingested to prevent and treat COVID-19. The means (and variances) of serious responses on both X and Y will be very close to the scales' lower ends, whereas careless responses will either bunch up around the scales' midpoints (if they're random) or cluster at the scales' higher ends (if they're produced by trolling).

Random-like or central tendency biassed responses can also deflate a true association[6]; extreme responses on either ends of the scales for the sake of trolling can inflate a true association; acquiescence-biassed responses can inflate a positive association and deflate a negative one. The possibilities are not endless but numerous; and all such cases call for conditioning on CR.

[5] If you're a visual learner, picture in your mind's eye a scatter-plot where the vast majority of the points belonging to diligent responders cluster in a small square in the corner of the plot's canvas, while careless responses cluster in the middle of the plot. This "elongates" the data point cloud and makes the regression line going through it less horizontal. See Stosic et al. (2024) if you want to picture it in your regular eye.
[6] The cloud in the scatter plot thickens.

CR as a mediator/collider

Say we have reason to believe that X will influence CR instead of the other way around, i.e., its effect on Y might be partially mediated by CR. In such a case, excluding CR may amount to conditioning on a collider (i.e., a variable that modifies the association between two variables that cause it):

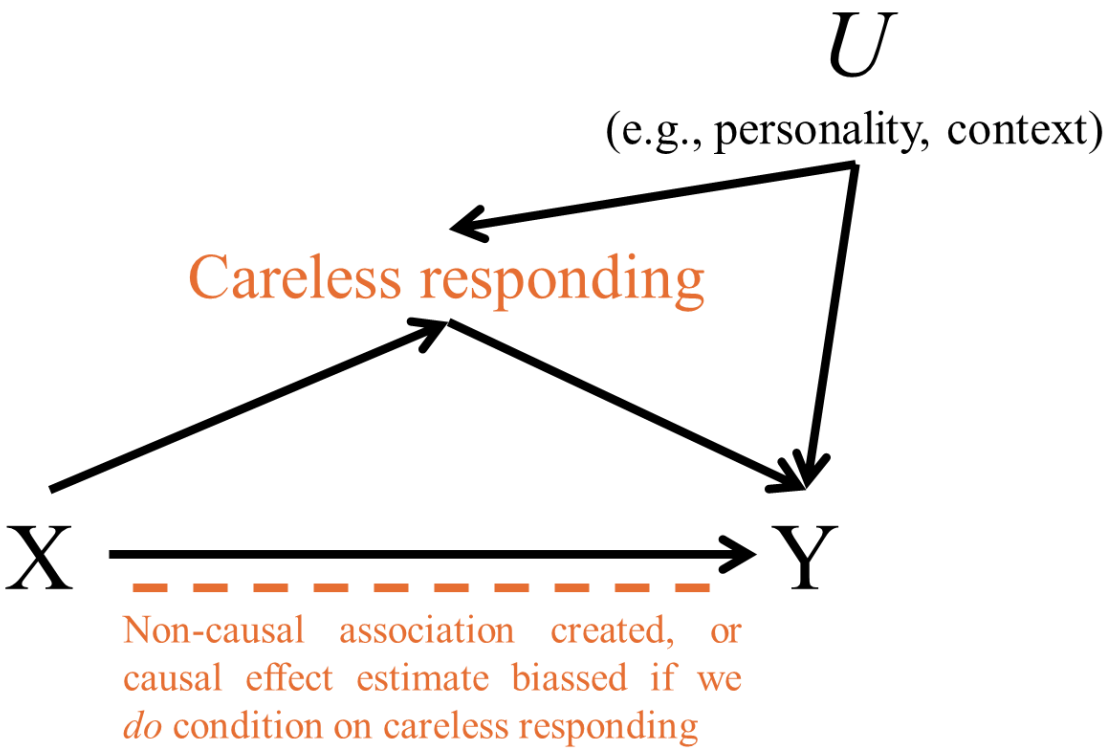

Figure 2: CR as a mediator/collider

This could occur if, for instance, our manipulation (X) is a simple flanker task in the control condition while participants additionally receive musical feedback on their performance in the experimental condition. Here the enrichment of the experimental condition reduces CR in it compared to the control group, which in turn affects Y (e.g., reaction time). Since we can't directly manipulate either Y or CR[7], they are likely influenced by variables we do not necessarily measure or account for (*U* in the DAG), such as the test takers' personality or the environment in which they're taking the survey. This so-called *backdoor path* remains closed (i.e., does not impact our estimate of the effect of X on Y) as long as we restrict our analysis to the bivariate effect of X on Y. But conditioning on CR, e.g., by excluding those deemed as showing it, opens the backdoor path and lets the confounding between CR and Y seep into our estimate of the effect of X on Y.

In the context of experiments this type of collider bias is known as post-treatment bias because it involves conditioning on a variable that is at least partially a descendant of (i.e., affected by) the

[7] Incorporating potential causes of CR into our DAGing is a difficult, but probably worthwhile enterprise. While personality and context can never be completely accounted for, we can attempt to partially get at their impact by, for instance, including a measure of personality which we assume is relevant (e.g., conscientiousness) and explicitly asking about the situation in which the participant is taking the survey. On the other hand, we invariably condition on survey characteristics whether we wish to or not, although there too we can experimentally manipulate some aspects of the survey design (e.g., reward for participation) to investigate their impact.

treatment. For example, say we want to test the effect of ritalin on ADHD symptoms in a new randomised experience sampling method study. It's likely that more participants in the control group will be flagged for CR than in the experimental condition. Purging CR would thus lead to excluding more of those with the worse symptoms in the control group, leading to an overestimation of the control group symptomatology mean, and consequently a deflated treatment effect estimate[8].

CR as a collider

There are also many ways we could end up with an inverted fork instead of a fork, with CR becoming a simple collider between X and Y themselves:

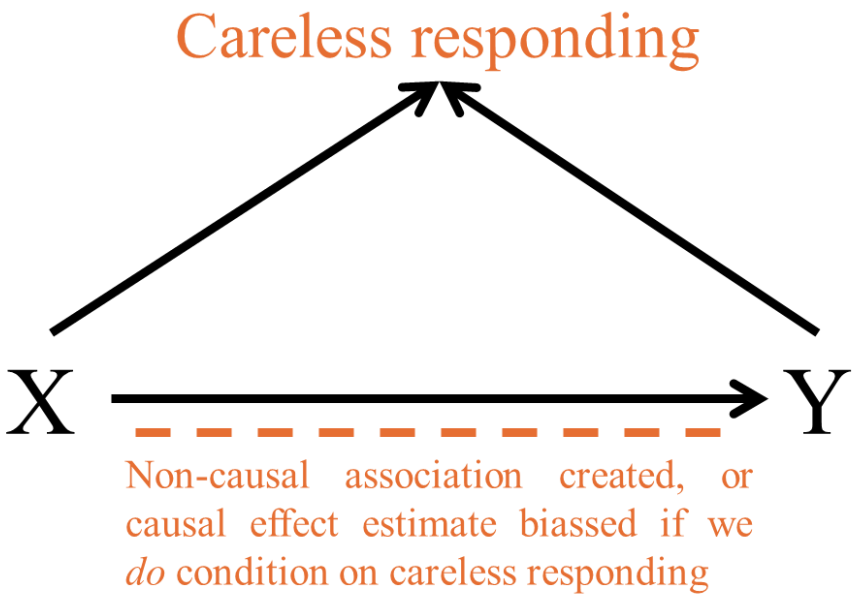

Figure 3: CR as a collider

This could be a problem if we are trying to estimate the effect of, say, conscientiousness (X) on scores of some cognitive test that incorporates an attention component (Y). If we use attention checks to exclude respondents who fail them and reasonably assume that both X and Y affect the probability of such a failure, we might introduce a spurious association between self-efficacy and performance on the cognitive test, or bias the estimate of any true effect that might exist. That is, because diligent responders will tend to be more conscientious and have a stronger latent ability to concentrate, the association of X and Y will be overestimated[9].

---

[8] Importantly, conditioning on CR is only problematic for internal validity (i.e., validity of the causal inference about the treatment) if we do it based on data collected *after* the start of the treatment (hence the *post*-treatment). If we collect data in a pre-trial phase for example, then conditioning on CR is fully harmless - for internal validity. The extent to which excluding the least attentive respondents from a study on ADHD is an external validity (i.e., generalisability/transportability) question that we also have to reckon with.

[9] You're again invited to picture in your mind's eye a scatter-plot, now showing a cloud of points strewn all over the place, indicating no association between conscientiousness and cognitive ability. Excluding careless respondents in this case approximately amounts to removing participants in the lower left corner of the plot, which introduces a spurious negative association between the two variables.

CR as a descendant/collider

Finally, let's say X is self-efficacy and Y is scores on a fluid intelligence test. It's reasonable to think of attention span as a facet of fluid intelligence, which would make the likelihood to pass an attention check for example a (partial) descendant of Y (i.e., a variable that is caused by Y):

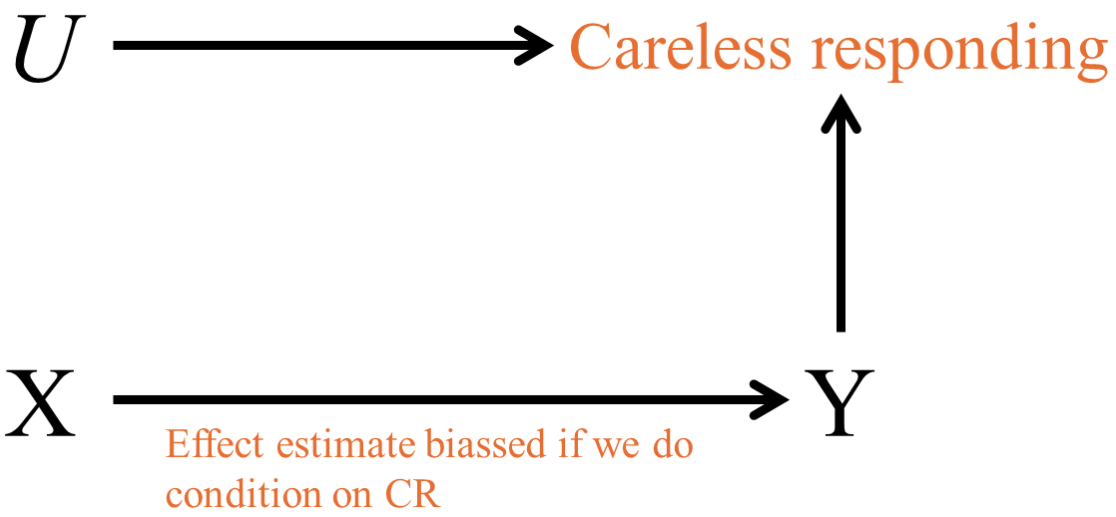

Figure 4: CR as a descendent of the outcome

Conditioning on a descendant of the outcome generally disrupts our effect size estimates in similar ways to conditioning on the outcome itself, i.e., a missing not-at-random (MNAR) scenario. However, here the estimation is disrupted further by the fact CR will be affected by other variables (e.g., study and context), introducing collider bias. This might not create an association out of thin air like the above scenarios could, but it very well could lead us to miss out on true effects by decimating Y's sample variance[10]. In the case of Y as a measure of cognitive ability, this happens because we would be removing those on the lower end of the latent scale.

So what do we actually do about it?

Sadly, there is no easy fix for CR-ridden data. As shown by the examples above, CR can play a role in many different ways, and the best course of action thus always depends on the specific case. Especially when it comes to online studies, how to measure CR and what to do about it are issues that deserve attention and careful thought. A reasonable workflow could look something like this:

1. DAG your Xs and Ys
2. Consider what types of CR might be plausibly associated with your Xs and Ys and incorporate them into your DAG(s)
3. Explore which CR measures are feasible and most likely to reflect the DAGed causal patterns. Do you want to implement external CR detection measures such as instructed response items (e.g., 'please choose "fully agree"') and bogus items (e.g., "I am hexagonal shaped")[11]? Which ad-hoc measures (e.g., long-string index for detecting straightlining) have plausible

[10] Although type of range restriction in the only outcome can also inflate effect sizes if we're talking about "standardized" effect sizes (e.g., Cohen's d or correlation coefficients, see Alsalti et al., 2024)

[11] Pssst they might be redundant at best if you have enough data to calculate ad-hoc measures (e.g., consistency measures, Yentes, 2020).

relationships with your Xs and Ys and what are the relationships' directions? If you want to exclude careless respondents, decide on thresholds for excluding CR (e.g., failing 3 instructed response items). If you're fortunate enough to be doing experimental work, make sure you assess CR before the manipulation. That way, it cannot be affected by your X. The remaining concern is that results from your sample of diligent responders might not generalize well to the population, so you're not scot-free. If you think controlling for CR might be disadvantageous in your case, do yourself a favor and spell it out in your pre-registration, so readers won't think it was a post hoc decision. In fact, always spell out how you will deal with CR in your pre-registration.

4. Conduct sensitivity analyses. Especially in complex multivariate scenarios (multiple predictors and multiple CR measures), things might turn out to be very different from what you expected (e.g., the prevalence of CR might turn out to be much larger based on the chosen thresholds). In this case, it helps to try out different thresholds for different CR measures and evaluate to what extent conditioning on CR influences the crucial hypothesis tests.

Acknowledgements

This comment ensued from a discussion at the Cake Club of Stosic et al's (2024) paper. We thank members of the Cake Club for their contributions to the discussion, as well as non Cake Club members Julia Rohrer and Anne Scheel for the helpful feedback on previous versions of this work.

References

Alsalti, T., Protzko, J., Lakens, D., Elson, M., & Arslan, R. C. (2024). *From Ells to Metres: Population norms should supersede sample-local standardisation*. OSF. https://doi.org/10.31234/osf.io/z34hg

Alvarez, R. M., & Li, Y. (2023). Survey Attention and Self-Reported Political Behavior. *Public Opinion Quarterly*, *null*, null. https://doi.org/10.1093/poq/nfac048

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

Matsuura, T., Hasegawa, A., Akiyama, M., & Mori, T. (2021). Careless Participants Are Essential for Our Phishing Study: Understanding the Impact of Screening Methods. *Proceedings of the 2021 European Symposium on Usable Security*, *null*, null. https://doi.org/10.1145/3481357.3481515

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. https://doi.org/10.1177/2515245917745629

Stosic, M. D., Murphy, B. A., Duong, F., Fultz, A. A., Harvey, S. E., & Bernieri, F. (2024). Careless Responding: Why Many Findings Are Spurious or Spuriously Inflated. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459241231581. https://doi.org/10.1177/25152459241231581

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, *56*(2), 804–825. https://doi.org/10.3758/s13428-022-02053-6

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*(Volume 74, 2023), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221095823. https://doi.org/10.1177/25152459221095823

Yentes, R. D. (2020). *In search of best practices for the identification and removal of careless responders*. North Carolina State University. https://search.proquest.com/openview/2fb9b886e3c9918372fa70a98caec76f/1?pq-origsite=gscholar&cbl=18750&diss=y