Response to reviewers for "Inattentive responding can induce spurious associations between task behavior and symptom measures"

Samuel Zorowitz[1], Johanne Solis[2], Yael Niv[1,3], Daniel Bennett[4]

[1]Princeton Neuroscience Institute, Princeton University, NJ, USA
[2]Rutgers-Princeton Center for Computational Cognitive Neuropsychiatry, Rutgers University, NJ, USA
[3]Department of Psychology, Princeton University, NJ, USA
[4]School of Psychological Sciences, Monash University, Victoria, Australia

Formatting note

In the document below, reviewers' comments are indicated in bold typeface. Quoted text from the revised manuscript is presented in quotation blocks; within these excerpts, text that is new is indicated in italicized typeface. All updates to text in the revised manuscript are detailed below.

Round 1 Reviews

Reviewer #1

The authors presented an interesting paper in which they argued that spurious correlations may arise between task behavior and psychiatric symptoms if online studies do not conduct attention or effort level check. Specifically, the authors launched an online study themselves using two common subject pool platforms: Amazon mTurk and Prolific. They found that behavior-symptom correlations were completely abolished when inattentive subjects were removed. The authors also suggested that excluding subjects based on task behavior alone was not sufficient, and that larger sample size further inflated the problem. I find this study timely

and important, as online studies are increasingly popular amongst behavioral researchers, particularly since the pandemic started. I also generally agree with the recommendation that online studies should include careful attention check. However, several major flaws in the rationale and design of this study and lack of novelty significantly diminished my enthusiasm.

1. A key assumption here is that people who failed attention check artificially inflated the self-reported mental health symptoms (which then subsequently inflated correlations with behavior). This would only be true if the "failed" subjects are actually healthy with low mental health symptoms. However, a very likely alternative is that these so called inattentive/low effort subjects indeed have higher mental health symptoms (as the authors acknowledged in the discussion). In fact, there is a very rich literature demonstrating that many psychiatric disorders and symptoms are strongly accompanied by deficits in attention as well as lack of effort. For example, attention deficits have been well established in depression (e.g. Paelecke-Habermann et al, 2005 JAD) and anxiety (Bishop, 2009 Nat Neurosci), the two main types of mental health symptoms measured in this study. Deficits in effort-based behaviors have also been recently documented in these disorders (see Treadway et al, 2012; Bishop and Gagne, 2018). Thus, there is a reason to believe that the "careless" subjects' carelessness might not be random and may be contributed by their underlying mental health issues. This possibility can actually be (at least partially) supported by the last analysis conducted by the authors, where they found that "careless" participants behaviorally showed less adaptive choice hysteresis and reduced outcome sensitivity (a finding that has been reported to be associated with depressive and other psychiatric symptoms; e.g. see Huys et al 2013 for a meta-analysis). Of course, the "ground truth" is unreachable in this case, since subjects are completely anonymous on these platforms (and the best way to validate the current set of analyses would be to conduct an in-person study where mental health status could be verified by trained clinicians). But based on the data available, the assumption that subjects who were flagged "careless /low effort" should be mentally healthier than their reported status may just not hold for the current study.

We thank the reviewer for this important point. We agree that the best way to validate our analyses would be to conduct an in-person study where mental health status could be verified by trained clinicians. In line with the reviewer's suggestion, we have now undertaken this additional study. The results of the new study (reported in a new section in the main text under the heading *"Individuals with and without diagnosed psychiatric disorders fail attention checks at similar rates"*) provide further evidence for the overall conclusions of our manuscript.

Briefly, in the additional study we embedded attention checks within the self-report measures of two ongoing studies of patients with major depressive disorder (total N=45 patients with diagnosis verified by a structured clinical interview, and N=20 controls whose lack of psy-

chopathology was also verified using the same interview). We did not find any evidence that patients with diagnosed major depressive disorder were more likely to fail attention checks than healthy controls, when completing the self-report symptom questionnaires online from their home (that is, under conditions as similar as possible to those in the main study we reported). Indeed, if anything, we found that non-psychiatric participants were more likely to be flagged by C/IE screening. Our results, though preliminary, suggest that rigorous C/IE screening is not likely to result in the differential exclusion of participants with elevated mental health symptoms.

The results of this additional study relate to a point that we originally made in the Discussion of our manuscript: if psychiatric participants would be more likely to fail attention checks in self-report surveys, this would entirely undermine a very large body of research in which self-report survey measures are considered a standard tool for psychiatric symptom measurement [1]. If it were truly the case that symptomatic participants were more likely to fail attention checks (i.e. because they are experiencing severe symptoms affecting their motivation or attention that render them unable to veridically complete a survey), then there would be little-to-no reason to use self-report symptom measures at all for these clinical populations. Reassuringly, our results suggest that this is not the case.

Of course, it may be the case that there are some psychiatric syndromes for which patients are unable to complete self-report surveys accurately. However, except for individuals experiencing acute mental health crises (e.g. mania, psychosis), we suggest that our results are in line with the common usage of self-report surveys in psychiatric research. As such, although we do not dispute that overcontrol bias is an important conceptual issue, we conclude that there are both empirical and theoretical reasons to believe that this issue does not undermine our results.

2. Lack of novelty: If the noise in the data is real, then the current findings become unsurprising because similar "spurious correlations" due to inadequate quality control have been well documented in other areas of science, including cognitive neuroscience (e.g. effect of poor motion control on imaging results, as shown in Friston et al 1996; Feller et al 2016 and many others). The presented data represents a specific manifestation of a general issue in experimental science and statistics, but is not something completely new.

We agree with the reviewer that our overarching point (i.e., that poor data quality controls can lead to spurious inferences) has been discussed, in the most general sense, elsewhere in science. Nevertheless, delineating specific instances of under-appreciated methodological biases can have major impact on methodological and empirical work within a field, as reflected in the reviewer's own suggested citations (as well as, e.g., [2, 3]). In addition, our manuscript is able, by very virtue of its specificity, to make actionable recommendations for researchers studying individual differences in cognitive processes (including in computational psychiatry). We would argue that these recommendations are a major novel contribution of our work that could not readily have been extracted from, for example, publications demonstrating motion-

induced artefacts in fMRI analyses. Indeed, since the publication of our initial manuscript on PsyArxiv, many researchers in computational psychiatry have changed their experiment designs to include infrequency items as we suggest (with the preprint already cited 24 times). If the need for this design feature were obvious from the extant literature, our paper should not have made such a difference in the practices in the field.

We would like to reiterate our case for the novelty of our findings. Though the perils of C/IE responding for individual-differences questionnaire research has been documented before [4, 5], we are unaware of any manuscript prior to ours addressing the risk of spurious correlations from C/IE responding for individual-differences cognitive science research (including computational psychiatry). To this point, our literature review revealed that, whereas the majority of behavioral researchers were applying quality controls to their behavioral data, only the minority of researchers were doing the same for their self-report data (and fewer still were following best-practices). This suggests that, although behavioral researchers were aware of the need for data quality assurance protocols in general, few were aware of the need for rigorous screening of their self-report data. We feel this suggests that the arguments and results we present are indeed novel for many individual-differences cognitive science researchers (including computational psychiatry researchers).

3. All analyses and conclusions were based on a single cognitive task. The lack of at least one more comparison task leaves the generalizability of the findings questionable. Although the chosen task is quite representative of value-based decision paradigms, it certainty does not represent the rich repertoire of cognitive processes and paradigms used in computational psychiatry research. Nevertheless, the authors made claims about how applicable the findings are to computational psychiatry in general.

We thank the reviewer for this suggestion. Indeed, our original conclusions rested on the findings from a single task. Therefore, in this revision, we have conducted a conceptual replication study where we repeated the entire experiment and analysis with a second task (the commonly-used, considerably more complex two-step task), as well as a second set of self-report measures (including personality scales), and more stringent quality control protocols (based on changes made by CloudResearch and Prolific to their respective platforms). The results of this additional study are reported in a new section in the main text under the heading *"Pattern of results generalizes to alternative tasks, self-report measures, and quality assurance protocols"*. The complete details of the methods and results for this study can be found in Appendix B of the supplementary materials. In brief, we successfully replicated the majority of findings from the original study in the second dataset. Of course, we have not confirmed our results across all task paradigms used in computational psychiatry, but we nevertheless hope that this new study goes some way towards establishing the robustness and generalizability of our original findings.

4. The claim about larger sample size further inflating false positive rates is based

on one particular index – the correlation between kappa (learning rate asymmetry) and 7-up. This particular correlation is small to start with. It is unclear if the sample size-amplified false positive rate finding is truly robust across effects with different sizes, or could be an effect size-dependent finding.

We thank the reviewer for this point. We wish to clarify two things. First, this result is not specific to the learning rate asymmetry and 7-down scale. As we show in a new supplementary figure (Appendix A, Figure S3), this finding generalizes to other pairs of variables.

Second, we note that this result reflects a more general point concerning statistical power. When an effect is truly non-zero, then an increase in sample size increases the power to detect that effect [6]. However, if a bias is present in the data (e.g. a non-zero correlation between behavioral and self-report measures, driven by the presence of C/IE responding), then increasing the sample size will also increase the probability of detecting it as a significant (spurious) effect [7]. To the reviewer's point, this is moderated by the true magnitude of the effect; larger effects require fewer observations for reliable detection (and vice-versa). If, in another dataset, the bias is larger than what was observed here (i.e. a stronger coupling between C/IE responding in task and self-report), then a smaller proportion of C/IE-responding participants would be necessary to induce a detectable spurious correlation. In sum, we suggest that there is little reason to suspect our results are effect-size dependent except insofar as they reflect general statistical principles.

Reviewer #2

This manuscript has multiple strengths. The writing is generally clear (though, I recommend keeping an eye on use of jargon and unnecessarily complex grammar. Writing could be more direct and I would encourage use of plain language). The figures are well done and informative. The study is technically well done. However, while the authors frame this research as significantly progressing the field, it is my view that the contribution is much narrower and I think this makes it a poor fit for Nature Human Behaviour.

1. It has been well-established elsewhere that inattentive responding leads to increased symptom report on psychiatric measures. It is also well-established that inattention impacts responding on behavioral measures; for example, one measure of inattentive responding is speeding through instruments. This paper extends this finding to speeding in non-survey measures. In particular, I recommend that you take a closer look at Curran (2016), who notes a correlation between speeding and other measures of inattention and (of relevance for the discussion section) provides alternative methods of using multiple measures to screen participants.

We thank the reviewer for this recommendation. We were indeed guided by the paper by Curran

(2016) in the design, analysis, and write-up of our manuscript. In line with this influence, the article by Curran (2016) article is cited at three separate points in our manuscript.

2. Essentially, the authors show that correlations between different measures of inattentiveness (i.e., unusually high symptom levels on psychiatric measures and performance on behavioral measures) should not be interpreted as meaningful. While this is certainly true, this is a very incremental point vs. a significant and novel contribution. In another point to narrowness, this spurious correlation should affect between-subjects measures only; did the authors look at outcomes that are within subjects and how might these be affected?

We respectfully disagree with the reviewer's view that the results here lack novelty (see also response to Reviewer 1, point #2). Though the perils of C/IE responding for individual-differences questionnaire research has been documented before [4, 5], we are unaware of any manuscript prior to ours addressing the risk of spurious correlations from C/IE responding for individual-differences cognitive science research (including computational psychiatry). To this point, our literature review revealed that, whereas the majority of behavioral researchers were applying quality controls to their behavioral data, only the minority of researchers were doing the same for their self-report data (and fewer still were following best practices). This suggests that, although behavioral researchers were aware of the need for data quality assurance protocols in general, few in this growing and impactful field were aware of the need for rigorous screening of their self-report data.

3. I recommend taking a look at Itay Sisso's tool to identify "bots" who may have completed the survey: http://mail.sjdm.org/pipermail/jdm-society/2018-November/008014.html **as the "bot" respondents would provide a gold-standard for higher versus lower quality workers.**

We thank the reviewer for pointing out this resource. We were aware of this tool before conducting our study. Unfortunately, Sisso's bot check tool requires the collection of participants' IP addresses, which is explicitly against Prolific's terms of service for researchers [8]. As such, we are unable to use this particular tool for detecting "bots" (or organizations of workers using private servers [9]). We should note, however, that we recruited participants through CloudResearch and Prolific, which both use IP addresses to monitor and filter out suspicious participants [10, 11]. Moreover, our custom experiment delivery software (NivTurk, https://nivlab.github.io/nivturk) has bot-checking functionality built into it. As such, we are confident that the results of the manuscript are not strongly affected by participants automatically completing the experiment. We have added additional text to the methods detailing the bot-checking functionality that our software employed:

> Participants were eligible if they resided in the United States or Canada; participants from MTurk were recruited with the aid of CloudResearch services [51]. (Note: This study was conducted prior to the introduction of CloudResearch's newest data

quality filters [52]). Following recent recommendations [53], MTurk workers were not excluded based on work approval rate or number of previous jobs approved. No other exclusion criteria were applied during recruitment. *It is important to note that both CloudResearch and Prolific use a number of tools (e.g. IP-address screening) to filter out the lowest quality participants. In addition, our custom experiment delivery software (NivTurk; see below) has bot-checking functionality built into it, and rejects from the start participants who are likely to not be human. We are therefore confident that our study is not strongly affected by participants using software to automatically complete the experiment.*

4. It is important to note that the authors deliberately let in lower quality workers and draw conclusions about what would be observed assuming no restrictions on worker quality or data cleaning. In practice, this combination of recruiting and data cleaning strategies is likely rare (and these are areas where detailed reporting are often lacking) and is not best practice.

We understand the reviewer's concern, and respectfully wish to clarify an inaccuracy in their appraisal of our methods. In our main analysis (in the section, *"Spurious symptom-behavior correlations produced by C/IE responding"*), we examine the consequences of multiple different types of data screening on correlations between behavioral and self-report measures. While it is true we examine a "no screening" condition (i.e. no exclusions are applied based on either self-report or behavioral screening measures, which our literature review corroborates as a rare strategy), we also examine conditions where we apply exclusions based on behavioral and/or self-report screening measures. Crucially, we found that screening only on behavioral screening measures (which is the most common screening strategy according to our literature review) is insufficient for ablating spurious correlations. In sum, we do evaluate data cleaning strategies common to the individual-differences cognitive science literature and find they are often lacking.

5. Screening participants out based on their response to self-report items or behavioral measures is one approach to managing quality concerns such as inattentiveness, but only an element of best practice if a researcher is looking to maximize sample size or perhaps sample diversity. This is one consideration among many and weeding participants out earlier, such as by IP address, is another approach that may help to circumvent these issues altogether. In practice, many researchers use multiple approaches in tandem.

We agree with the reviewer that some researchers use multiple approaches to screen for C/IE responding, and we also recommend this strategy in the manuscript (recommendation #1 in Box 1). However, we found in our literature review that the majority of researchers used only a single method to screen their data. As such, we believe our results and recommendations are relevant to many individual-difference researchers.

We also agree with the reviewer that there approaches to screening out participants even earlier during recruitment, such as through the use of IP addresses. We again note that, unfortunately, screening by IP address is not permitted at some commonly used online labor platforms (e.g., Prolific). Furthermore, we note that major online labor platforms perform this sort of screening as a service for researchers (including the two platforms studied here; i.e. CloudResearch, Prolific). Despite these data quality assurance protocols, however, we still found ample evidence of C/IE responding and risk for spurious correlations.

6. The recommendations the authors provide for best practice may not be used frequently in the field of computational psychiatry, but they, and other strategies, are well described by others (see in particular work by Jesse Chandler) and, as noted above, are not universally applicable depending on the goals of a research study. Redundancy with the existing literature and narrow applicability of recommendations further limits the scope and novelty of the current manuscript.

Taken together, this is a well designed study and well written manuscript. If it were modified to better include the existing literature and generally be more circumspect, it would likely be of interest to a narrower journal in computational psychiatry.

We respectfully disagree with the reviewer's suggestion that our recommendations are narrowly applicable. With the growing use of online platforms for data collection, the use of self-report surveys alongside online behavioural tasks is becoming more frequent not only in computational psychiatry, but also in fields as broad as cognitive psychology, behavioral economics, business studies, and psycholinguistics. The recommendations that we make in this manuscript complement those of Chandler and colleagues, but apply broadly to all of the fields above. Indeed, the breadth of the potential consequences of careless responding was our primary motivation for submitting this manuscript to a generalist journal such as *Nature Human Behavior*.

Reviewer #3

This paper addresses an important question for studies that use online testing to assess associations between cognitive processing and symptom questionnaire data, an approach that has increased dramatically (not least due to the COVID-19 pandemic) as online participant pools have been established.

The key finding is that such designs may produce spurious correlations – specifically, that participants who do not engage attentively in questionnaires, making random and/or "straight-lined" responses, will typically score higher than those who do attend, when symptoms are rare and population scores are therefore skewed. If those same participants also perform more noisily/poorly/differently on cognitive tests, then a spurious correlation between symptoms and cognitive

processing may be observed. This should not happen for questionnaires without skew. The authors suggest several recommendations to minimize these effects.

The writing is generally clear and the authors do a good job of explaining what they did and why. I applaud the authors on a timely and detailed piece of work, which is technically sophisticated and accompanied by clearly annotated open code and data. I do however have several concerns which reduced my enthusiasm for the manuscript, as there remains a lot of ambiguity as to what is driving these effects. These are listed immediately below, followed by some more minor points.

Major concerns

1) The most serious concern relates to the potential for selection bias to influence the results. Essentially, the key result of the study is that when participants who fail attention checks (C/IE) are removed from the analysis, this reduces some symptom-behaviour correlations. However, because of the nature of the questionnaires used in the study (i.e. depression, anxiety, and mania), performance on attention checks cannot be assumed to be orthogonal to the actual construct of interest. This is because difficulty concentrating is *part of the diagnosis* of all three of these syndromes. Consequently, the authors' approach may result in genuine associations being removed/attenuated, as a large section of the population of interest (e.g. those with genuine low mood, hypomania, anxiety) have been excluded from the data. To put it another way, these participants may be failing attention checks *precisely because* they are depressed, hypomanic or anxious, and therefore suffer concentration difficulties. This is analogous to the common statistical error of "controlling for" cognitive impairment in case-control studies of schizophrenia – by doing so, one removes important variance that is inherent to the construct of interest (Miller and Chapman, Journal of Abnormal Psychology 2001).

We thank the reviewer for this important point (see also Reviewer 1, point #1). We agree with the reviewer that the best way to validate our analyses would be to conduct an in-person study where mental health status could be verified by trained clinicians. Indeed, this is precisely what we have now done. Specifically, we embedded attention checks into the self-report measures of two ongoing studies of patients with major depressive disorder (total N=45 patients with diagnosis verified by a structured clinical interview, and N=20 controls whose lack of psychopathology was also verified using the same interview). The results of this study are reported in a new section in the main text under the heading *"Individuals with and without diagnosed psychiatric disorders fail attention checks at similar rates"*. The complete methodological and analytic details of the study can be found in Appendix C of the supplementary materials. In brief, we did not find any evidence that patients with diagnosed major depressive disorder were more likely to fail attention checks than healthy controls (if anything, non-psychiatric participants were more likely to be flagged by C/IE screening). Our

results, though preliminary, suggest that rigorous C/IE screening is unlikely to result in the differential exclusion of participants with elevated mental health symptoms.

A more convincing demonstration of the hypothesized spurious correlations would require data from a questionnaire that resulted in substantial skew due to rare endorsement, but where difficulty concentrating was *not* an inherent part of the construct in question. An example might be computer programming experience.

We thank the reviewer for this excellent suggestion, and agree that a stronger demonstration of our hypotheses would be to demonstrate the main result using a questionnaire where difficulty concentrating was not an inherent part of the construct in question. To do so, we conducted a conceptual replication study where we repeated the entire experiment and analysis with a new task (the two-step task), a second set of self-report measures, and more stringent quality control protocols (based on changes made by CloudResearch and Prolific to their respective platforms). Crucially, in this new study, we used self-report personality measures (i.e. HEXACO greed avoidance and artistic interests subscales) that are (a) not related to psychiatric symptoms, including difficulty concentrating, and (b) not hypothesized to have any relation to the behavior of interest (i.e. model-based choice on the two-step task). The results of this study are reported in a new section in the main text under the heading *"Pattern of results generalizes to alternative tasks, self-report measures, and quality assurance protocols"*. The complete methodological and analytic details of the study can be found in Appendix B of the supplementary materials. In brief, we observed a mean-shift in scores on the greed avoidance scale between attentive participants and participants flagged for C/IE responding. In the absence of screening, this translated into a significant, spurious correlation between greed avoidance and model-based planning behavior; this correlation was ablated when proper screening and exclusion protocols were applied. We hope that this new study demonstrates the robustness and generalizability of our original findings.

2) The authors do touch on the above point in their discussion. They suggest that any participant who fails even a single attention check must presumably not be responding veridically throughout the entire data collection procedure, thus that all of their data is untrustworthy, undermining the suitability of online testing to examine syndromes such as depression. I felt this discussion lacked nuance. In fact, it is likely that participants will adopt several strategies to answer self-report questionnaires (e.g. skimming the questions before starting), which may be incentivized by how studies are designed and rewarded, and indeed the nature of the questionnaire itself.

We agree that it is somewhat simplistic to assume that a participant who fails an attention check must therefore not be responding veridically for any part of the study. Nevertheless, we suggest that, given the potential for spurious correlations that we have documented, it is prudent to take a conservative approach to data collection by excluding these participants. Though it may be the case that a participant is inattentive on some questionnaires but attentive

on others, upon observing a failed attention check we cannot determine exactly how inattentive a participant has been. In this manuscript, we therefore suggest a conservative approach to experimental rigor in which any participant suspected of inattentive responding in any part of the experiment is excluded, precisely because we cannot know exactly how much of their dataset is likely to be contaminated.

Notwithstanding the above, one response to this legitimate concern is to take a graded approach to screening and excluding participants [12]. That is, participants could be screened with respect to a multitude of measures and only the consistently flagged participants be removed, thereby reducing the risk of inducing bias. Another possibility is to use sensitivity analysis as an alternative to exclusion, testing whether full-sample observed correlations are robust to the exclusion of participants flagged by measures of C/IE responding. We note these alternative possibilities in the discussion of our manuscript. For example, we write:

> Notwithstanding the above, one response to this legitimate concern is to take a graded approach to screening and excluding participants [41]. That is, participants could be screened with respect to a multitude of measures and only the consistently flagged participants be removed, thereby reducing the risk of inducing bias. Another possibility is to use sensitivity analysis as an alternative to exclusion, testing whether full-sample observed correlations are robust to the exclusion of participants flagged by measures of C/IE responding.

3) Other than a model-agnostic analysis examining trial-by-trial effects, the authors do not show to what extent C/IE participants and the rest of sample differ on behavioural measures (although they do this for questionnaire measures). If C/IE participants also differ on behavioural measures, it is likely that a statistical artifact, known as collider bias (also driven by selection bias), may occur when they are excluded, as participants who have *both* high depression (for example) and poor performance are being removed through this procedure. The effect of this would then be to either reduce real associations between depression and cognitive variables, or even induce counter-intuitive associations where none were present before.

We thank the reviewer for this very interesting point. We respectfully disagree, however, that collider bias poses a threat in this study. To illustrate our arguments, we visualize below two candidate causal models of the relationships between two latent variables, major depression and a tendency towards C/IE responding, and three observed indicators, self-report scores, task performance, and failures on infrequency items (Figure 1). Before we discuss each model, we note that a collider variable is defined as a variable that is affected by an independent (exposure) and dependent (outcome) variable [13, 14]. In turn, collider bias is defined as a spurious association between those two antecedent variables that is caused by controlling for a collider.

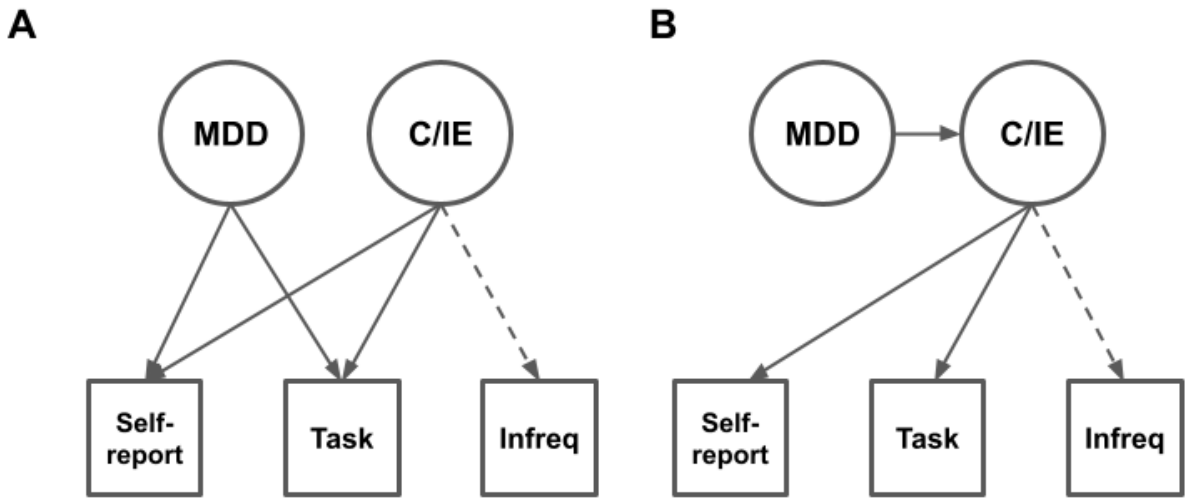

Figure 1: Directed acyclic graphs (DAG) depicting the hypothesized causal relations between latent variables (circles) and observed indicators (squares). (A) A causal model where major depression (MDD) and a tendency towards C/IE responding are independent common causes that affect self-report scores and task performance, but only C/IE tendencies affect the probability of failing infrequency-item checks. Here, excluding participants based on infrequency items (dotted line) controls only for C/IE tendencies, thereby unconfounding the correlation between task and self-report measures. (B) A causal model where major depression (MDD) is a common cause—mediated by a tendency towards C/IE responding—that affects self-report scores, task performance, and the probability of failing infrequency-item checks. Here, excluding participants based on infrequency items (dotted line) would be tantamount to controlling for depression, thereby attenuating correlations between task and self-report measures (i.e. overcontrol bias).

The first candidate model is where major depression (MDD) and a tendency towards C/IE responding are independent common causes that affect self-report scores and task performance, but where only C/IE tendencies affect the probability of failing infrequency-item checks (Figure 1A). To elaborate, latent depression straightforwardly affects self-report scores as individuals experiencing depression are more likely to endorse having symptoms; latent depression also affects task performance through changes in cognitive processing (e.g. lower reward sensitivity). In turn, a tendency towards C/IE responding affects self-report scores, task performance, and infrequency-item failure rates via the mechanisms proposed in this manuscript (e.g. random responding). In this model, excluding participants based on infrequency items controls only for C/IE tendencies, thereby unconfounding the correlation between task and self-report measures caused by latent depression.

The second candidate model, as suggested by the reviewer, is where major depression (MDD) is a common cause—mediated by a tendency towards C/IE responding—that affects self-report scores, task performance, and infrequency-item failure rates (Figure 1B). In this model, latent depression affects attentiveness and/or motivation, which in turn affects performance on all three outcome measures. (We note that this illustration omits possible direct paths between depression and self-report scores & task performance, caused by other facets of the disorder

[e.g. anhedonia]; this omission, however, does not affect our conclusions.) Here, excluding participants based on infrequency items would be tantamount to controlling for depression, thereby attenuating correlations between task and self-report measures (i.e. overcontrol bias).

We add three brief remarks. First, model 1 motivates much of our manuscript, whereas model 2 represents the reviewer's aforementioned hypothesis. Second, model 1 is more consistent with the results of our new patient study. That is, in this study, depression status is seemingly uncorrelated with the probability of failing attention checks and, as such, a tendency towards C/IE responding. Third, none of the outcome variables in either candidate model are collider variables. That is, none of the outcome variables, directly or indirectly, affect the latent variables (e.g. task performance does not change depression status). As such, we do not believe we are at risk for collider bias.

Therefore, it would be important to include a data table analogous to Table 2, but including the behavioural outcome measures shown in Figure 4. Apologies if this was already included, but I couldn't see it. This is particularly important because the example given in Figure 1 appears to depend on poorer task accuracy in the C/IE group, but we have no evidence presented in the paper that this is actually the case.

We thank the reviewer for this point. We agree such a table would be instructive for readers and have included it in the supplementary materials (see Appendix A, Table S2). As can be found there, we observed a number of significant differences in task performance between attentive and C/IE participants, including a significant decrease in accuracy in the C/IE group.

4) Interestingly, the study shows minimal evidence that C/IE participants are generally inattentive on tasks, as assessed for example through chance performance. The model-agnostic analysis presented in Figure 6 suggests that the groups are really quite similar in their task strategies, with no clear difference in reward/punishment sensitivity. A complement to the above point is that if C/IE are performing very similarly to the rest of the sample, it is hard to understand how their inclusion could induce spurious correlations.

While it is true that relatively few participants showed performance on the task at or below chance level, we do not necessarily agree that C/IE participants performed quite similarly. As shown in Table S2, there are quite a number of metrics on which C/IE participants were different than attentive participants (e.g. choice accuracy, proportion of win-stay and lose-shift choices, sensitivity to positive prediction errors). We hope that the inclusion of this table addresses the reviewer's concern by illustrating that C/IE participants differed in several important aspects of task behaviour compared with more attentive participants.

5) Despite the significant difference between the groups in Cronbach's alpha, it is still high (in some cases identical) in the C/IE sample for almost all questionnaires (other than BIS/BAS) which appears inconsistent with the suggestion that these

participants are making completely random responses. In any case, this is not a strong argument for the authors' position (as I understand it), as what they define as "C/IE" responding could equally involve random selection or "straight-lining" (i.e. choosing the same response for all items), which would produce either very low or high alpha, respectively.

Indeed, on average, the internal consistency of scale scores remains high among participants flagged for C/IE responding. We disagree, however, that this is inconsistent with our suggestion that these participants are responding without regard to the actual questions. Our rationale is based on a study by DeSimone et al. (2018) [15]. In Figure 2 of this paper, the authors show that, even when 20% of a sample is composed of truly random responding participants (i.e. approximately the same proportion of participants we flagged for C/IE responding), Cronbach's alpha decreases visibly but not dramatically. This suggests that we should not expect *colossal* reductions in internal consistency even with non-negligible numbers of C/IE participants. (In fact, Cronbach's alpha does not reach unacceptably low values [$\alpha < 0.8$] until more than half the sample is composed of C/IE participants.)

We do agree with the reviewer that C/IE responding, as we have defined it, could equally involve true random responding or straight-lining. As they note, these two response behaviors have opposite impacts on internal consistency [15]. Averaging across scales, however, we found that only approximately 10% of participants flagged for C/IE responding show evidence of straight-lining (i.e. choosing the same response for all items). Thus, we can expect that the majority of C/IE participants were engaging in heuristic response strategies closer to random responding, which should in turn lead to a decrease in internal consistency. Indeed, this is what we observed on average across scales. In sum, we believe this analysis is an effective sanity check to show C/IE participants are appropriately flagged as such.

Minor

Results

P2 – it isn't clear why this is specifically an issue for online research. Why would such spurious associations not arise for in-person testing?

We agree with the reviewer that this is not necessarily an issue specific to online research. We have removed the word "online" to make this point clearer. The text now reads:

> Here we wish to draw special attention to an underappreciated feature of ~~online~~ psychiatric research using self-report symptom surveys.

P2 – "complete the survey accurately". Since there is no objective truth for many such questionnaires, this is not an appropriate description – I suggest "attentively".

We agree with the reviewer and have modified the text as suggested. The sentence now reads:

> Because of the positive skew in the ground-truth symptom distribution, participants who respond carelessly to the symptom survey are more likely to report higher levels of symptom endorsement relative to participants who complete the survey ~~accurately~~ *attentively*.

P4 - it would be useful to clarify in the text why instructed-item checks are not useful, especially as avoiding them is listed as a recommendation in Box 1.

We agree with the reviewer that this would be a helpful clarification. We have expanded the text in our recommendations (Box 1) to include this point:

> When collecting self-report questionnaire data, include attention-check items that flag participants who may be engaging in C/IE responding. We recommend following best-practice guidelines in using infrequency-item checks rather than instructed-item checks, *as multiple studies have now shown that online participants are habituated to and circumvent the latter* [18-20]. Participants flagged by suspicious responses on attention-check items should either be excluded from further analysis, or assessed using sensitivity analyses to ensure that observed full-sample correlations are robust to their exclusion.

P5 - the hypothesis could be worded more clearly. The 'mean-shift' terminology could be complemented by explanation of what a mean-shift would result in (e.g. higher scores would be likely in C/IE participants responding at random to a right skewed questionnaire).

We thank the reviewer for this suggestion. We have updated the text, which now reads:

> We hypothesise that spurious behavior-symptom correlations may emerge due to a mean-shift in the average level of symptom endorsement in participants engaging in C/IE responding relative to attentive participants. In turn, a mean-shift is expected to occur when the overall rate of symptom endorsement is low; *that is, comparably higher scores are more likely for C/IE participants responding at random on a right skewed questionnaire.* In line with our predictions, the average level of symptom endorsement was noticeably exaggerated in C/IE-responding participants for the symptom measures where symptom scores were most positively-skewed (7-up, 7-down, GAD-7; see Figure 2).

P7 – it isn't really appropriate to term BIS/BAS "symptom" measurements, really these are personality constructs

We agree that the BIS/BAS constructs are better understood as personality measurements rather than symptom measures (their frequent use in clinical research notwithstanding). We have therefore updated our description of the questionnaires to reflect this:

> Prior to completing the reversal learning task, participants completed five self-report symptom *and personality-trait* measures.
>
> This translates to a tendency to endorse more severe symptoms on the 7-up/7-down and GAD-7 scales (where the rightmost options indicate greater frequency of symptoms) but less ~~severe~~ extreme symptoms *or personality traits* on the SHAPS and BIS (where the rightmost options indicate lower frequency of symptoms *or personality traits*) despite these inventories measuring strongly correlated constructs (i.e. depression and anhedonia, anxiety and behavioral inhibition).

P11 – Figure 4, it would be easier to understand the results if signed, not absolute, correlations were indicated (e.g. using both blue and red colours).

We thank the reviewer for this helpful suggestion. At the reviewer's request, we have included such a figure in the supplement (Appendix A, Figure S1). We elected not to make this change in the main text because we felt that a version of Figure 4 using a diverging colormap (e.g. red-blue) was more difficult to parse and detracted from our overall point about the absolute strength of associations.

P12 - it is not completely clear why the κ parameter was chosen for the bootstrapped correlation analyses. It would be interesting to see if the effects still hold for another potentially 'spuriously-correlated' parameter, such as learning rate or inverse temperature; and for a parameter that wasn't considered to have a spurious relationship and continued to be significant, such as the 7up and inverse temperature.

We thank the reviewer for this helpful suggestion. We agree that the choice of the κ parameter was not well-motivated. We have added an additional figure (Appendix A, Figure S2) to show that the results are not specific to κ and indeed generalize to other pairs of variables. Furthermore, we have amended the main text to reflect this change:

> Next, we investigated how spurious correlations depended on sample size. To do so, we performed a bootstrapping analysis where we held fixed the proportion of participants engaging in C/IE responding (i.e. 5%, 10%, 15%, 20%) and increased the total number of participants. Across all analyses, we measured the correlation and between the 7-down depression scale and learning-rate asymmetry (κ), which we previously identified as likely exhibiting a spurious association. *(The following results are not specific to learning-rate asymmetry and generalize to other model parameters; Figure S2).*

In addition, we have added a figure (Appendix A, Figure S3) showing the effects of sample size and proportion of C/IE participants on the true positive rate for a correlation that we expect

to be non-spurious for theoretical reasons (the correlation between learning rate asymmetry, κ, and GAD-7 scores). As can be observed, the true positive rate is essentially unaffected by the proportion of C/IE participants, supporting the hypothesis that this reflects a non-spurious association.

P13 - the authors should temper their discussion on spurious correlations and ensure they caveat that the false-positive rate did not depend only on sample size, but proportion of C/IE respondents in the population (e.g. at 5% the FPR actually seemed stable).

We agree with this suggestion and have tempered our language regarding the relationship between false-positive rate, sample size, and proportion of C/IE participants. The text now reads:

> Instead, our results suggest that, ~~holding the proportion of C/IE responding constant,~~ when there is significant C/IE responding, the false-positive rate for behavioral-symptom correlations will become increasingly inflated as the sample size increases.

P14 - the substantially greater perseveration shown by C/IE participants may indicate that they were pursuing a different strategy on the task which may have also been at play in their responses to questionnaires (e.g. "straight-lining"), and may therefore form a distinct sub-population who could be studied more.

We thank the reviewer for this interesting suggestion. We wish to clarify that C/IE participants do not appear to be more perseverative on average (Table S2). Instead, as the model-agnostic analyses (now moved to Appendix A) suggest, C/IE participants fail to maintain a reward-maximizing policy (e.g. stay with the previous choice following a reward).

Discussion

P17 - the authors argue that the elevated rate of symptoms in their sample compared to epidemiological estimates is evidence that some of these participants are inattentive. However, the authors should acknowledge that depression and anxiety symptoms may be more enriched in those doing online research – those who participate in these studies are not chosen at random from the population, but self-select.

We thank the reviewer for this important point. We have amended the discussion to note this:

> Notwithstanding the above, one response to this legitimate concern is to take a graded approach to screening and excluding participants [41]. That is, participants could be screened with respect to a multitude of measures and only the consistently flagged participants be removed, thereby reducing the risk of inducing

bias. Another possibility is to use sensitivity analysis as an alternative to exclusion, testing whether full-sample observed correlations are robust to the exclusion of participants flagged by measures of C/IE responding. We note that the strict screening approach used in the present study did not preclude us from identifying symptomatic participants or behavior-symptom correlations. Indeed, we found in our sample roughly 10% of participants endorsing symptoms consistent with clinical levels of depression, and approximately 20% consistent with clinical levels of acute anxiety. These estimates are within the realm of epidemiological norms [11, 30, 31]. *(We should note, however, that some studies have found elevated rates of psychiatric symptomology in online participants even after controlling for C/IE responding [13].)*

Methods

P22 - it would be useful for the reader to have more information about the narrative search. It isn't clear what permutations of the search terms were used, or why a more systematic search wasn't performed. It is also unclear how the categories of different checks were determined.

We thank the reviewer for this suggestion. We have added more detail concerning the methods of our literature review, specifically addressing why we could not perform a systematic review and where the full search terms and their permutations can be found. We have also clarified that the categories of different checks were determined based on previous taxonomies of screening methods (e.g. [16]).

To characterize common data screening practices in online computational psychiatry studies, we performed a narrative literature review [50]. We identified studies for inclusion through searches on Google Scholar using permutations of query terms related to online labour platforms (e.g. "mechanical turk", "prolific", "online"), experimental paradigms (e.g. "experiment", "cognitive control", "reinforcement learning"), and symptom measures (e.g. "psychiatry", "mental illness", "depression"). *We note that it was not feasible to conduct a systematic review, which requires the use of a publication database with reproducible search, because we required Google Scholar's full-text search in order to identify papers by recruitment method (e.g., Mechanical Turk).* We included in the review studies that (a) recruited participants online through a labour platform, (b) measured behavior on at least one experimental task, and (c) measured responses on at least one self-report symptom measure. Through this approach, we identified for inclusion 49 studies spanning 2015 through 2020. *The complete list of studies, and search terms used to find them, are included in the Github repository for this study.*

Two of the authors (S.Z., D.B.) then evaluated whether and how each of these

> studies performed data quality screening for both the collected task and self-report data. Specifically, we confirmed whether a study had performed a particular type of data screening, *with screening categories determined based on previous taxonomies of screening methods (e.g. [9]).* In addition, we assessed the total number of screening measures each study used and if and how monetary bonuses were paid to participants. This review was not meant to be systematic, but instead to provide a representative overview of common practices in online behavioral studies.

P22 - why is the N excluded for Prolific/mTurk participants who had done the study on the other platform not the same for each platform?

The difference in numbers of excluded participants (i.e., MTurk: N=19; Prolific: N=1) reflects on which platform the participant admitted to having already completed the experiment. That is, N=19 MTurk participants admitted to completing the experiment twice (presumably after previously completing the experiment first on Prolific). We agree this demarcation is confusing and extraneous, and have thus removed it from the text.

P22 - why did the authors not choose to select participants based on Prolific or mTurk scores? These should be reported in the paper. Do these scores correlate with C/IE responding?

If by "MTurk/Prolific scores" the reviewer is referring to platform-specific metrics like HIT Approval Rate, we did not use these as inclusion criteria because previous research has found they are not discriminating between low- and high-quality participants [17]. This is already noted in the methods under the "Sample" section:

> Following recent recommendations, MTurk workers were not excluded based on work approval rate or number of previous jobs approved [53].

Though we agree it would be interesting correlate these platform metrics with measures of C/IE responding, we are unable to do so with these data at this point. This is because our IRB requires us to discard the mapping between participants' platform IDs (i.e. MTurk worker ID, Prolific ID) and anonymized IDs after participants are paid. As such, we are presently unable to map participants' platform metrics to their anonymized data.

P28 - it would be useful to add a subheading to clarify that the measures being described relate to the bandit task.

We thank the reviewer for this helpful suggestion. We have added subheadings to better distinguish the self-report screening measures from the task-based screening measures.

P28 - what the authors term "choice variability" does not seem to be a meaningful measure of inattentive responding on this task, as values close to either 1.00

(always choosing the same bandit) or 0.33 (always choosing a different bandit) could indicate inattention. The extent to which this is the case will of course vary depending on the precise reward history, which is why a computational approach is instructive. Equally, these extremes could be caused by low or high learning rates, and could even be adaptive.

We thank the reviewer for this insightful point. We agree with the reviewer that, in principle, extreme values in either direction on this metric could indicate inattentive behavior. As such, one might wonder if we should have used a bidirectional coding of this variable; that is, recode the variable such that values of 0.33 and 1.00 in the native scale have the same resulting value. In practice, however, we found that such an approach actually *reduces* the correlation between choice variability and the other screening measures (unidirectional: $\rho_\mu = 0.042$; bidirectional: $\rho_\mu = 0.027$), especially for the other task-based screening measures (unidirectional: $\rho_\mu = 0.097$; bidirectional: $\rho_\mu = 0.045$). This is likely because the vast majority of participants do not exhibit large choice perseveration (mean variability = 0.437, IQR = 0.381 – 0.467). As such, we elected not to include this suggested analysis in the revised manuscript, so as not to bias the correspondence analyses in support of our arguments (that is, to avoid artificially reducing the correlation between task and self-report screening measures).

Round 2 Reviews

Reviewer #1

I thank the authors for making the effort to provide new data and revise the manuscript per previous comments. In general, the authors did a great job with the new data they have. However, there are still major remaining issues that they did not address in this version of the manuscript.

1. My main concern about the novelty and significance of this work still remains (which is echoed by reviewer 2). The take-home message of the study — which was essentially that high quality data must be obtained to generate reliable results — was evident prior to the actual conduction of the study. In fact, many experimental areas, such as experimental psychology, neuroscience, and biomedical sciences have already proven that. For instance, in neuroimaging, spurious correlations come up due to motion [18, 19]; similar issues have been raised in genetics [20]. In the context of behavioral research, these attention checks are typically called "catch trials" or "sanity check". The practice of using these procedures to filter out noisy data has been going on for decades. Even in the context of using online subjects, there is now a very rich literature on this very same point (e.g., [21, 22] and many more).

We appreciate the reviewer's ongoing concerns about the novelty of our work, but we respectfully disagree. We agree with the reviewer that our "take-home message" (i.e., that poor data quality controls can lead to spurious inferences) has been discussed, in the most general sense, elsewhere in science. Nevertheless, this has not precluded previous studies that explicated particular methodological biases from having major impact on research within a field (e.g., [2, 3]). Moreover, we argue that a general understanding of the need for high-quality data is not sufficient to obviate our findings. By very virtue of its specificity, our manuscript is able to offer concrete recommendations for researchers studying individual differences in cognitive processes. Indeed, since the publication of our initial manuscript on PsyArxiv, many researchers in computational psychiatry have changed their experiment designs to include infrequency items as we suggest (with the preprint already cited 29 times). If the insights and results we present were obvious from the extant literature, then our paper should not have made such a difference in the practices in the field.

2. In the revised paper, the authors added data demonstrating that patients with MDD show comparable attention failure rate as HCs. However, one of the key points in my previous comment is that patients with known attention deficits, such as ADHD (and SCZ or addiction), are likely to show heightened failure rate. The conclusion of the paper should be limited to the conditions they've tested (e.g., depression).

We thank the reviewer for this point. We first wish to note that, in the previous round of reviews, the reviewer did not explicitly mention attention deficits in patients with ADHD, schizophrenia, and/or substance use issues; instead, they wrote:

> In fact, there is a very rich literature demonstrating that many psychiatric disorders and symptoms are strongly accompanied by deficits in attention as well as lack of effort. For example, attention deficits have been well established in depression (e.g. Paelecke-Habermann et al, 2005 JAD) and anxiety (Bishop, 2009 Nat Neurosci), the two main types of mental health symptoms measured in this study. Deficits in effort-based behaviors have also been recently documented in these disorders (see Treadway et al, 2012; Bishop and Gagne, 2018).

To address the reviewer's concerns, we therefore investigated the sensitivity of attention checks in patients diagnosed with major depressive disorder. While we agree that additional studies are necessary to validate attention checks in other patient populations, we respectfully disagree that we neglected to address the reviewer's stated concerns.

Second, we would like to reiterate our concerns with this line of reasoning. If psychiatric participants were in general more likely to fail attention checks in self-report instruments, this would mean their self-reports on such instruments are also unreliable, undermining a very large body of research in which these instruments are considered a standard tool for psychiatric

symptom measurement [1]. If it were truly the case that symptomatic participants were more likely to fail attention checks (i.e., because they are experiencing severe symptoms affecting their motivation or attention that render them unable to veridically complete a survey), then there would be little-to-no reason to use self-report symptom measures at all for these clinical populations. Of course, it may be the case that there are some acute psychiatric states for which patients are unable to complete self-report surveys accurately (e.g., mania, psychosis). Barring these, we suggest that our results are in line with the common usage of self-report surveys in psychiatric research. We mention this concern in the manuscript, in the Discussion:

> Experimenters should also carefully consider whether an online study is truly appropriate for the research question. In particular, if the project aims to study syndromes associated with considerable difficulty in task or survey engagement (e.g., severe ADHD, acute mania), symptomatic participants are likely to produce responses that cannot be distinguished from C/IE responding. In such a case, correlational research in online samples is likely not the best approach for the research question.

Regardless, we have also tempered our language describing the results of the clinical study in the Discussion. The text now reads:

> Though our final sample was small, we did not find evidence that *depressed* patients were more likely to fail attention checks than healthy controls (if anything, non-psychiatric participants were more likely to be flagged by C/IE screening). These results provide preliminary evidence that rigorous C/IE screening is unlikely to result in overcontrol bias. *However, further research with larger samples is necessary to validate attention checks in depressed and other patient populations.*

3. The authors also added a second task (two-step) to prove the point that their findings are generalizable across tasks. I applaud the authors for being able to replicate the findings with another task. But again, the two-step task is also an RL task as the original task (albeit more complex) and the study still falls short in terms of proving the findings are generalizable across experimental contexts.

We thank the reviewer for this point. We chose these tasks in particular due to their prevalence in computational psychiatry research, which makes heavy use of RL tasks. We agree that further study is required to better understand the scope of these issues. It is worth noting, however, that that similar results were reported (after the publication of our preprint) by an independent group using non-RL tasks and an alternative set of symptom measures [23]. Though more research is required, this suggests that our findings are likely generalizable beyond the context of RL tasks.

4. The overall attention failure is quite high for this study – in fact higher than many published studies [24]. This might imply a potential hidden factor which is boredom (if the main tasks are not perceived as interesting), that plays a role (even more important role) in this particular experimental context.

We thank the reviewer for this point. We wish to clarify three things. First, the quality of data collected from online labor platforms is always in flux (e.g., [25, 26]). That is, attention check failure rates on one platform at a single (distant) point in time (e.g., data collected on MTurk in 2013–2014; [24]) may not be an appropriate benchmark for data quality. Second, we do not believe our observed failure rates (i.e., 22% and 14% of participants failing one or more infrequency items in the original and replication studies) is abnormally high. Chandler et al. (2020) find evidence for C/IE responding in 18% of MTurk participants [27]. Similarly, Barends & de Vries (2019) find evidence of C/IE responding in 12-17% of MTurk participants [28]. In recent data quality research published by Prolific, attention check failure rates on Prolific and MTurk (using CloudResearch) was 31% and 53%, respectively [29]. Finally, our behavioral tasks were "gamified" and incorporated design elements previously used to sustain engagement in children [30, 31]. Of course, participants may still have found our tasks boring, but arguably then C/IE responding would be worse for the majority of online studies that do not utilize task "gamification".

Reviewer #2

I appreciate the authors' consideration and responses to my suggestions. I maintain that this is a technically well-executed and -written manuscript and am satisfied with their responses to my more minor points. However, my main concern with this paper was and remains that I find it to be too narrow in scope for Nature Human Behavior. The authors did not indicate that they made any substantive changes to the manuscript to address this point and while they make arguments for the paper's novelty, I ultimately continue to feel that this paper contributes an incremental step.

We thank the reviewer for taking the time to consider our revisions and for praising the technical execution of our work. We appreciate the reviewer's concerns about the scope of our work. To try to at least partially address this, we have noted in the Discussion that our results and conclusions are not applicable only to computational psychiatry research, but to any online individual-differences cognitive science research:

> We conclude with a list of concrete recommendations for future online studies involving correlations between task behavior and self-report instruments. *We note that these recommendations are not limited to computational psychiatry studies, but are applicable to any online individual-differences cognitive science research involv-*

ing similar methods (e.g., behavioral economics, psycholinguistics).

Reviewer #3

The authors have done a great job of responding to the reviewers' comments, and in particular the addition of the second online experiment provides convincing evidence that the attenuation of spurious associations is not driven by over-correction, which was my main concern. Unfortunately, in this second study the pattern of results with the artistic interest scale was difficult to interpret due to the negative skew on this measure, but the results relating to the greed avoidance scale, which is positively skewed like the symptom measures in the initial study, are very clear. I thank the authors for their work in conducting this additional study, which has greatly strengthened the paper.

There are two remaining substantial suggestions:

1. It seems a shame to relegate the results of the replication study to the Appendix, where some readers may never see them. At the very least Figures S7, S8 and especially S9, and Table S11, and possibly some of the supplementary text describing the results, could be promoted to the main manuscript to allow readers to compare between the two online studies more easily.

We thank the reviewer for this suggestion. While we sympathize with the concerns, our preference is to retain the material in Appendix B, as is, for two reasons. First, the main text is already quite long (over 7,000 words, not including the Methods or Figure/Table captions) and we worry about depleting our readers' attentions before they reach the Discussion (i.e., where our recommendations are). Second, most of the figures and tables from the replication study (e.g., Figures S7, S9; Table S11) require detailed explanation to be meaningfully interpreted. Promoting these to the main text would therefore necessitate adding much of the methods section from Appendix B alongside them, which we worry would counter-productively lengthen and clutter the main text. However, we would be willing to make this change if the editorial staff agree that having the material in the main text is essential, and allow us to extend the overall length of the article.

2. There are several discrepancies between the results of the two online studies, to the extent that these merit a specific paragraph in the Discussion section of the main paper. The main issues to cover are: i) The distributions were quite different for some variables in the initial and replication studies (in the second study there was much lower skew for GAD; and C/IE responders did not score significantly higher on anxiety or depression – which is a point that the authors should discuss, as part of the argument for how the spurious correlations arise is through an overall shift towards higher responding on common self-report mental

health scales). It's not clear what the explanation is for this, perhaps the authors could speculate. ii) In the second study, exclusion based on WSLS behaviour alone was effective at attenuating spurious correlations (which does somewhat contradict one of the authors' key take-home messages in the original submission!). Perhaps this is due to the more difficult task, which would provide more sensitivity for exclusions based on behaviour alone? What do the authors think? iii) The lower rate of C/IE responding in the second study (which the authors ascribe to new data quality filters on the platforms).

We thank the reviewer for this important point. To start, we wish to clarify one misunderstanding. In the replication study, we employed a 7-item measure of general anxiety from the HiTOP group; this is different from the GAD-7, which we used in the original study. This alternative anxiety scale was expected to yield an approximately symmetric total score distribution, and thus was hypothesized *not* to show differences in scores between attentive and C/IE responding participants. We note this in the methods section in Appendix B:

> Participants also completed an alternative 7-item measure of general anxiety symptoms over the last year (e.g., "I was overwhelmed by anxiety."; [13]). This scale is expected to elicit moderate rates of symptom endorsement, thereby resulting in a symmetric score distribution. We therefore expected the depression and mania measures to be at greater risk for spurious correlations with behavior on the two-step task than the anxiety measure.

As such, we do not believe it additionally noteworthy that the anxiety scores were less skewed than in the original study, or that participants flagged for C/IE responding did not score significantly higher on the anxiety scale. That scores on the depression scale (7-down) in the replication study were not more inflated in C/IE relative to attentive participants is perplexing, though this may simply reflect the overall smaller number of participants flagged for C/IE responding in the replication study.

We have attempted to address the other discrepancies in the Discussion:

> This study highlights the need for more research on the prevalence of C/IE responding in online samples and its interactions with task-symptom correlations. Many open questions remain, including under what conditions task- and symptom-screening measures might better correspond, what screening measures are most effective and when, and under what conditions spurious correlations are more likely to arise. *For example, we found that screening on task behavior alone was insufficient to prevent putatively spurious correlations for one task (reversal learning) but was sufficient for another task (the two-step task). This discrepancy may reflect differences in the tasks (e.g., the two-step task may be more challenging and thus more sensitive to C/IE responding) or differences in the screening measures (e.g.,*

choice accuracy across 90 trials may be a noisier measure than win-stay lose-shift choice behavior across 200 trials).

One especially pressing question is how sample size affects the likelihood of obtaining spurious correlations. The results of a bootstrapping analysis in our data suggest that false positive rates are likely to increase with sample size. As computational psychiatry studies move towards larger samples to characterize heterogeneity in symptoms (and to increase statistical power), it will be important to understand how sample size may exaggerate the effects of systematic error. *It will also be important to understand how this is moderated by overall C/IE responding rates, which we observed to vary across platforms and time, and which will presumably continue to evolve with changing labor platform and researcher screening practices.*

A few final minor suggestions:

3. In the replication study it's notable that one of the associations with depression that was attenuated when using exclusions based on C/IE only seems to return when using both types of exclusions, and was also evident using the WSLS only exclusion – this is the association between depression and sensitivity to reward on the preceding trial (Figure S9). This seems to be quite a plausible association, and it's not clear whether the authors are claiming that it is spurious or not. This should be clarified, as if the authors don't believe it's spurious (and personally I think it's real), presumably this would be a downside to excluding based on C/IE alone, as it's an example of a real association that is actually obscured by their recommended exclusion procedure. Reviewing Figure 4, a similar pattern is actually evident in the initial study, between anxiety and positive learning rate – again, this only becomes non-significant when excluding using C/IE. Are the authors claiming this is spurious, even though it is significant when using both types of exclusion?

We thank the reviewer for this interesting point, though we wish to clarify the correlation in question (i.e., between depression and sensitivity to reward). Before exclusions are applied, the correlation is $\rho = -0.104$, $p = 0.019$. Under the three exclusion strategies, the correlation is slightly attenuated (WSLS-only: $\rho = -0.097$, $p = 0.044$; infrequency-only: $\rho = -0.091$, $p = 0.058$; both: $\rho = -0.097$, $p = 0.049$). As is apparent, the effect is on the border of significance under all three exclusion strategies and just happens to be, likely by chance, non-significant in one case. We do not believe this is a meaningful difference; that is, in this instance, the difference between "significant" and "not significant" is not itself statistically significant [32]. As such, this situation does not present a clear downside to excluding based on C/IE responding on surveys alone. (We believe a similar explanation underlies the pattern of correlations between anxiety and positive learning rate in the original study.) We have added new supplementary tables (S14–S17) to facilitate understanding of this point.

4. The authors should temper their conclusions in relation to the study conducted

in depressed participants (Discussion, bottom of p17). While this is a useful addition to the manuscript, the finding of a null result in such a small sample is quite ambiguous, and indeed the Bayes factors in favour of the null are hardly convincing. Additionally, the context of online vs clinical recruitment is obviously very different, and patients referred from a clinic may well be more personally invested in the research and thus potentially more attentive for this reason.

We thank the reviewer for this point (see also Reviewer #1, point #2). We have tempered our language describing the results of the clinical study in the Discussion. The text now reads:

> Though our final sample was small, we did not find evidence that *depressed* patients were more likely to fail attention checks than healthy controls (if anything, non-psychiatric participants were more likely to be flagged by C/IE screening). These results provide preliminary evidence that rigorous C/IE screening is unlikely to result in overcontrol bias. *However, further research with larger samples is necessary to validate attention checks in depressed and other patient populations.*

Regarding the reviwer's last point, we wish to clarify that not all of the patients were recruited via clinician referral; many patients who agreed to participate were recruited via online ads (a process similar to how participants are recruited online from MTurk or Prolific). In any case, we agree that patient populations may be more invested in studies that investigate the conditions they are struggling with. This further suggests that our proposed methods of screening inattentive or careless participants is not likely to present an over-control bias and systematically exclude patients. We note this in the Discussion of Appendix C:

> Indeed, whereas healthy controls may be primarily motivated to participate for monetary purposes, patients may be motivated to participate to further scientific research that may ultimately benefit them (or others suffering from the same conditions). That is, patients may have more "stakes in the game," and may therefore be more motivated to provide higher-quality responses.

5. p58 of the supplement states "However, an unexpected and statistically significant mean-shift" – since this measure (greed avoidance) is somewhat skewed, why would the mean shift be unexpected? Surely the more surprising finding here is that the attentive and C/IE groups don't differ on depression score, which is clearly skewed.

We thank the reviewer for the question. To clarify, the mean-shift is unexpected because previous research involving large samples observed that the total score distribution for the greed avoidance scale was relatively symmetric. As such, we *a priori* expected minimal difference in these scores between attentive and C/IE responding participants, as we note in the methods:

Finally, participants completed a 6-item measure of greed avoidance, which measures attitudes towards wealth and status (e.g., "I am out for my own personal gain"; [14]). Based on previous studies, this scale is expected to elicit moderate rates of endorsement, thereby resulting in a symmetric score distribution. We therefore expected the artistic interests scale to be at greater risk for spurious correlations with behavior on the two-step task than the greed avoidance scale.

References

1. Demetriou, C., Ozer, B. U. & Essau, C. A. in *The Encyclopedia of Clinical Psychology* 1–6 (Jan. 2015).
2. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences* **113,** 7900–7905 (2016).
3. Bradley, V. C. *et al.* Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* **600,** 695–700 (2021).
4. Chandler, J., Sisso, I. & Shapiro, D. Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology* **129,** 49 (2020).
5. Arias, V. B., Garrido, L., Jenaro, C., Martinez-Molina, A. & Arias, B. A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods,* 1–17 (2020).
6. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 2013).
7. Meng, X.-L. 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,** 685–726 (2018).
8. *Prolific IDs, data collection and security* en. https://researcher-help.prolific.co/hc/en-gb/articles/360009377494-Prolific-IDs-data-collection-and-security. Accessed: 2022-11-9.
9. Dennis, S. A., Goodson, B. M. & Pearson, C. A. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting* **32,** 119–134 (2020).
10. *Enhancing Data Quality* en. https://go.cloudresearch.com/en/knowledge/enhancing-data-quality. Accessed: 2022-11-9.
11. *How does Prolific prevent duplicate participant accounts?* en. https://researcher-help.prolific.co/hc/en-gb/articles/360009092774-How-does-Prolific-prevent-duplicate-participant-accounts-. Accessed: 2022-11-9.
12. Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A. & King, K. M. Detecting random responders with infrequency scales using an error-balancing threshold. en. *Behav. Res. Methods* **50,** 1960–1970 (Oct. 2018).
13. Elwert, F. & Winship, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology* **40,** 31–53 (2014).
14. Wysocki, A. C., Lawson, K. M. & Rhemtulla, M. Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science* **5,** 25152459221095823 (2022).
15. DeSimone, J. A., DeSimone, A. J., Harms, P. & Wood, D. The differential impacts of two forms of insufficient effort responding. *Applied Psychology* **67,** 309–338 (2018).
16. Curran, P. G. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* **66,** 4–19 (2016).

17. Robinson, J., Rosenzweig, C., Moss, A. J. & Litman, L. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PloS one* **14,** e0226394 (2019).

18. Murphy, K., Birn, R. M. & Bandettini, P. A. Resting-state fMRI confounds and cleanup. *Neuroimage* **80,** 349–359 (2013).

19. Fellner, M.-C. *et al.* Spurious correlations in simultaneous EEG-fMRI driven by in-scanner movement. *Neuroimage* **133,** 354–366 (2016).

20. Sullivan, P. F. Spurious genetic associations. *Biological psychiatry* **61,** 1121–1126 (2007).

21. Barends, A. J. & De Vries, R. E. Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and individual differences* **143,** 84–89 (2019).

22. Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J. & Litman, L. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods* **51,** 2022–2038 (2019).

23. Sulik, J., Ross, R. M., Balzan, R. & McKay, R. Delusion-like Beliefs and Data Quality: Are Classic Cognitive Biases Artefacts of Carelessness? (2021).

24. Hauser, D. J. & Schwarz, N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* **48,** 400–407 (2016).

25. Moss, A. & Litman, L. *After the Bot Scare: Understanding What's Been Happening With Data Collection on MTurk and How to Stop It* https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/. Accessed: 2023-3-25. Sept. 2018.

26. Charalambides, N. *We recently went viral on TikTok - here's what we learned* en. https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned. Accessed: 2023-3-25. Aug. 2021.

27. Chandler, J., Sisso, I. & Shapiro, D. Participant carelessness and fraud: Consequences for clinical research and potential solutions. en. *J. Abnorm. Psychol.* **129,** 49–55 (Jan. 2020).

28. Barends, A. J. & de Vries, R. E. Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Pers. Individ. Dif.* **143,** 84–89 (June 2019).

29. Eyal, P., David, R., Andrew, G., Zak, E. & Ekaterina, D. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods,* 1–20 (2021).

30. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science* **27,** 848–858 (2016).

31. Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D. & Hartley, C. A. Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology* **6** (2020).

32. Gelman, A. & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* **60,** 328–331 (2006).