Meta-Psychology, 2022, vol 6, MP.2020.2637
https://doi.org/10.15626/MP.2020.2637
Article type: Commentary

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Danielsson H., Carlsson, R.
Reviewed by: Shirlene Wang, Oscar Olvera Astivia
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/X5QHF

Careless Response Processes are Heterogeneous: Comment on Goldammer et al. (2020)

Alexander J. Denison
Department of Psychology, University of South Florida; Department of Education and Human Development, Clemson University

Brenton M. Wiernik
Department of Psychology, University of South Florida

Abstract

Goldammer et al. (2020) examined the performance of careless response detection indices by experimentally manipulating survey instructions to induce careless responding, then compared the ability of various indices to detect these induced careless responses. Based on these analyses, Goldammer et al. concluded that metrics designed to detect overly consistent response patters (i.e. longstring and IRV) were ineffective. In this comment, we critique this conclusion by highlighting critical problems with the experimental manipulation used. Specifically, Goldammer et al.'s manipulations only encouraged overly inconsistent, or random, responding and thus did not induce the full range of behavior that is present in natural careless responding. As such, it is unsurprising that metrics designed to detect overly consistent responding appeared to be ineffective. Because the full range of careless behavior was not induced, Goldammer et al.'s study cannot address the utility of longstring or other consistency metrics. We offer recommendations for alternative experimental manipulations that may produce more naturalistic and diverse careless responding.

Keywords: careless responding, insufficient effort responding, data cleaning, methodology

Comment on Goldammer et al. (2020)

Goldammer et al. (2020) presented two studies aimed at investigating the performance of various statistical methods for detecting careless respondents. Their first study experimentally manipulated response patterns using various instruction sets and examined the performance of commonly used careless response detection methods. Based on these results, Goldammer et al. provided recommendations about which detection methods are effective and ineffective. In this comment, we discuss several important limitations of the careless response manipulations used in this study, whether these manipulations produced behavior consistent with natural careless responding, and the appropriateness of Goldammer et al.'s recommendations.

Defining Careless Responding

Careless or insufficient effort responding[1] is part of a larger construct of carelessness or inattentiveness, which has been used to describe response behaviors that arises when individuals are not motivated to give honest, thoughtful responses to questions (Curran, 2016; Johnson, 2005). Careless responding occurs when this inattentiveness results in individuals answering items in a content non-responsive manner – i.e. without paying attention to the content or instructions of the items (Curran, 2016; Meade & Craig, 2012; Nichols et al., 1989). This is distinct from other aberrant response patterns such as faking, which is content-responsive. When faking, individuals provide invalid data, but their responses are contingent on the content of the items they are responding to (Nichols et al., 1989).

Content non-responsive behavior has a long history of study in psychology, dating back over 30 years to examinations of detecting such behavior in the MMPI (Baer et al., 1997; Berry et al., 1992; Nichols et al., 1989) and examining factors arising from negatively keyed items (Schmitt & Stuits, 1985). The construct of careless responding grew out of these early investigations as a way to describe content non-responsive behavior that was due to a lack of motivation. While early studies focused on overly inconsistent or "random" responding, it is now well accepted that this behavior can take on two different forms—overly consistent or overly inconsistent responding.

Overly inconsistent responding is often conceptualized as randomness or random responding, as it is assumed that individuals select their response to each item completely at random. While this behavior is discussed as being completely random, researchers generally seem to use the word random in the colloquial and not the statistical sense of the word (Curran, 2016). Instead, this behavior may be better conceptualized as highly inconsistent responses and is generally characterized by a high degree of variance within response strings, such as using every response option at least once on a question block (Curran, 2016; Marjanovic et al., 2015). Conversely, overly consistent responses follow some pattern, such as responding with the same anchor point to every item or varying responses in some pattern, such as "1, 2, 1, 2, 1, 2…" or "1, 2, 3, 4, 5, 4, 3, 2, 1…" (Curran, 2016; Meade & Craig, 2012).

While careless responding is the most commonly investigated construct when examining content-nonresponsive behavior, it is important to note that not all studies in the literature investigate this construct in its entirety. Specifically, there are some studies that explicitly investigate random responding (e.g., Berry et al., 1992; Credé, 2010; Osborne & Blanchard, 2011) with no investigation of consistent responding. Although random, or more aptly termed inconsistent, responding is one piece of the careless responding construct, it is not equivalent to the construct of careless responding as a whole. Any investigation of purely random responding will overlook a large chunk of the careless responding construct that involves response consistency. The conflation of random responding with *all* careless responding is something most authors are careful to avoid. Indeed, in all of the aforementioned examinations of random responding the authors are careful to separate random and careless responding, and at most talk about carelessness as the underlying cause of random responding.

These two conceptualizations of careless response behavior are reflected in the metrics that have been developed to catch these respondents. As Goldammer et al. note there are two general categories of indirect careless response metrics. The first category is *invariability metrics*, which detect overly consistent response behavior. This category includes longstring analysis and intraindividual variability. The second category is *consistency metrics*, which detect overly inconsistent behavior. This category includes a person's response reliability and semantic or psychometric antonyms/synonyms. See Goldammer et al. (2020) or Curran (2016) for a description of these metrics.

Inducing Careless Responding

Given the above definition of careless responding as content non-responsive, it is critical to ensure that any manipulation intended to induce careless behavior is consistent with this definition. It is also important to consider whether experimentally induced behavior provides a good proxy for what that behavior is like in the real world. For comparison, research on the aberrant response pattern of faking has consistently found that faking induced through researcher instructions does not approximate non-directed or "real-world" faking (Kuncel et al., 2011; Viswesvaran & Ones, 1999). It is questionable whether either of Goldammer et al.'s manipulations induce behavior that is a good proxy for real-world careless responding or that is even consistent with the definition of careless responding.

Goldammer et al.'s (2020) first manipulation is "opposite responding," wherein participants are instructed to respond to items using the opposite scale point they

[1]The terms careless responding and insufficient effort responding are used to refer to the same underlying construct in almost all cases. This behavior has also been called inattentive responding, random responding, and several other terms. In this paper we opt to use the term careless responding when discussing the construct.

would normally choose. This response pattern clearly violates the definition of careless responding as it is *not* content non-responsive behavior; participants are instructed to attend to the item content and then to respond in a specific aberrant way based on what they would normally answer. That is, respondents must attend to the content of a given item to generate their initial response, then reverse that response. Although these data may be considered invalid, as half of their responses will not align with the other half, they are produced from content-responsive, not careless, behavior.

This is not to say that careless response indices will be unable to detect such behavior; in fact, we can see that the metrics designed to detect inconsistent response patterns do a good job of capturing this "opposite responding" pattern. However, this result is unsurprising because this manipulation, by design, makes a participant's responses inconsistent and introduces a high degree of variance. Conversely, it would be nearly impossible for the "opposite responding" manipulation to produce an overly *consistent* response pattern that would be detected by longstring analysis, for example. In fact, longstring would only detect the opposite responding pattern if a participant happened to respond to all items at the midpoint. Consistency indices were not designed to detect an "opposite response" pattern, and it is neither unexpected nor informative that they, in fact, do not detect such behavior. Thus, the results from this manipulation do not have any bearing on the utility of inconsistency or consistency indices to detect careless responding because the responses are not careless.[2]

Goldammer et al.'s second manipulation is "random responding." Participants in this condition were instructed on alternating pages to either "complete the questions below exactly as they apply to you" or to "choose any response option, no matter whether it applies to you or not" (p. 4). It is again unclear what real-world response pattern this represents. Under what conditions would participants repeatedly start and stop attending to item content? Even if participants were partially careless, it is unlikely they would exhibit this in alternating blocks of responses. By repeatedly changing the instructions from careful to random and vice versa, this manipulation will increase the within-person variability of responses, but it is not clear that these responses will resemble real-world random responding, much less careless responding as a whole.

Even if these instructions are assumed to reasonably approximate random responding, these responses would nevertheless not reflect the full scope of response behavior exhibited in real-world careless responding. Previous work has extensively documented that careless responding is not simply responding randomly to items, but can also manifest as certain forms of patterned or consistent responding (Curran, 2016; Jaso et al., 2021; Johnson, 2005; Meade & Craig, 2012). For example, a respondent might try to complete a survey quickly by clicking the same response option repeatedly. Invariability metrics are designed specifically to capture this type of behavior because we know overly consistent responding is a type of behavior careless respondents engage in. Because the manipulation used by the authors is unlikely to produce overly consistent behavior, it is again unsurprising that the invariability metrics failed to detect these respondents.[3] As such, the data have no bearing on the utility of invariability indices because invariability indices were not designed to detect the random or inconsistent behavior induced by the authors manipulation.

Are Invariability Metrics Useless?

A key takeaway from our examination of both of Goldammer's manipulations is that the response patterns they produce will be highly inconsistent. Thus, it is unsurprising that Goldammer et al. (2020) conclude that metrics designed to detect overly inconsistent behavior (consistency metrics) are effective at detecting careless responding, whereas the invariability metrics designed to detect *overly consistent responding* are not. This finding does not mean that invariability metrics are useless, but rather that the manipulations were not designed to produce consistent responding.

Goldammer et al.'s (2020) conclusions in their Study 1 and subsequent recommendations against the use of invariability indices are especially problematic because we know that actual careless respondents do sometimes produce overly consistent response patterns. For example, Johnson (2005) identified that 3.5% of partic-

[2] It is somewhat difficult to determine what sort of real-world response pattern the "opposite responding" is intended to reflect. Rarely would a person be expected to actively switch their scale point use partway through a survey. One possibility would be if the anchor points for scales were reversed midway through a multi-part survey and the participant did not notice this reversal. However, such a change would go against general best practice for survey design (Stern et al., 2007) and would likely be better simulated by instructing respondents to respond normally and randomly reverse-coding items for a subset of respondents.

[3] The actual participation instructions used did not explicitly state that participants should respond at random, but it seems highly likely that participants would interpret these instructions this way. Indeed, Golammer et al. appeared to assume as much, as their descriptions of the manipulation clearly communicate that they thought participants were engaging in random responding.

ipants in their dataset responded by selecting the same response option repeatedly throughout the survey; similar patterns have been noted in other studies of careless responding (Curran, 2016; Jaso et al., 2021; Meade & Craig, 2012). Therefore, applying only consistency metrics in a real dataset (as done in Goldammer et al.'s Study 2) will overlook a potentially large portion of careless respondents who are engaging in overly consistent response behavior. Even if few or no overly consistent respondents are found, there is little downside to computing these metrics.

How Can We Experimentally Study Careless Responding?

Above, we critique Goldammer et al.'s manipulations as unlikely to produce response behavior that resembles real careless responding. This raises the question of how careless responding *could* be experimentally studied. We offer several possibilities.

First, rather than instructing participants to respond in a specific way, researchers can instead instruct participants to optimize their responses based on a specific goal that is relevant for the population being studied. For example, study participants recruited for extra credit in university classes or platforms like MTurk might aim to complete surveys as quickly as possible to receive their compensation. To simulate such a strategy, participants could be instructed to "answer items as quickly as possible" or to "answer items as quickly as possible, while still appearing to respond carefully." (cf. Huang et al., 2012). This type of instruction may more accurately reflect the types of thought processes that lead to real-world careless responding, and participants might engage in a variety of behaviors to achieve this goal (e.g., random responding, patterned responding). This approach is not foolproof. It may be difficult to write instructions that capture the myriad of factors that research participants simultaneously weigh when choosing a response strategy (cf. directed "fake good" instructions do not produce the same response strategies test takers use in real high-stakes settings; Kuncel et al., 2011; Viswesvaran & Ones, 1999). Response speed is also not the only goal participants optimize toward when responding carelessly, otherwise response time would be the only metric needed to detect carelessness.

A second approach might be manipulations designed to decrease the probability that individuals respond carelessly. For example, participants in one condition could be warned that "The researchers will be able to detect if you have responded carelessly. You will not receive compensation if you respond carelessly." (cf. Gibson, 2019; Huang et al., 2012). Researchers could also employ a virtual presence (such as a human, or more abstract entity) to make participants feel monitored, in addition to warning them (cf. Ward & Pond, 2015). When participants are made to feel monitored or warned, carelessness rates should be lower. Accordingly, careless response indices would be expected to detect lower rates of carelessness in these conditions, compared to control conditions without a warning. Similarly, performance of careless responding indices could be compared across data collected in comparatively high-stakes (e.g., a job application) versus low-stakes (e.g., an extra credit or MTurk study) settings; careless responding rates would be expected to be lower in high-stakes contexts. However, the exact prevalence of careless responding in each sample would still be unknown, which could make comparisons between conditions difficult.

A third approach might be to directly ask participants whether they responded to a survey carelessly and then examine whether careless responding indices can detect participants who responded affirmatively to this item (cf. a single item asking participants about their data quality can effectively detect a high percentage of careless respondents; Meade & Craig, 2012). While this solves the problem of determining who is careless, it's possible that the worst cases of carelessness will still be missed. That is, if someone is truly paying no attention to any items, they may answer no to this question by chance.

Infrequency, trap, or instructed response items could also be used. These are items that have a correct answer that any conscientious respondent should be answer correctly (e.g., "select option 5") but are answered incorrectly by careless participants because they are not paying attention (Curran & Hauser, 2019; Huang et al., 2015). While these items are effective at detecting careless participants, a potential disadvantage is that careless participants could circumvent these items if they are skimming questions in an attempt to not be caught. This method also does not solve the problem of inducing careless responding in the first place, so relies on this behavior being naturally present in the data.

Last, a fourth approach might be to induce content non-responsive responses by using items with nonsensical or blank content (cf., Maul, 2017, but see also Curran and Hauser, 2019; Rhemtulla et al., 2017). While this will produce behavior that is non-responsive to the item content, since that content does not exist, this may produce content-nonresponsive behavior that differs from careless responding to substantive scales.

As outlined above, each of these approaches has potential advantages and disadvantages, and experimental studies of careless responding should triangulate re-

sults across multiple approaches. The largest problem with inducing careless behavior is the variety of potential forms and motivations behind this behavior (e.g., some participants may want to finish as fast as possible, while others may want to exert as little effort as possible), which makes a precise operationalization difficult. The fact that this behavior is caused by a lack of motivation or effort also means that any instructed manipulations may not induce naturalistic careless responding because participants are now exerting conscious effort to produce this behavior. This is again why using a variety of approaches to triangulate this behavior seems most useful and why the process that elicits this behavior should capture how a lack of motivation will influence response patterns.

While addressing the above problem is beyond the scope of this paper, applied researchers are advised to ask a different question in the interim: what constitutes a conscientious response pattern? By doing so researchers can uncover response patterns that are theoretically impossible for a conscientious respondent to produce and screen for them with careless response metrics. For example, identical responses to every item on a positive affect scale might be theoretically possible, if unlikely, but identical responses to every item, or even to half the scale, on the BFI-10 does not make theoretical sense (Rammstedt & John, 2007). Thus, different theoretically impossible longstring cut scores could be produced for these different scales. In doing so, researchers should consider a variety of careless responding metrics, including *both* consistency metrics and invariability metrics, to identify the full range of potentially invalid responses.

Author Contact

Alexander J. Denison
https://orcid.org/0000-0002-4291-8325
Brenton M. Wiernik
https://orcid.org/0000-0001-9560-6336
Correspondence concerning this article should be addressed to Alexander J. Denison or Brenton M. Wiernik. Email: adeniso@clemson.edu or brenton@wiernik.org

Conflict of Interest and Funding

The authors have no conflicts of interest or funding to disclose.

Author Contributions

Conceptualization: Alexander J. Denison and Brenton M. Wiernik.
Investigation: Alexander J. Denison.
Project Administration: Alexander J. Denison.
Supervision: Brenton M. Wiernik.
Writing - Original Draft Preparation: Alexander J. Denison.
Writing - Review Editing: Alexander J. Denison and Brenton M. Wiernik.

Open Science Practices

This article is a commentary and as such had no data or materials to share, and it was not pre-registered. The entire editorial process, including the open reviews, is published in the online supplement.

References

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the mmpi–a [[Online; accessed 2020-08-15]]. *Journal of Personality Assessment*, *68*(1), 139–151. https://doi.org/10.1207/s15327752jpa6801_11

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). Mmpi-2 random responding indices: Validation using a self-report methodology. [[Online; accessed 2020-09-15]]. *Psychological Assessment*, *4*(3), 340–345. https://doi.org/10.1037/1040-3590.4.3.340

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research [[Online; accessed 2020-02-04]]. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/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

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items [[Online; accessed 2019-09-24]]. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

Gibson, A. (2019). *Stop what you're doing, right now! effects of interactive messages on careless responding* (Doctoral dissertation). Ohio, Wright State University. https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=3260&context=etd_all

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies [[Online; accessed 2020-03-05]]. *The Leadership Quarterly*, 101384. https://doi.org/https://doi.org/10.1016/j.leaqua.2020.101384

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions [[Online; accessed 2021-05-11]]. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10/gd86zr

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys [[Online; accessed 2018-08-21]]. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/https://doi.org/10.1007/s10869-011-9231-8

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. [[Online; accessed 2021-10-17]]. *Psychological Methods*. https://doi.org/10.1037/met0000312

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories [[Online; accessed 2018-08-21]]. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/https://doi.org/10.1016/j.jrp.2004.09.009

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication [tex.ids: Kuncel2011PleaProcess, Kuncelpleaprocesspersonality2011a]. *Human Performance*, *24*(4), 373–378. https://doi.org/10.1080/08959285.2011.597476

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (isd): An index that discriminates between conscientious and random responders [[Online; accessed 2020-03-12]]. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10/gddnqh

Maul, A. (2017). Rethinking traditional methods of survey validation [publisher: Routledge $_{e}$*print* : *https* : *//doi.org*/10.1080/15366367.2017.1348108]. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data [[Online; accessed 2018-08-21]]. *Psychological Methods*, *17*(3), 437–455. https://doi.org/https://doi.org/10.1037/a0028085

Nichols, D., Greene, R., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the mmpi: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 12. https://doi.org/https://doi.org/10.1002/1097-4679(198903)45:2{\textless}239::AID-JCLP2270450210{\textgreater}3.0.CO;2-1

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results [publisher: Frontiers]. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00220

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german [[Online; accessed 2021-10-22]]. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rhemtulla, M., Borsboom, D., & Bork, v. R. (2017). How to measure nothing [publisher: Routledge $_{e}$*print* : *https* : *//doi.org*/10.1080/15366367.2017.1369785]. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 95–97. https://doi.org/10.1080/15366367.2017.1369785

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? [[Online; accessed 2020-03-12]]. *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/https://doi.org/10.1177/014662168500900405

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey [[Online; accessed 2013-08-10]]. *Survey Research Methods*, *1*(3), 121–138. https://doi.org/10.18148/srm/2007.v1i3.600

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement [[Online; accessed 2020-04-08]]. *Educational and Psychological Measurement*, *59*(2), 197–210. https://doi.org/https://doi.org/10.1177/00131649921969802

Ward, M., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on internet-based surveys [[Online; accessed 2020-02-03]]. *Computers in Human*

Behavior, *48*, 554–568. https://doi.org/https://doi.org/10.1016/j.chb.2015.01.070