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

Abstract

A common research design in the field of computational psychiatry involves leveraging the power of online participant recruitment to assess correlations between behavior in cognitive tasks and the self-reported severity of psychiatric symptoms in large, diverse samples. Although large online samples have many advantages for psychiatric research, some potential pitfalls of this research design are not widely understood. Here we detail circumstances in which entirely spurious correlations may arise between task behavior and symptom severity as a result of inadequate screening of careless or low-effort responding on psychiatric symptom surveys. Specifically, since many psychiatric symptom surveys have asymmetric ground-truth score distributions in the general population, participants who respond carelessly on these surveys will show apparently elevated symptom levels. If these participants are similarly careless in their task performance, and are not excluded from analysis, this may result in a spurious association between greater symptom scores and worse behavioral task performance. Here, we demonstrate exactly this pattern of results in two independent samples of participants (total $N = 779$) recruited online to complete a self-report symptom battery and one of two common cognitive tasks. We show that many behavior-symptom correlations are entirely abolished when participants flagged for careless responding on surveys are excluded from analysis. We also show that exclusion based on task performance alone is often insufficient to prevent these spurious correlations. Of note, we demonstrate that false-positive rates for these spurious correlations *increase* with sample size, contrary to common assumptions. We offer guidance on how researchers using this general experimental design can guard against this issue in future research; in particular, we recommend the adoption of screening methods for self-report measures that are currently uncommon in this field.

1 Introduction

In recent years, online labour markets (e.g. Amazon Mechanical Turk, Prolific, CloudResearch) have become increasingly popular as a source of research participants in the behavioral sciences [1], in no small part due to the ease with which these services allow for recruitment of large, diverse samples. The advantages of online data collection have also begun to be recognized in psychiatric research [2], where this method offers several distinct advantages over traditional approaches to participant recruitment. The ability to assess psychiatric symptom severity in large general-population samples makes possible large-scale transdiagnostic analysis [3, 4], and facilitates recruitment from difficult-to-reach participant populations [5]. Online labour markets also facilitate re-recruitment, making them an attractive option for validating the psychometric properties of assessment tools [6] or studying clinical processes longitudinally [7].

With the advantages of online data collection also come specific drawbacks. Since participants recruited from online labour markets are typically completing experiments in their homes, they may be more likely to be distracted or multi-tasking during an experiment. They may also be more likely to use heuristic response strategies with the intention to minimise expenditure of time and cognitive effort (e.g., responding randomly on self-report surveys or behavioral tasks). Here, we will refer to such inattentive or low-effort behaviors as careless/insufficient effort (C/IE) responding [8, 9]. Among researchers using online labour markets, a common view is that poor-quality data resulting from C/IE responding can simply be treated as a source of unsystematic measurement error that can be overcome with increased sample sizes [3, 10]. Common practice in online behavioral research is to mitigate poor-quality data using the same screening methods that are typically used in in-person data collection (e.g., excluding participants who perform at- or below-chance on behavioral tasks). However, these methods may be specifically inappropriate for online psychiatry studies, as we detail below.

Here we wish to draw special attention to an underappreciated feature of psychiatric research using self-report symptom surveys. In such surveys, participants rate their endorsement of various psychiatric symptoms and, since most individuals in the general population tend to endorse no or few symptoms in many symptom domains, the resulting ground-truth symptom distributions tend to be heavily positively skewed [11, 12]. In this situation, the assumption that C/IE responding merely increases unsystematic measurement noise becomes untenable. 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 attentively [10, 13, 14]. Consequently, unless C/IE survey responses are carefully identified and removed, a considerable proportion of putatively symptomatic individuals in an online sample may, in fact, be participants who have not engaged with the experiment with sufficient attention or effort.

When participants complete both symptom surveys and behavioral tasks—a common study design in computational psychiatry—this artifact has the potential to induce spu-

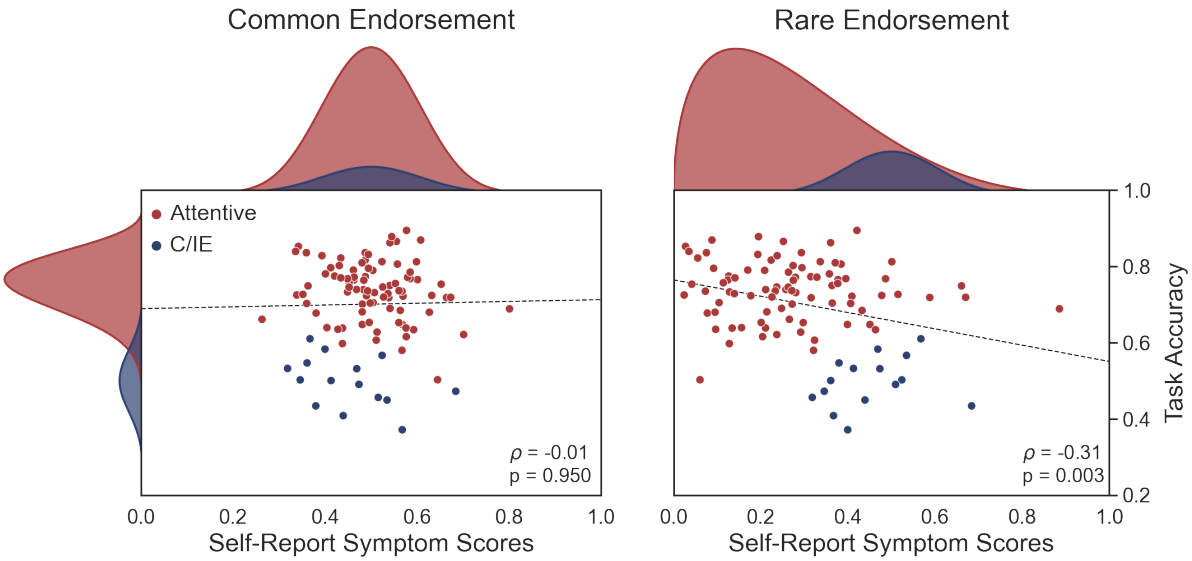

Figure 1: Simulated example of how spurious behavior-symptom correlations can arise when symptom endorsement is rare. *Left:* When symptoms are moderately common in the general population, C/IE respondents (blue) are indistinguishable from attentive participants (red) in self-report measures (x-axis, marginal distribution shown on top). Despite the lower task performance of C/IE respondents (y-axis), no correlation arises between symptom scores and task performance (dots are participants drawn from the shown distributions, with 15% C/IE participants; dashed line shows the (lack of) correlation.) *Right:* When symptoms are rare in the general population, careless respondents appear symptomatic in self-report measures. As a result, self-report symptom scores show a significant correlation with task performance.

rious correlations between symptom self-reports and task behavior. That is, while C/IE behavior is traditionally thought of as a source of noise that can result in type II (false negative) errors, here we suggest that in large-scale online psychiatric studies it can instead result in type I (false positive) errors. Concretely, if the same participants who engage in C/IE responding on surveys (and who therefore inaccurately report high levels of psychiatric symptoms) also respond with insufficient effort on behavioral tasks, this can cause experimenters to observe an entirely spurious correlation between greater symptom severity and worse task performance (see Figure 1). A similar effect has been well documented in personality psychology, where the presence of C/IE responding can induce correlations between questionnaires, and bias estimated factors in factor analysis [8, 10, 15–17].

Here, we demonstrate the real risk that C/IE responding can lead to spurious symptom-task correlations in computational psychiatry research. First, we asked to what extent recent studies in computational psychiatry screen participants based on self-report symptom data. We found that the majority of these studies did not screen participants' survey data at all, and that very few followed best-practice recommendations for survey data screening. We then asked whether behavioral screening alone was sufficient to identify participants engaging in C/IE responding on psychiatric symptom surveys. In

two new datasets from two separate online labour marketplaces, we found that screening based on task behavior fails to completely identify participants engaging in C/IE responding on surveys. Lastly, we investigated whether, under these circumstances, C/IE responding led to spurious correlations between symptom severity and task performance for positively-skewed symptom measures. Consistent with the logic set out above, we confirmed that failure to appropriately screen out C/IE survey responding in the proof-of-concept datasets that we collected would have produced a number of spurious correlations between task behavior and self-reported symptoms that are abolished when data are screened more thoroughly.

2 Results

2.1 Screening for C/IE responding is common for task behavior but not for self-report surveys

First, we sought to what extent recent online studies screen participants in a way that would reduce the risk of spurious correlations due to C/IE participants. We performed a narrative literature review of 49 online human behavioral studies, and evaluated whether and how each study performed task and self-report data screening (see Methods for details of the literature search).

Among studies that we reviewed, approximately 80% (39/49) used at least one method to identify C/IE responding in task behavior (see Table 1). Of these, just over half relied on a single screening method, with considerable heterogeneity in behavior screening methods across studies. Most common (46% of all studies) was identifying participants whose performance was statistically indistinguishable from chance-level on some measure of accuracy. Almost as common (38%) was screening based on low response variability (i.e., excluding participants who predominantly responded in the same fashion across trials, such as using only a single response key).

In contrast, only a minority (19/49, or 39%) of studies screened for C/IE responding in self-report symptom measures. The most common survey screening method was the use of attention checks, which are prompts for which most responses are unlikely given attentive responding. Participants who do not give the correct response to these prompts are therefore likely to be engaged in C/IE responding. Attention checks can be subdivided into instructed items (in which participants are explicitly told which response to select; e.g., 'Please select "Strongly Agree"'), and infrequency items (in which some responses are logically invalid or exceedingly improbable; e.g., endorsing 'Agree' for the question 'I competed in the 1917 Summer Olympic Games'). Of those studies that specified what type of attention check was used, instructed items were the most common method. As we discuss further below, this is notable because best-practice recommendations for data collection in personality psychology explicitly counsel *against* using instructed-item attention checks [18–20]. Only a handful of studies employed statistical or so-called unobtrusive

	Task Screening		Self-Report Screening	
Frequency	N=39 (80%)		N=19 (39%)	
Measure	Accuracy	18 (37%)	Attention Check	17 (35%)
	Variability	15 (31%)	Instructed	10 (20%)
	Response Time	7 (14%)	Infrequency	2 (4%)
	Comprehension Check	5 (10%)	Unspecified	5 (10%)
	Other	16 (33%)	Unobtrusive	4 (8%)

Table 1: The prevalence and types of task and self-report data screening practices in a sample (N=49) of recent online behavioral studies.

screening methods such as outlier detection or personal consistency.

In sum, whereas screening for C/IE responding in task behavior was relatively common for online behavioral studies, screening of self-report survey data was far less prevalent. Although this pattern may seem troubling, low rates of survey data screening are not necessarily an issue if screening on task behavior alone is sufficient to remove participants engaging in C/IE responding. That is, screening on survey data may be redundant if there is a high degree of correspondence between task- and survey-based screening methods.

In the next section, we explicitly test this hypothesis in a large sample of online participants completing a battery of self-report surveys and a behavioral task. Specifically, we measure the empirical correspondence between common task- and survey-based screening methods—as identified in our literature review—so that results are informative with respect to typical study designs in online psychiatry research.

2.2 Careless participants appear symptomatic when the overall level of symptom endorsement is low

To measure the correspondence of screening measures estimated from task and self-report behavior, we conducted an online behavioral experiment involving a simple decision-making task and a battery of commonly used self-report psychiatric symptom measures (see Methods). A final sample of 386 participants from the Amazon Mechanical Turk (N=186) and Prolific (N=200) online labour markets completed a probabilistic reversal-learning task and 5 self-report symptom measures. The reversal-learning task required participants to learn through trial-and-error which of three options yielded reward most often, and was modeled after similar tasks used to probe reinforcement-learning deficits in psychiatric disorders [21, 22]. The five self-report measures were the 7-up (which measures symptoms of hypomania), the 7-down (which measures symptoms of depression), the GAD-7, (which measures generalized anxiety symptoms), the BIS/BAS (which measures reward and punishment motivations), the SHAPS (which measures anhedonia symptoms), and the PSWQ (which measures worry symptoms), and were chosen based on previous literature to have a variety of expected response distributions (symmetric and

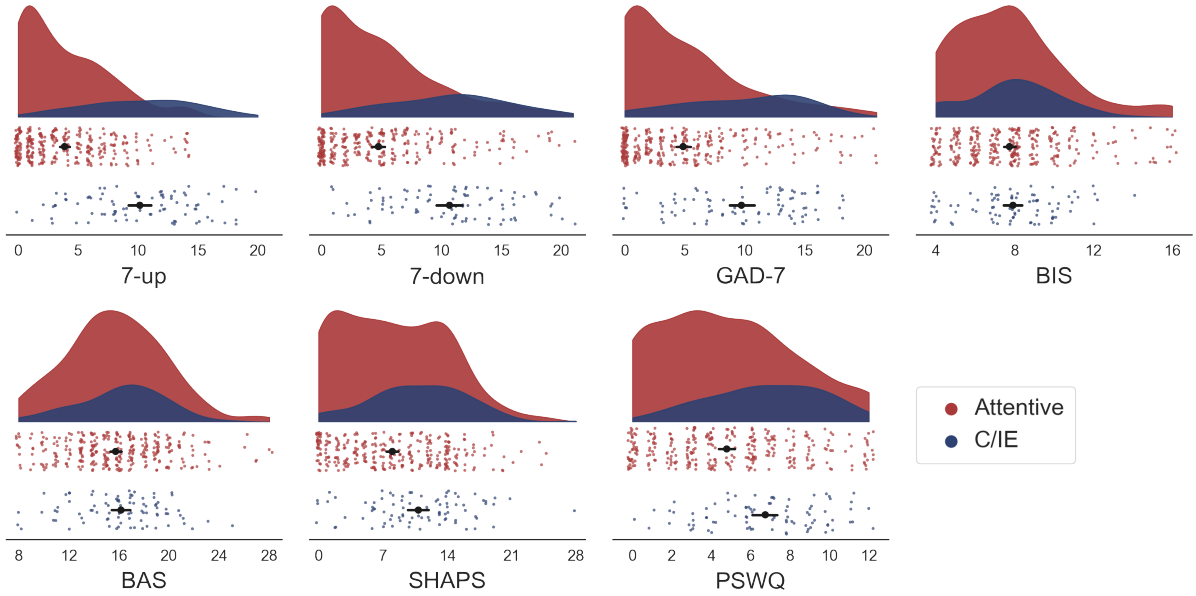

Figure 2: Raincloud plots of total symptom scores in attentive (red) and C/IE (blue) participants. Each colored dot represents the symptom score for one participant. Black circles: average score within each group (error bars denote 95% bootstrap confidence interval). Shaded plots: estimated distribution of responses for each group of participants. The scales are ordered approximately according to their estimated skew (see Table 2) from top-left (7-up) to bottom-right (PSWQ). The average level of symptom endorsement is most markedly different between groups in measures with lowest overall rates of endorsement.

asymmetric). In line with current best-practice recommendations in personality psychology [23], each self-report instrument included one 'infrequency' item that could be used to identify C/IE responses in survey data (see Methods for a list of infrequency items). The entire experiment (surveys and task) was designed to require 10 minutes on average to complete (observed mean = 10.28 minutes). To minimize any influence of fatigue on survey responding, participants completed the surveys prior to beginning the task.

To assess the overall quality of the data, we examined the number of participants flagged by the choice accuracy and infrequency item screening measures. Only 26 participants (7%) were flagged as exhibiting choice behavior at or below statistically chance levels in the reversal-learning task. In contrast, 85 participants (22%) endorsed a logically invalid or improbable response on one or more of the infrequency items when completing the self-report symptom measures. This discrepancy in the proportion of participants flagged by each method is consistent with previous research, which found varying levels of sensitivity to C/IE responding across screening methods [24]. The proportion of participants flagged for C/IE responding was marginally but significantly greater on Mechanical Turk compared to Prolific for both task (MTurk: N=18/186; Prolific: N=8/200; $z = 2.22, p = 0.026$) and survey data (MTurk: 50/186; Prolific: 35/200; $z = 2.22, p = 0.026$).

We hypothesise that spurious behavior-symptom correlations may emerge due to a mean-

		Total Score			Cronbach's α		% Clinical Cutoff	
Subscale	Skew	Attentive	C/IE	t-value	Attentive	C/IE	Before	After
7-up	0.81	3.87	10.15	-13.31^*	0.84	0.84	13.0%	4.0%
7-down	0.76	4.75	10.68	-9.99^*	0.94	0.88	17.4%	9.3%
GAD-7	0.75	4.86	9.73	-7.88^*	0.92	0.87	25.9%	17.3%
BIS	0.78	7.74	7.92	-0.54	0.83	0.62	-	-
BAS	0.17	15.73	16.16	-0.91	0.84	0.71	-	-
SHAPS	0.26	8.02	10.85	-4.04^*	0.90	0.81	17.9%	14.6%
PSWQ	0.19	4.78	6.74	-4.78^*	0.93	0.81	7.3%	7.0%

Table 2: Descriptive statistics of the self-report symptom measures between attentive and C/IE participants. Skew: the empirical skewness of the distribution of total symptom scores. Total score: the average symptom score across attentive and C/IE participants. Stars indicate statistical significance at $p < 0.05$. Cronbach's α: a measure of response consistency, where values closer to 1 indicate greater consistency in responses. % Clinical Cutoff: the percentage of participants reaching threshold for clinical symptomology before and after screening based on the infrequency measure. (The BIS/BAS scales do not have clinical thresholds.)

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). In contrast, where there was higher rates of symptom endorsement overall, the distributions of symptom scores between the two groups of participants were less noticeably distinct. Permutation testing confirmed that observed mean-shifts in symptom scores for C/IE participants were statistically significant for the majority of symptom measures (Table 2).

Hereafter, we use the infrequency-item method as a primary means of identifying C/IE responding in our data. To verify this approach, we conducted three validation analyses. The first analysis compared estimated internal consistency of self-report measures between the C/IE and attentive groups. The logic is that, if C/IE responding manifests as a tendency to respond randomly, we should expect to see a decrease in the consistency of a measure in the C/IE responding group [24–26]. In line with this reasoning, we observed a reduction in Cronbach's α in the C/IE group for the majority of survey instruments (Table 2). A permutation test confirmed that the average decrease in internal consistency across measures was greater than would be expected by chance given the difference in participant numbers between groups ($t = 3.69, p = 0.021$).

Second, we quantified the degree to which participants responded to self-report symptom

surveys in a stereotyped fashion; that is, we determined if participants exhibited patterns in their responses that were independent of the contents of the survey items. We fit a random-intercept item factor-analysis model [27] to self-report data (see Methods), and for each participant we estimated an intercept parameter that quantified their bias towards using responses on the left or right side of the response scale, regardless of what that response signifies for a particular self-report measure (e.g., low on one symptom scale versus high on another). We observed a credible difference between the average value of this intercept for the two groups ($\Delta\text{intercept} = 0.67, 95\%\ \text{HDI} = [-0.78, -0.55]$), such that C/IE participants were biased towards using the right-half of survey response options. 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 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).

Finally, we compared the proportion of participants meeting the cutoff for clinical levels of psychopathology before and after excluding participants based on their responses to the infrequency items. Previous studies have found that applying such measures reduced the prevalence of clinical symptomology in online samples towards ground truth rates from epidemiological studies [13]. On the most positively-skewed measures, the fraction of participants reaching clinical levels of symptom endorsement prior to screening was greater than what would be expected (Table 2). For example, 13.0% of participants scored at or above clinical thresholds for (hypo)mania on the 7-up scale in our sample prior to screening, compared with a 12-month prevalence of 5% in the general population [28, 29], but this rate was reduced to 4.0% (in line with the population prevalence estimates) after exclusion of C/IE respondents. We observed a similar pattern for both major depressive disorder and anxiety[1] (population prevalence estimates of 7% and 5% respectively; [11, 31, 32]). In line with previous research, we interpret these inflated rates of clinical symptomology in our sample prior to screening as suggestive of C/IE responding [13].

2.3 Low correspondence between task and self-report measures of C/IE responding

Next, we evaluated the degree of correspondence between behavioral and self-report screening measures in order to determine whether screening on behavior alone was sufficient to identify and remove careless participants. In line with the literature review, we computed multiple measures of C/IE responding from each participant's task behavior and survey responses (see Methods for description of measures). To measure the degree of correspondence between these behavioral and self-report screening measures,

[1]Interestingly, compared to previous literature the proportion of participants meeting threshold on the GAD-7 was elevated. We suspect this may reflect elevated rates of state anxiety during the COVID-19 pandemic [30], when these data were collected.

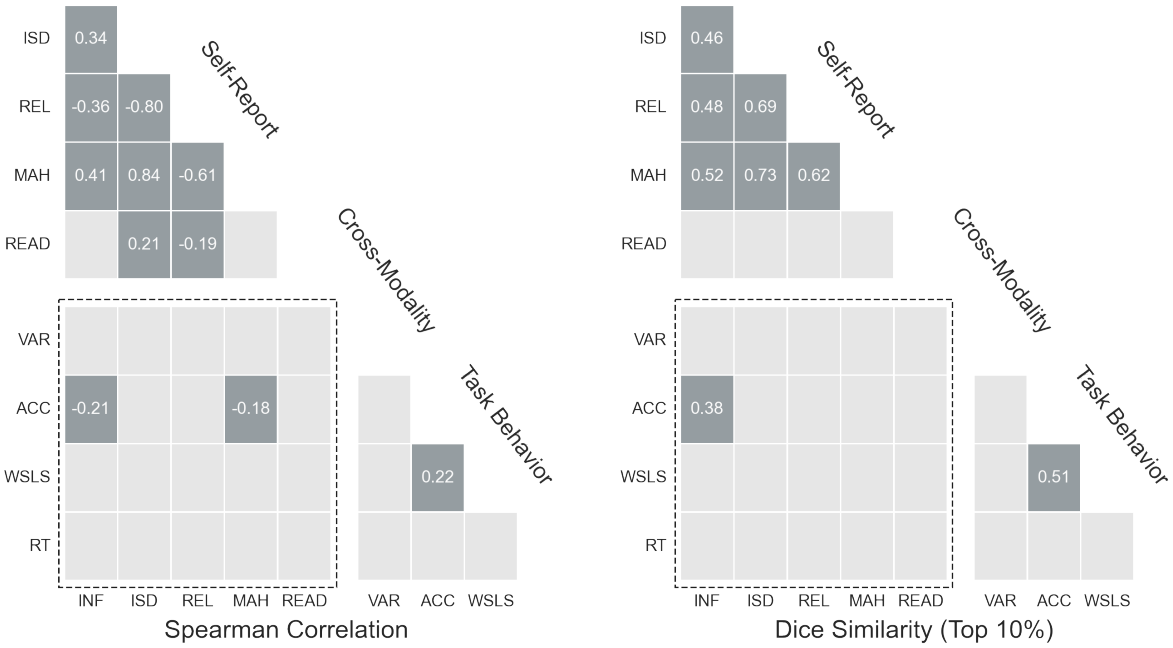

Figure 3: Similarity of task and self-report data screening measures. Each tile corresponds to the Spearman rank correlation (left) and Dice similarity coefficient (right) between two screening measures. Acronyms: INF = infrequency item; ISD = inter-item standard deviation; REL = personal reliability; MAH = Mahalanobis distance; READ = reading time; VAR = choice variability; ACC = choice accuracy; WSLS = win-stay lose-shift rate; RT = suspicious response times. Similarity scores have been thresholded after correcting for multiple comparisons. Numbers denote the strength of statistically significant correlations. Cross-modality correlations between task-behavior (left) and infrequency-item self-report measures (bottom) are in the dashed rectangle.

we performed two complementary analyses. First, we computed pairwise correlations on the unthresholded (continuous) measures using Spearman's rank correlation. The resulting pairwise similarity matrices are presented in Figure 3 (left panel). After correcting for multiple comparisons, there were few significant correlations between the behavioral and self-report screening measures. Only choice accuracy showed significant associations with any self-report measure (specifically, the infrequency and Mahalonobis distance measures). Crucially, the sizes of these observed correlations were roughly half those observed for the correlations between the self-report measures. This is worrisome as it suggests that, although there is some relationship between C/IE responding on tasks and self-report inventories, the relationship is not strong enough to ensure reliable detection of careless participants using task data alone.

Second, we used the Dice similarity coefficient to quantify agreement between different screening methods in the set of participants flagged for exclusion (Figure 3, right panel). This approach quantifies the degree of overlap between the set of would-be excluded participants based on different screening measures under a common exclusion rate. Though some measures have relatively clear threshold cutoffs (e.g., chance level performance for task accuracy), the majority of the measures evaluated here do not. As such, we evaluated the measures with respect to the top 10% of "suspect" participants flagged by

each measure, corresponding roughly to the fraction of participants having performed at chance levels on the reversal-learning task.[2] Results were largely consistent with the correlation analysis: few pairs of task and self-report screening measures achieved levels of agreement greater than what would be expected by chance. The only significant cross-modality pair identified — between the infrequency item and choice accuracy measures — has a similarly coefficient less than 0.4. In other words, when these two measures are used to identify the top 10% of participants most strongly suspected of C/IE responding, they agree on only two out of every five participants. Screening on choice accuracy alone (the most common method identified in our literature review) would fail to identify the majority of participants most likely engaging in C/IE responding as determined by the infrequency items.

Taken together, these results suggest that measures of C/IE responding in task and self-report data do not identify the same set of participants. This means that solely excluding participants on the basis of poor behavioral performance—the most common approach in online studies—is unlikely to identify participants who engage in C/IE responding on self-report surveys.

2.4 Spurious symptom-behavior correlations produced by C/IE responding

Here we examine the potential consequences of screening only on task behavior in our data. To do this, we estimated the pairwise correlations between the symptom scores of each of the self-report measures and several measures of performance on the reversal learning task. This analysis emulated a typical computational psychiatry analysis, in which the results of primary interest are the correlations between task behavior and self-reported psychiatric symptom severity.

For each participant, we computed both descriptive and computational-model-based measures of behavior on the reversal learning task (see Methods). To understand the effects of applying different forms of screening, we estimated the correlations between each unique pairing of a self-report symptom measure and measure of behavior under four different conditions: no screening, screening only on task behavior (i.e. only participants whose choice accuracy was above chance), screening only on self-report responses (i.e. only participants who responded correctly on all infrequency items), or both. The resulting pairwise behavior-symptom correlations following each screening procedure are presented in Figure 4. We note that we did not correct these correlation analyses for multiple comparisons, since our purpose was to demonstrate the extent of this issue across multiple behavioral measures and self-report symptoms. Any one of these correlations considered individually can be thought of as emulating a conventional analysis where fewer statistical tests would be performed.

[2]Results of the same analysis repeated for the top 25% of "suspicious" participants (corresponding roughly to the fraction of participants flagged by the infrequency-item measure) produced similar results, and are included in the supplement.

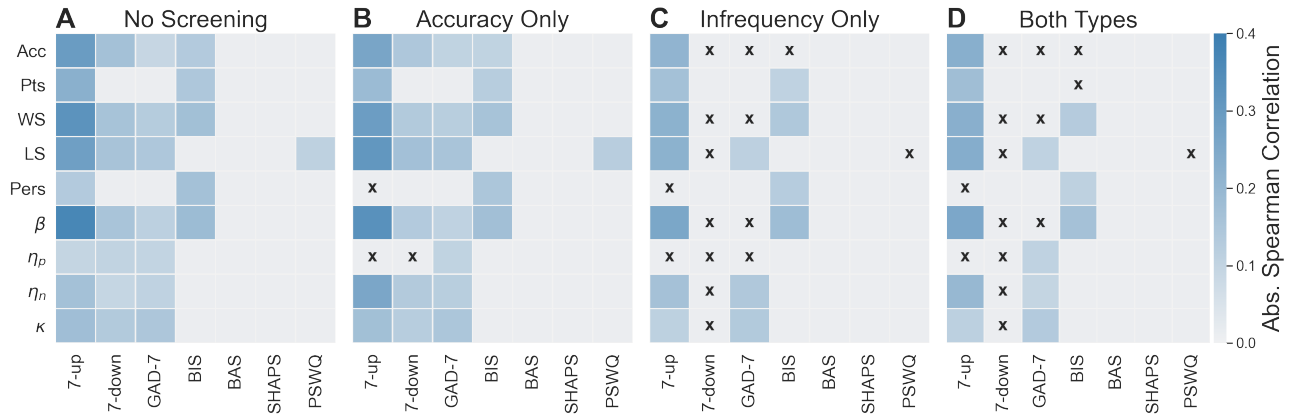

Figure 4: Absolute Spearman rank correlations between task behavior (y-axis) and symptom measures (x-axis) under different regimes of data screening and participant exclusions. Only statistically significant correlations are shown ($p < 0.05$ not corrected for multiple comparisons; all signed correlations are shown in Supplementary Figure S1). Black Xs indicate significant correlations ablated under screening. No Screening = no exclusions; Accuracy Only = exclusions based on chance-level performance in the reversal-learning task; Infrequency Only = exclusions based on invalid or improbable responses to infrequency items; Both Types = exclusions based on the previous two measures. Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry.

When no rejections based on C/IE responding was applied (i.e. all participants were included in the analysis; Figure 4A), many significant correlations emerged between measures of task behavior and symptom scores, in particular for 4 of the self-report instruments (7-up, which measures symptoms of hypomania; 7-down, which measures symptoms of depression; GAD-7, which measures generalized anxiety symptoms; and BIS, which measures tendencies related to behavioral inhibition). Consistent with our predictions, the majority of these correlations involved symptom measures with asymmetric total score distributions. Attending to only the most skewed measures (i.e. 7-up, 7-down, GAD-7), symptom endorsement was correlated with almost every behavioral measure. That is, significant correlations were not restricted only to general behavioral measures often used as proxies for participant effort (e.g. accuracy, inverse temperature β) but also to measures of specific theoretical interest, such as asymmetry of learning from positive and negative reward prediction errors (κ). Conversely, we found few significant correlations among symptom measures with more symmetric distributions. This is despite the fact these scales measure similar symptoms and syndromes (e.g. anxiety as measured by the GAD-7 and worry as measured by the PSWQ; depression as measured by the 7-down and anhedonia as measured by the SHAPS).

Next, we excluded participants from analysis based on task-behavior screening (i.e. choice accuracy, removing the 7% of participants exhibiting behavior indistinguishable from chance; Figure 4B). The pattern of correlations was largely unchanged: we again found many significant correlations between measures of behavior and asymmetric symptom

measures, but almost no significant correlations involving symmetric symptom measures. This suggests that rejection of participants based on the most common form of behavioral screening (i.e. performance accuracy) had little effect on behavior-symptom correlations as compared to no screening.

In stark contrast, when we rejected participants based on self-report screening (removing 22% of participants who endorsed one or more invalid or improbable responses on the infrequency items; Figure 4C), the number of significant correlations was markedly reduced, particularly for several of the most skewed symptom measures (7-down, GAD-7) and proxy measures of task attentiveness (e.g. accuracy, inverse temperature). This pattern of correlations was largely similar when rejections were applied based on both task and self-report screening measures (Figure 4D). We also note that with stricter screening, the remaining significant correlations were, for the most part (but not always), weaker.

These findings suggest that many of the significant behavior-symptom correlations observed without strict participant screening may indeed be spurious correlations driven by C/IE responding. Importantly, screening based on task behavior alone did not adequately protect against spurious symptom-behavior correlations in the presence of skewed distributions of symptom endorsement. For instance, consider the 7-down scale, a measure of trait depression: had we not screened participants based on infrequency items, we would have erroneously concluded that there were many significant associations between reversal-learning task performance and self-reported depression. Screening on self-report data allowed us to identify that each of these depression-behavior correlations was likely to be entirely spurious.

One possible objection to this interpretation is that the reduction in significant correlations following self-report screening was a result of the reduced sample size after removal of C/IE respondents (which comprised over 20% of the sample). To test this alternative hypothesis, we performed the same correlation analysis after removing random subsets of participants, fixing the sample size to that obtained after excluding C/IE respondents. In this case, the pattern of significant correlations was more similar to that before screening than after screening using the infrequency measure ($t = 262.49$, $p < 0.001$; Figure S2, compare to Figure 4A). Thus, the reduction in significant correlations following screening was unlikely to be driven solely by a reduction in statistical power.

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 pairs of variables; Figure S3).

The outputs of the bootstrapping analysis are presented in Figure 5. We found that, although estimated correlation magnitudes were independent of sample size (x-axis, left panel), the absolute magnitude of the behavior-symptom correlation increased with the

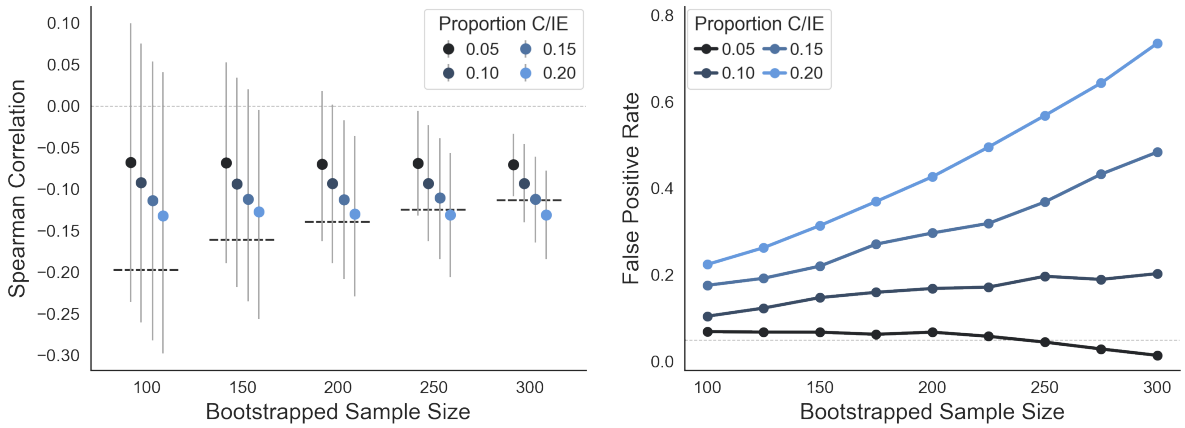

Figure 5: False positive rates for spurious correlations *increase* with sample size. *Left:* Bootstrapped Spearman correlations between learning rate asymmetry (κ) and depression scores (7-down) as a function of sample size and proportion of C/IE participants. The grey error bars indicate 95% bootstrap confidence intervals. The thick dashed lines indicate the threshold for statistical significance for the Spearman correlation at the corresponding sample size. Markers are jittered along the x-axis for legibility. *Right:* False positive rates for learning rate asymmetry (κ) and depression scores (7-down) as a function of sample size and proportion of C/IE participants. False positive rate was calculated as the proportion of bootstrap samples in which the correlation between κ and 7-down was statistically significant. The horizontal dotted line denotes the expected false positive rate at $\alpha = .05$.

proportion of C/IE participants (different coloured circles, left panel). Crucially, we found false-positive rates for spurious correlations *increased* with increases in sample size in our data for all but the smallest rates of C/IE responding (right panel). This runs counter to a common assumption that larger sample sizes are protective against spurious correlations because they serve to mitigate measurement error. Although this assumption is correct for unsystematic measurement error, it no longer holds in the regime of systematic measurement error (where larger sample sizes reduce the variance of estimates, but do not alter their bias). Instead, our results suggest that, except for low rates of C/IE responding, the false-positive rate for behavioral-symptom correlations will become increasingly inflated as the sample size increases.

2.5 Pattern of results generalizes to alternative tasks, self-report measures, and quality assurance protocols

One possible concern with the results presented so far is that they are specific to one instantiation of our experimental design. With more stringent quality assurance protocols during participant recruitment, or perhaps a different task or set of self-report measures, one might wonder if spurious correlations would remain such a threat.

To evaluate the generalizability of our findings, we therefore conducted a conceptual repli-

cation experiment in which an independent sample of N=393 participants (N=193 from MTurk using CloudResearch, N=200 from Prolific) completed a more difficult cognitive task, the well-known "two-step task" [33], and an alternate set of self-report measures (see Appendix B for details). Importantly, participants were recruited *after* CloudResearch and Prolific implemented new protocols to improve data quality on their respective platforms. As a final control measure, participants completed self-report symptom measures as before, but also personality measures with no hypothesized relationship to model-based planning behavior on the two-step task.

For the sake of brevity, we report here only the main pattern of findings (all results are reported in Appendix B). In the replication sample, 55 out of 393 participants (14%) endorsed a logically invalid or improbable response on one or more of the infrequency items when completing the self-report measures. This is roughly two-thirds of the number of participants who were flagged for C/IE responding in the original study, suggesting that the newer quality assurance protocols used by the online platforms are at least partially effective.

In the self-report symptom measures, we replicated the finding that total scores were noticeably exaggerated in participants suspected of C/IE responding, but only for symptom measures where overall rates of symptom endorsement were the lowest. Similarly, we again found that task-based screening and self-report screening measures showed low correspondence; that is, excluding participants on the basis of poor behavioral performance would not have identified and removed participants who engaged in C/IE responding on self-report surveys.

Finally, when we did not apply any exclusions, we observed spurious correlations between performance on the two-step task and total scores for both symptom and personality self-report measures with a mean-shift in scores between attentive participants and participants suspected of C/IE responding. In contrast with our original findings, however, we found that excluding participants based on self-report *or* task screening measures was sufficient to ablate these spurious correlations.

In sum, we replicated most of the main findings from the original study in an independent sample of participants completing a different task and other self-report measures. Although we found that screening on task behavior was sufficient to protect against spurious correlations in the replication sample, it is difficult to generalize and predict when or why this might be the case for other datasets. As such, we still believe that screening for C/IE responding in both task and self-report measures is the best approach to protect oneself against the possibility of spurious correlations.

2.6 Individuals with and without diagnosed psychiatric disorders fail attention checks at similar rates

One major concern with performing rigorous screening and exclusion of participants based on C/IE detection methods is that we might inadvertently introduce an overcontrol bias [34]. That is, to this point we have treated the tendency towards C/IE responding as independent from psychopathology. However, to the extent that C/IE responding reflects lack of motivation [35], avoidance of effort [36, 37], or more frequent lapses of attention [38, 39], one might hypothesise a true underlying association between psychopathology and careless responding in online studies. It is thus plausible that rigorous screening of C/IE responding might lead to the differential exclusion of truly symptomatic participants.

To explore this possibility, we embedded attention checks into the self-report measures of two studies of patients with major depressive disorder (see Appendix C for details). Specifically, N=45 psychiatric patients (confirmed to meet criteria for a diagnosis of major depressive disorder through a structured clinical interview) and N=20 healthy controls, all recruited through the Rutgers-Princeton Center for Computational Cognitive Neuropsychiatry (i.e., personal recruitment, not on online labour platforms), completed a series of self-report symptom measures, online, on their computers from the comfort of their homes. In total, 16 of 65 (24.6%) participants failed one or more attention checks. Subdivided by group, 6 of 20 (30%) healthy participants and 10 of 45 (22%) psychiatric participants were flagged for C/IE responding.

Using these data, we computed pairwise Bayes factors comparing three candidate models: attention check failure rates are equal between healthy and psychiatric participants (M1); failure rates are greater in psychiatric patients (M2); and failure rates are greater in healthy participants (M3). The model assuming equal rates of failure between healthy and psychiatric participants was 2.88 times more likely than the model assuming greater rates for psychiatric patients. In turn, the model assuming lower rates of failure for psychiatric patients was 1.27 times more likely than the model assuming equal rates. Finally, the model assuming lower rates of failure for psychiatric patients was 3.65 times more likely than the model assuming higher rates for psychiatric patients. Only the final comparison exceeds the cutoff value of 3, which is conventionally treated as the minimal amount of evidence required to treat a difference in model fit as meaningful. Although the size of the sample precludes any definitive conclusion, it is noteworthy that the model least consistent with the data was the one where psychiatric patients are more likely to fail infrequency item attention checks. These data suggest, therefore, that it is unlikely that individuals with high psychiatric symptom severity were disproportionately flagged for C/IE responding in the main analyses. Accordingly, we conclude that the screening measures we are suggesting are not likely to result in overcontrol bias and false-negative correlations between tasks and symptom measures.

3 Discussion

In this manuscript, we highlighted a particular set of circumstances, common in computational psychiatry research done on large online samples, in which spurious correlations may arise between task behavior and self-reported symptomology. When the ground-truth prevalence of a symptom is low in the general population, participants who respond carelessly on measures assessing this symptom may erroneously appear as symptomatic. A less careful pattern of responding on tasks used to measure cognitive constructs can then masquerade as a correlation between alteration in these constructs and symptom dimensions. We found repeated evidence for this pernicious pattern in participants recruited from two popular online labour platforms. False-positive rates for these spurious correlations *increased* with sample size, because the correlations are due to measurement bias, not measurement noise. Importantly, we found that screening on task behavior alone was insufficient to identify participants engaging in C/IE responding and avoid the false-positive correlations. Unfortunately, a literature review identified this type of screening as the most common practice in online computational psychiatry studies. We recommend instead to screen and exclude participants based on responding on surveys, a practice that abolished many spurious behavior-symptom correlations in our data.

One way of conceptualizing our results is through the lens of rational allocation of mental effort [40]. In any experiment, attentive responding is more effortful than careless responding. As such, participants completing an online task must perform a cost-benefit analysis—implicitly or otherwise—to decide how much effort to exert in responding. The variables that factor into such calculations are presumably manifold and likely include features of the experiment (e.g., task difficulty, monetary incentives), facets of the participant (e.g., subjective effort costs, intrinsic motivation, conscientiousness), and features of the online labour market itself (e.g., opportunity costs, repercussions for careless responding).

Viewed from the perspective of effort expenditure, our results suggest that participants appraised the cost/benefit trade-off differently for behavioral tasks and self-report surveys. Specifically, we found that only 7% of participants were at chance-level performance in our task, compared to more than 22% of participants who failed one or more attention-check items in the self-report surveys (a finding that qualitatively replicated in a second study involving a different task). Moreover, different measures of C/IE responding were weakly or not at all correlated between task behavior and self-report responses. This suggests the motivation for effortful responding was greater in the behavioral task, though precisely why is unclear. One possibility is that we gave participants a monetary incentive for attentive responding only during the task (a common practice, according to our literature review). A second possibility is that participants expected fewer consequences for C/IE responding during the self-report surveys, a reasonable assumption in light of how infrequently previous experiments have screened self-report data. Alternatively, participants may have found the gamified behavioral task more engaging or the self-report inventory more tedious. Regardless of the reason, this discrepancy reinforces our observations concerning the inadequacy of behavioral-task screening as a stand-alone method

for identifying C/IE responding. Since, in general, participants may appraise costs and benefits of effortful responding differently for behavioral tasks and self-report surveys, screening for C/IE responding on one data modality may in general be unsuitable for identifying it in the other. We therefore recommend screening on each component of an experiment.

One complicating factor for our argument is that C/IE responding may manifest in other ways than simply random responding for both behavioral tasks and self-report surveys. Indeed, there are more ways to respond carelessly than to respond attentively to a task or self-report inventory (e.g., random response selection, straight-lining, zig-zagging, acquiescence bias) [9]. The specific response strategy a participant adopts is likely to reflect the idiosyncratic integration of multiple perceived benefits (e.g. time saved, effort avoided) and costs (e.g. loss of performance bonuses, risk of detection and pay forfeited). As has been previously documented [24], the presence of multiple response strategies makes it clear why certain screening measures are more or less likely to correlate. For example, the inter-item standard deviation and personal reliability measures are both sensitive to statistically random responding, but less sensitive to straight-lining. Most importantly, a diversity of heuristic response strategies highlights the need for many screening measures of C/IE responding, each sensitive to different heuristic strategies.

Here we have focused on the potential for C/IE responding to result in spurious symptom-behavior correlations when rates of symptom endorsement are low, a case common to online computational psychiatry research. Beyond this, we should emphasize that a diversity of heuristic response strategies entails that there is more than one mechanism by which spurious correlations can emerge. To the extent that the only prerequisite is a mean-shift between attentive and careless participants, ours is not the only situation where one might expect spurious correlations to emerge [16]. For example, random responding on items with *high* base-rate endorsement could yield spurious correlations with precisely the opposite pattern observed here. Conversely, straight-lining may actually suppress correlations when symptom endorsement is low. In sum, without more understanding about the various types of heuristic responding and when each is likely to occur in a sample, it is difficult to predict *a priori* the patterns of systematic bias that may arise for a given study. This is further impetus for experimenters to be wary of C/IE responding and to use a variety of screening measures to detect it.

One objection to the rigorous screening and exclusion of participants based on C/IE detection methods is that we might inadvertently introduce an overcontrol bias. That is, to the extent that C/IE responding might reflect symptoms common to psychopathology (e.g. low motivation, effort avoidance, inattentiveness), rigorous screening of C/IE responding might lead to the differential exclusion of truly symptomatic participants. To explore this possibility, we embedded attention checks into the self-report measures of two studies of patients with major depressive disorder. Though our final sample was small, we did not find evidence for the proposition that 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.

Given that the results of our patient study are preliminary and warrant further investigation, researchers might still be wary of possible overcontrol bias. However, when using self-report questionnaires for screening, for overcontrol to seriously impact results it would have to be the case that symptomatic participants frequently endorse improbable or impossible responses to infrequency-item checks (e.g., responding 'Agree' to "I competed in the 1917 Olympic Games"). In this case, and even if such participants truly are experiencing severe symptoms of motivation or attention, there is likely to be limited utility in measuring these symptoms using a self-report measure that they are unable to complete veridically. A similar rationale underlies the widespread use of semi-structured interviews and other clinician-report measures rather than self-report measures for in-clinic psychiatric research. We would therefore argue that, if the psychiatric phenomenon being studied is such that this issue warrants concern, the research question may be better suited to an in-person study design involving participants in the clinic who meet full diagnostic criteria than a correlational design involving an online convenience sample.

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].) We also observed some positive correlations between anxiety and choice behavior that were consistent with effects found in previous literature [42–44]. For example, we found higher lose-shift rates and higher learning rates following negative prediction errors correlated with self-reported anxiety. This suggests that the screening methods we employed were not so aggressive as to attenuate behavior-symptom correlations that would be expected from the literature.

There are several notable limitations to this proof-of-concept study. We used a small set of screening measures, and did not employ other recommended procedures (e.g. logging each key/mouse interaction during survey administration to detect form-filling software or other forms of speeded responding [45]). Thus, we cannot be confident that all of the flagged participants were indeed engaging in C/IE responding; similarly, we cannot be certain that we correctly excluded all participants engaged in C/IE responding. We studied behavior-symptom correlations for only two tasks and two short self-report symptom batteries. It remains to be seen how generalizable our findings are, although our study

design was inspired by experiments prevalent in the online computational psychiatry literature. As suggested above, future studies may find greater correspondence between task and self-report screening measures for more difficult behavioral experiments. Finally, we should note that, unlike previous studies in which some participants were explicitly instructed to respond carelessly [45], we do not have access to "ground truth" regarding which participants were engaging in C/IE responding. Future work testing the efficacy of different screening metrics for identifying instructed C/IE responding may help to identify some of the issues that we have identified here.

The present study highlights the need for future work 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. 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.

4 Conclusions

Moving forward, we would strongly recommend that experimenters employ some form of self-report screening method, preferably one recommended by the best-practices literature (see Box 1 for a list of concrete recommendations). Our literature review found that, to date, the majority of online studies assessing behavior-symptom correlations have not used self-report screening, and our results demonstrate that stand-alone task-behavior screening is likely to be insufficient. Our results also demonstrate that inadequate screening is likely not merely to result in increased measurement error, as commonly assumed, but may also induce spurious correlations between behavioral task metrics and self-reported psychiatric symptom levels. For these reasons, we encourage experimenters to use a variety of data-quality checks for online studies and to be transparent in their reporting of how screening was conducted, how many participants are flagged under each measure, and what thresholds are used for rejection.

More broadly, we encourage experimenters in computational psychiatry to be mindful of the myriad reasons why participants may perform worse on a behavioral task. Wherever possible, researchers are encouraged to design experiments where the signature of some symptomology could not also be explained by C/IE responding (e.g. [46, 47]). Finally, we conclude by noting that it is preferable to prevent C/IE responding than to account for it after the fact [48]. As such, we recommend researchers take pains to ensure their experiments promote engagement, minimize fatigue and confusion, and reimburse participants fairly and ethically.

Box 1: Recommendations for future research

Here we offer several concrete recommendations for future research investigating symptom-behavior correlations in online samples.

- Use multiple screening methods to detect different types of C/IE responding. At a minimum, we recommend screening of both behavioral and self-report data. Within self-report data, we recommend using methods sensitive to multiple distinct patterns of C/IE responding (e.g., random responding, straight-lining, side bias) and, if possible, to log all page interactions (e.g., mouse clicks, keyboard presses).
- 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.
- We found that spurious correlations predominantly affected self-report instruments for which the expected distributions of symptom scores were asymmetric (either positively or negatively skewed). As such, symmetrically-distributed measures of a given construct should be preferred to asymmetrically-distributed measures (though this will often be infeasible given that the frequency of many psychiatric symptoms in the general population is itself positively skewed).
- Scales with reverse-coded items can be used to quantify the consistency of participants' responses between reverse-coded and non-reverse-coded measures of the same latent construct. With some care, this may be used to identify C/IE responding even for measures that do not include attention-check items [49]. Similarly, it may be beneficial to include multiple self-report surveys of the same construct to measure consistency across scales.
- In our experience, we have found it instructive to review discussions on public forums for participants of online labour markets (e.g. Reddit, TurkerNation). Doing so helps an experimenter identify what screening methods would-be participants are already aware of and prepared to answer correctly. (Several examples of workers discussing common attention checks can be found at the Github repository for this project.)
- Consider carefully whether the online methodology is truly appropriate for the research question. In particular, if the project studies 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 this case, correlational research with online samples is likely not the best methodology for the research question.

This is not an exhaustive list, and one overarching recommendation is that researchers

studying individual differences in psychiatric symptom endorsement should engage meaningfully with methodological research from the psychological measurement literature, where many of these questions have long been studied [9, 13, 16, 19, 24].

5 Methods

All code, data, and analysis materials are publicly available at https://github.com/nivlab/sciops.

5.1 Literature Review

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 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.

5.2 Experiment

5.2.1 Sample

409 total participants were recruited to participate in an online behavioral experiment in late June - early July, 2020. Specifically, 208 participants were recruited from Amazon Mechanical Turk (MTurk) and 201 participants were recruited from Prolific. This study was approved by the Institutional Review Board of Princeton University (#5291), and all participants provided informed consent. Total study duration was approximately 10 minutes per participant. Participants received monetary compensation for their time (rate USD $12/hr), plus an incentive-compatible bonus up to $0.25 based on task performance.

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.

The data from multiple participants who completed the experiment were excluded prior to analysis. Three participants (all MTurk) were excluded due to missing data. In addition, we excluded 20 participants who disclosed that they had also completed the experiment on the other platform. This left a final sample of N=386 participants (MTurk: N=186, Prolific: N=200) for analysis. The demographics of the sample split by labour market is provided in Table S1. Notably, the participants recruited from MTurk were older ($M = 7.7$ yrs, $t = 6.567, p < 0.001$) and comprised of fewer women ($z = 6.567, p = 0.011$).

5.2.2 Experimental Task

Participants performed a probabilistic reversal learning task, explicitly designed to be similar to previous computational psychiatry studies [21, 22]. On every trial of the task, participants were presented with three choice options and were required to choose one. After their choice, participants were presented with probabilistic feedback: a reward (1 point) or a non-reward (0 points). On any trial one choice option dominated the others. When chosen, the dominant option yielded reward with 80% probability; the subordinate options yielded reward with only 20% probability. The dominant option changed randomly to one of the two previously subordinate options every 15 trials. Participants completed 90 trials of the task (1 learning block, 5 reversal blocks).

As a cover story, the probabilistic reversal learning task was introduced to participants as a fishing game in which each choice option was a beach scene made distinguishable by a colored surfboard with unique symbol. Participants were told they were choosing which beach to fish at. Feedback was presented as either a fish (1 point) or trash (0 points). Participants were instructed to earn the most points possible by learning (through trial-and-error) and choosing the best choice option. Participants were also instructed that the best option could change during the task, but were not informed about how often or when this would occur (see the Supplement for the complete instructions). Prior to beginning the experiment, participants had to correctly answer four comprehension questions about the instructions. Failing to correctly answer all items forced the participant to start the instructions over.

The task was programmed in jsPsych [54] and distributed using custom web-application

software. The experiment code is available at https://github.com/nivlab/sciops, and the web-software is available at https://github.com/nivlab/nivturk. A playable demo of the task is available at https://nivlab.github.io/jspsych-demos/tasks/3arm/experiment.html.

5.2.3 Symptom Measures

Prior to completing the reversal learning task, participants completed five self-report symptom and personality-trait measures. The symptom measures were selected for inclusion based on their frequency in clinical research, and for having an expected mixture of symmetric and asymmetric score distributions.

Seven-Up/Seven-Down. The Seven-Up/Seven-Down (7u/7d; [55]) scale is a 14-item measure of lifetime propensity towards depressive and hypomanic symptoms. It is an abbreviation of the General Behavior Inventory [56], wherein only items that maximally discriminated between depression and mania were included. Items are scored on a 4-point scale from 0 ("Never or hardly ever") to 3 ("Very often or almost constantly"). Total symptom scores on both subscales range from 0 to 21, and are usually strongly right-skewed, with few participants exhibiting moderate to high levels of symptom endorsement.

Generalized Anxiety Disorder-7. the Generalized Anxiety Disorder-7 (GAD-7; [57]) is a 7-item measure of general anxiety. The GAD-7 assesses how much a respondent has been bothered by each of seven core anxiety symptoms over the last 2 weeks. Items are scored on a 4-point scale from 0 ("not at all") to 3 ("nearly every day"). Total scores on the GAD-7 range from 0 to 21, and are usually right-skewed, with few participants exhibiting moderate to high levels of symptom endorsement.

Behavioral Inhibition/Behavioral Activation Scales. the Behavioral Inhibition and Behavioral Activation Scales (BIS/BAS; [58]) are a measure of reward and punishment sensitivity. The original 42-item measure was recently abbreviated to a 14-item measure [59], which we use here. Items are scored on a 4-point scale from 1 ("very true for me") to 4 ("very false for me"). Total scores on the BAS subscale range from 8 to 32, whereas total scores on the BIS subscale range from 4 to 16. Previous reports have found total scores to be symmetrically distributed [60]. Importantly, in order to maintain presentation consistency with the other symptom measures, the order of the BIS/BAS response options was reversed during administration such that "very false for me" and "very true for me" were the left- and rightmost anchors, respectively.

Snaith-Hamilton Pleasure Scale. the Snaith-Hamilton Pleasure Scale is a 14-item measure of anhedonia [61]. Items are scored on a 4-point scale from 0 ("strongly agree") to 3 ("strongly disagree"), where higher scores indicate greater pathology. Total scores on the SHAPS range from 0 to 42, and have previously been found to be somewhat right-skewed [62, 63], with only the minority of participants exhibiting moderate to high levels of symptom endorsement. Importantly, as with the BIS/BAS, the order of the SHAPS response options was reversed during administration such that "strongly disagree" and

"strongly agree" were the left- and rightmost anchors, respectively.

Penn State Worry Questionnaire. the Penn State Worry Questionnaire is a measure of worry symptoms [64]. The original 16-item was recently abbreviated to a 3-item measure [65], which we use here. Items are scored on a 5-point scale from 0 ("not at all typical of me") to 4 ("very typical of me"), where higher scores indicate greater pathology. Total symptom scores range from 0 to 12 and are usually uniformly distributed.

5.3 Analysis

All statistical models fit as part of the analyses (described in detail below) were estimated within a Bayesian framework using Hamiltonian Monte Carlo as implemented in Stan (v2.26) [66]. For all models, four separate chains with randomised start values each took 2000 samples from the posterior. The first 1500 samples from each chain were discarded. As a result, 2000 post-warmup samples from the joint posterior were retained. Unless otherwise noted, the $\hat{R}$ values for all parameters was less than 1.1, indicating acceptable convergence between chains, and there were no divergent transitions in any chain.

5.3.1 Validation analyses

To validate the infrequency items as a sensitive measure of C/IE responding, we performed three complimentary analyses. We describe each in turn below.

Cronbach's α. We compared the average Cronbach's α, a measure of internal consistency, between attentive and C/IE participants. To control for the unbalanced numbers of participants in these groups, we performed a permutation test. First, we estimated Cronbach's α was estimated for each subscale and group. Next, we computed the average difference in Cronbach's α between the two groups. Then we created a null distribution for this statistic by repeating the same analysis but permuting group membership (i.e. randomly assigning participants to either group), holding fixed the sizes of both groups. This procedure was performed 5000 times. To compute a p-value, we tallied the number of null statistics equal to or (absolutely) greater than the observed test statistic.

Random intercept item factor analysis. We employed random intercept item factor analysis [27] to detect heuristic patterns of responding. In the model, the probability of observing response level k (of K total levels) from participant i on item j is defined as:

$$p(\, y_{ij} = k \,) = \begin{cases} 1 - \text{logit}^{-1}(\mu_i + x_j \cdot \theta_i - c_{j,1}) & \text{if } y = 1 \\ \text{logit}^{-1}(\mu_i + x_j \cdot \theta_i - c_{j,y-1}) - \text{logit}^{-1}(\mu_i + x_j \cdot \theta_i - c_{j,y}) & \text{if } 1 < y < K \\ \text{logit}^{-1}(\mu_i + x_j \cdot \theta_i - c_{j,K-1}) - 0 & \text{if } y = K \end{cases}$$

where μ_i is an intercept for participant i; θ_i is a vector of latent factor scores for participant i; x_j is a vector of factor loadings for item j; c_j is a vector of ordinal cutpoints for item j; and y_{ij} is the observed response for participant i on item j.

In this analysis, we did not estimate the factor loadings but instead treated them as observed. Specifically, we defined the factor loading for each item as a one-hot vector where the only nonzero entry denoted that item's corresponding subscale. That is, all of the items from a given subscale were assigned to their own unique factor (which was fixed to one). As such, the model estimated one factor score per participant and subscale (akin to the 1-parameter ordinal logistic model).

Crucially, each participant's responses were also predicted by a random intercept term, μ_i, which was not factor specific but instead was fit across all items. This intercept then reflects a participant's overall bias towards a response level. In our analysis, we coded the response levels such that the smallest value indicated endorsing the leftmost anchor (irrespective of semantic content) and the largest value indicated endorsing the rightmost anchor (irrespective of semantic content). Because the leftmost response option corresponds to symptomology on some scales (SHAPS), and a lack of symptomology for others (GAD-7, 7-up/7-down), we would not expect a consistent nonzero bias in this random intercept term for an attentive participant.

Clinical cutoffs. We compared the proportion of participants in our sample reaching the threshold for clinical symptomology before and after applying exclusions. For the GAD-7, previous research has suggested a clinical cutoff score of 10 or higher [11, 32]. Though the 7-up/7-down scales do not have firmly established clinical cutoffs recent work has suggested a cutoff score of 12 or higher [67], which we use here. Finally, the original authors of the SHAPS scale recommended as a cutoff a score of 3 or more when the items are binarized (1 = "strongly disagree", "agree"; 0 = "strongly agree", "agree"). We use this scoring approach in Table 2.

5.3.2 Correspondence of screening measures

To measure the correspondence of task- and self-report-based screening measures, we estimated a number of standard measures of data quality from each participant's task behavior (four in total) and self-report responses (five in total). Beginning first with the self-report data, we describe each below.

Self-report screening measure: Infrequency items. Infrequency items are questions for which all or virtually all attentive participants should provide the same response. We embedded four infrequency items across the self-report measures. Specifically, we used the following questions:

1. Over the last two weeks, how much time did you spend worrying about the 1977

Olympics? (Expected response: *Not at all*)

2. Have there been times of a couple days or more when you were able to stop breathing entirely (without the aid of medical equipment)? (Expected response: *Never or hardly ever*)
3. I would feel bad if a loved one unexpectedly died. (Expected response: *Somewhat true for me* or *Very true for me*)
4. I would be able to lift a 1 lb (0.5 kg) weight. (Expected response: *Agree* or *Strongly agree*)

Prior to conducting the study, the infrequency items were piloted on an independent sample of participants to ensure that they elicited one dominant response. In the main study, we measured the number of suspicious responses made by each participant to these questions. For thresholded analyses, participants were flagged if they responded incorrectly to one or more of these items.

Self-report screening measure: Inter-item standard deviation. The inter-item standard deviation (ISD) is an estimate of a participant's response consistency on a self-report measure [68], defined as:

$$ISD = \sqrt{\frac{\sum_{i=1}^{k}(y_j - \bar{y})^2}{k-1}}$$

where y_i is a participant's response to item i, x_i is a participant's average score across all items, and k is the total number of items for a self-report measure. A composite ISD measure was estimated per participant by summing across each of the seven self-report scales. Larger ISD values indicate lower response consistency.

Self-report screening measure: Personal reliability. The personal reliability coefficient is an estimate of a participant's response consistency on a self-report measure, estimated by correlating the average scores from split-halves of their responses. To avoid any item-order bias, a participant's personal reliability coefficient for a particular self-report measure was computed from the average correlation from 1000 random split-halves. A composite reliability measure was generated per participant by averaging across each of the seven self-report scales. Smaller reliability coefficients indicate lower response consistency.

Self-report screening measure: Mahalanobis D. The Mahalanobis distance is a multivariate outlier detection measure, which estimates how dissimilar a participant is relative to all others. For a participant i, the Mahalanobis D is defined as:

$$D = \sqrt{(X_i - \bar{X})^T \cdot \Sigma_{XX}^{-1} \cdot (X_i - \bar{X})^T}$$

where $(X_i - \bar{X})$ represents the vector of mean-centered item responses for participant i and Σ_{XX}^{-1} represents the inverted covariance matrix of all items. Greater Mahalanobis D values indicate larger deviations from the average pattern of responding.

Self-report screening measure: Reading time. The reading time is the total number of seconds spent filling out a particular self-report measure, adjusted for that measure's total number of items [13]. A total reading time estimate was estimated for each participant by summing across the adjusted time for each of the seven self-report measures. Shorter scores are indicative of less time having been spent on each item.

Task-based screening variable: Choice variability. Choice variability was defined as the fraction of trials of the most used response option per participant. Choice variability could range from 0.33 (all response options used equally) to 1.00 (only one response option used). Values closer to 1.00 are indicative of more careless responding during the task.

Task-based screening variable: Choice accuracy. Choice accuracy was defined as the fraction of choices of the reward-maximizing response option. For a task with 90 trials and three response options, a one-tailed binomial test at $\alpha = 0.05$ reveals chance-level performance to be 37 or fewer correct choices (41%). Lower accuracy values are indicative of more inattentive responding during the task.

Task-based screening variable: Win-Stay Lose-Shift. Win-stay lose-shift (WSLS) measures a participant's tendency to stay with a choice option following a reward versus shifting to a new choice option following a non-reward. WSLS thus measures a participant's sensitivity to reward feedback on the screen. WSLS was estimated per participant via regression, where the current choice (stay, switch) predicted by the previous trial's outcome (reward, non-reward) and a stationary intercept. Here we used the first (slope) term to represent a participant's WSLS tendency. Lower values of this term indicate less sensitivity to reward feedback and are thus indicative of more careless responding during the task.

Task-based screening variable: Response times. "Suspicious response time" was defined as the proportion of trials with an outlier response time, here measured as responses faster than 200ms. Greater proportions of outlier response times are indicative of more careless responding during the task.

Correspondence Analysis. We measured the correspondence of the above screening measures via two complimentary approaches. First, we computed pairwise correlations on the unthresholded (continuous) measures using Spearman's rank correlation. Second, we estimated the pairwise rate of agreement on the binarized measures using the Dice

similarity coefficient (looking at the top 10% and 25% most suspicious respondents for each measure). The former approach estimates two measures' monotonic association, whereas the latter approach estimates their agreement as to which participants were most likely engaging in C/IE responding. For significance testing, we used permutation testing wherein a null distribution of similarity scores (i.e. Spearman's correlation, Dice coefficient) was generated for each pair of screening measures by iteratively permuting participants' identities within measures and re-estimating the similarity. P-values were computed by comparing the observed score to its respective null distribution. We corrected for multiple comparisons using family-wise error rates [69].

5.3.3 Correlations between behavior and symptom measures

To quantify the effects of both task and self-report data screening on behavior-symptom correlations, we estimated the pairwise correlations between the symptom scores of each of the self-report measures and several measures of performance on the reversal learning task. For each participant, we computed both descriptive and model-based measures of behavior on the reversal learning task. We describe each in turn below.

Descriptive measures. Descriptive task measures included the following: accuracy (the fraction of choices of the reward-maximizing response option), points (the total number of points accumulated over the game), win-stay rates (the fraction of trials on which a participant repeated the previous trial's choice following a reward outcome), lose-shift rates (the fraction of trials on which a participant deviated from the previous trial's choice following a non-reward outcome), and perseveration (the number of trials on which a participant continued to choose the previously dominant response option following a reversal in task contingencies).

Model-based measures. The model-based measures were derived from a common reinforcement learning model of choice behavior, the risk-sensitive temporal difference learning model [70]. In this model, the expected value of a choice option, $Q(s)$, is learned through cycle of choice and reward feedback. Specifically, following a decision and reward feedback, the value of the chosen option is updated according to:

$$Q_{t+1}(s) = Q_t(s) + \eta \cdot \delta_t$$

where η is the learning rate bounded in the range $[0, 1]$ (controlling the extent to which value reflects the most recent outcomes) and δ is the reward prediction error, defined as:

$$\delta_t = r_t - Q_t(s)$$

In the risk-sensitive temporal difference learning model, there are separate learning rates for positive and negative prediction errors, such that positive and negative prediction

errors have asymmetric effects on learning. For example, the effect of negative prediction errors on learned values is larger than that of positive errors if $\eta_p+ < \eta_n$, and vice versa if $\eta_p > \eta_n$.

Finally, decision-making according to the model is dictated by a softmax choice rule:

$$p(y_t = s) = \frac{\exp\left(\beta \cdot Q(s)\right)}{\sum_i^S \exp\left(\beta \cdot Q(s)\right)}$$

where β is the inverse temperature, controlling a participant's sensitivity to the expected value of the choice options. In sum then, the model-based approach describes a participant's choice behavior as a function of three parameters (β, η_p, η_n).

We fit the reinforcement learning model to each participants' choice behavior using Stan (details above). Notably, 11 participants (3% of sample) had parameter estimates with poor convergence, i.e. $\hat{R} > 1.1$; their parameters were removed from the correlation analysis. Participants' parameters were fit individually (i.e. not hierarchically) so as to prevent bias during parameter estimation from partial-pooling between attentive and C/IE participants. Parameters were sampled using non-centred parameterisations (i.e., all parameters were sampled separately from a unit normal before being transformed to the appropriate range). Of note, the learning rates were estimated via an offset method such that $\eta_p = \eta + \kappa$ and $\eta_n = \eta - \kappa$, where κ is an offset parameter controlling the extent of an asymmetry between the two learning rates. This parameter was also entered into the behavior-symptom correlation analyses.

We confirmed the model adequately fit participants' choice behavior through a series of posterior checks (Figure S5). In particular, we confirmed the model recapitulated the group-average learning curves for each block of the experiment. Moreover, we confirmed that the model was able to recover reasonably well the choice accuracy for each participant.

The model-based measures included for analysis were: choice sensitivity (β, inverse temperature), positive learning rate (η_p), negative learning rate (η_n), and learning rate asymmetry ($\kappa = \frac{\eta_p - \eta_n}{\eta_p + \eta_n}$, normalized difference between η_p and η_n). We chose these measures as they have been previously used to assess performance in clinical samples [22, 42, 71, 72].

Correlation analysis. Behavior-symptom correlations (after various forms of screening and exclusion) were estimated using Spearman's rank correlation. Significance testing was performed using the percentile bootstrap method [73] so as to avoid making any parametric assumptions. These correlation analyses were not corrected for multiple comparisons, since our overarching purpose was to demonstrate the extent of this issue across multiple behavioral measures and self-report symptoms. Any one of these correlations considered individually can be thought of as emulating a conventional analysis where fewer statistical tests would be performed.

6 Acknowledgements

The authors are grateful to Agnes Norbury, Alexandra Pike, and Oliver Robinson for helpful discussion. The research reported in this manuscript was supported in part by the National Institute of Mental Health (NIMH) under award number 5R01MH119511-02, and by the National Center for Advancing Translational Sciences (NCATS), a component of the National Institute of Health (NIH), under award number UL1TR003017. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SZ was supported by an NSF Graduate Research Fellowship. DB was supported by an Early Career Fellowship from the Australian National Health and Medical Research Council (#1165010).

References

1. Stewart, N., Chandler, J. & Paolacci, G. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences* **21,** 736–748 (2017).

2. Chandler, J. & Shapiro, D. Conducting clinical research using crowdsourced convenience samples. *Annual review of clinical psychology* **12** (2016).

3. Gillan, C. M. & Daw, N. D. Taking psychiatry research online. *Neuron* **91,** 19–23 (2016).

4. Rutledge, R. B., Chekroud, A. M. & Huys, Q. J. Machine learning and big data in psychiatry: toward clinical applications. *Current opinion in neurobiology* **55,** 152–159 (2019).

5. Strickland, J. C. & Stoops, W. W. The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology* **27,** 1 (2019).

6. Enkavi, A. Z. *et al.* Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences* **116,** 5472–5477 (2019).

7. Kothe, E. & Ling, M. Retention of participants recruited to a one-year longitudinal study via Prolific (2019).

8. Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M. & DeShon, R. P. Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology* **27,** 99–114 (2012).

9. Curran, P. G. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* **66,** 4–19 (2016).

10. Chandler, J., Sisso, I. & Shapiro, D. Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology* **129,** 49 (2020).

11. Lowe, B. *et al.* Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical care,* 266–274 (2008).

12. Tomitaka, S. *et al.* Distributional patterns of item responses and total scores on the PHQ-9 in the general population: data from the National Health and Nutrition Examination Survey. *BMC psychiatry* **18,** 1–9 (2018).

13. Ophir, Y., Sisso, I., Asterhan, C. S., Tikochinski, R. & Reichart, R. The turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science* **8,** 65–83 (2020).

14. King, K. M., Kim, D. S. & McCabe, C. J. Random responses inflate statistical estimates in heavily skewed addictions data. *Drug and alcohol dependence* **183,** 102–110 (2018).

15. Robinson-Cimpian, J. P. Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher* **43,** 171–185 (2014).

16. Huang, J. L., Liu, M. & Bowling, N. A. Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology* **100,** 828 (2015).

17. 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).

18. 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).

19. Thomas, K. A. & Clifford, S. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* **77,** 184–197 (2017).

20. 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).

21. Waltz, J. A. & Gold, J. M. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. *Schizophrenia Research* **93,** 296–303 (2007).

22. Mukherjee, D., Filipwicz, A. L. S., Vo, K., Satterthwaite, T. D. & Kable, J. W. Reward and punishment reversal-learning in major depressive disorder. *Journal of Abnormal Psychology* **129,** 810–823 (2020).

23. Huang, J. L., Bowling, N. A., Liu, M. & Li, Y. Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology* **30,** 299–311 (2015).

24. DeSimone, J. A. & Harms, P. Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology* **33,** 559–577 (2018).

25. Maniaci, M. R. & Rogge, R. D. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* **48,** 61–83 (2014).

26. 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).

27. Maydeu-Olivares, A. & Coffman, D. L. Random intercept item factor analysis. *Psychological methods* **11,** 344 (2006).

28. Merikangas, K. R. *et al.* Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of general psychiatry* **64,** 543–552 (2007).

29. Merikangas, K. R. & Lamers, F. The 'true' prevalence of bipolar II disorder. *Current opinion in psychiatry* **25,** 19–23 (2012).

30. Yarrington, J. S. *et al.* Impact of the COVID-19 Pandemic on Mental Health among 157,213 Americans. *Journal of Affective Disorders* (2021).

31. Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M. & Wittchen, H.-U. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International journal of methods in psychiatric research* **21,** 169–184 (2012).

32. Hinz, A. *et al.* Psychometric evaluation of the Generalized Anxiety Disorder Screener GAD-7, based on a large German general population sample. *Journal of affective disorders* **210,** 338–344 (2017).

33. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69,** 1204–1215 (2011).

34. Elwert, F. & Winship, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology* **40,** 31–53 (2014).

35. Barch, D. M., Pagliaccio, D. & Luking, K. Mechanisms underlying motivational deficits in psychopathology: similarities and differences in depression and schizophrenia. *Behavioral neuroscience of motivation,* 411–449 (2015).

36. Cohen, R., Lohr, I., Paul, R. & Boland, R. Impairments of attention and effort among patients with major affective disorders. *The Journal of neuropsychiatry and clinical neurosciences* **13,** 385–395 (2001).

37. Culbreth, A., Westbrook, A. & Barch, D. Negative symptoms are associated with an increased subjective cost of cognitive effort. *Journal of abnormal psychology* **125,** 528 (2016).

38. Kane, M. J. *et al.* Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General* **145,** 1017 (2016).

39. Robison, M. K., Gath, K. I. & Unsworth, N. The neurotic wandering mind: An individual differences investigation of neuroticism, mind-wandering, and executive control. *The Quarterly Journal of Experimental Psychology* **70,** 649–663 (2017).

40. Kool, W. & Botvinick, M. Mental labour. *Nature human behaviour* **2,** 899–908 (2018).

41. 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).

42. Huang, H., Thompson, W. & Paulus, M. P. Computational dysfunctions in anxiety: Failure to differentiate signal from noise. *Biological psychiatry* **82,** 440–446 (2017).

43. Harlé, K. M., Guo, D., Zhang, S., Paulus, M. P. & Yu, A. J. Anhedonia and anxiety underlying depressive symptomatology have distinct effects on reward-based decision-making. *PloS one* **12,** e0186473 (2017).

44. Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L. & Sharot, T. Updating beliefs under perceived threat. *Journal of Neuroscience* **38,** 7901–7911 (2018).

45. Buchanan, E. M. & Scofield, J. E. Methods to detect low quality data and its implication for psychological research. *Behavior research methods* **50,** 2586–2596 (2018).

46. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood instability. *Nature communications* **6,** 1–10 (2015).

47. Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M. & Daw, N. D. Excessive deliberation in social anxiety. *bioRxiv,* 522433 (2019).

48. Ward, M. & Meade, A. W. Applying social psychology to prevent careless responding during online surveys. *Applied Psychology* **67,** 231–263 (2018).

49. Emons, W. H. Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement* **33,** 599–619 (2009).

50. Grant, M. J. & Booth, A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal* **26,** 91–108 (2009).

51. Litman, L., Robinson, J. & Abberbock, T. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* **49,** 433–442 (2017).

52. Litman, L. *New Solutions Dramatically Improve Research Data Quality on MTurk* https://www.cloudresearch.com/resources/blog/new-tools-improve-research-data-quality-mturk/. (Accessed: 2021-02-23).

53. 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).

54. De Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods* **47,** 1–12 (2015).

55. Youngstrom, E. A., Murray, G., Johnson, S. L. & Findling, R. L. The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment* **25,** 1377–1383 (2013).

56. Depue, R. A. *et al.* A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. *Journal of Abnormal Psychology* **90,** 381–437 (1981).

57. Spitzer, R. L., Kroenke, K., Williams, J. B. & Lowe, B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* **166,** 1092–1097 (2006).

58. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of Personality and Social Psychology* **67,** 319–333 (1994).

59. Pagliaccio, D. *et al.* Revising the BIS/BAS Scale to study development: measurement invariance and normative effects of age and sex from childhood through adulthood. *Psychological ssessment* **28,** 429–442 (2016).

60. Cooper, A., Gomez, R. & Aucote, H. The behavioural inhibition system and behavioural approach system (BIS/BAS) scales: Measurement and structural invariance across adults and adolescents. *Personality and individual differences* **43,** 295–305 (2007).

61. Snaith, R. *et al.* A scale for the assessment of hedonic tone: the Snaith–Hamilton Pleasure Scale. *The British Journal of Psychiatry* **167,** 99–103 (1995).

62. Franken, I. H., Rassin, E. & Muris, P. The assessment of anhedonia in clinical and non-clinical populations: further validation of the Snaith–Hamilton Pleasure Scale (SHAPS). *Journal of affective disorders* **99,** 83–89 (2007).

63. Leventhal, A. M. *et al.* Measuring anhedonia in adolescents: a psychometric analysis. *Journal of personality assessment* **97,** 506–514 (2015).

64. Meyer, T. J., Miller, M. L., Metzger, R. L. & Borkovec, T. D. Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy* **28,** 487–495 (1990).

65. Kertz, S. J., Lee, J. & Bjorgvinsson, T. Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychological Assessment* **26,** 1146–1154 (2014).

66. Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual* https://mc-stan.org.

67. Youngstrom, E. A., Perez Algorta, G., Youngstrom, J. K., Frazier, T. W. & Findling, R. L. Evaluating and Validating GBI Mania and Depression Short Forms for Self-Report of Mood Symptoms. *Journal of Clinical Child & Adolescent Psychology,* 1–17 (2020).

68. Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R. & Greenglass, E. The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences* **84,** 79–83 (2015).

69. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *Neuroimage* **92,** 381–397 (2014).

70. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience* **32,** 551–562 (2012).

71. Brolsma, S. C. *et al.* Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. *Psychological Medicine,* 1–11 (2020).

72. Ritschel, F. *et al.* Neural correlates of altered feedback learning in women recovered from anorexia nervosa. *Scientific reports* **7,** 1–10 (2017).

73. Wilcox, R. R. & Rousselet, G. A. A guide to robust statistical methods in neuroscience. *Current protocols in neuroscience* **82,** 8–42 (2018).

Appendix A

Signed correlation analysis

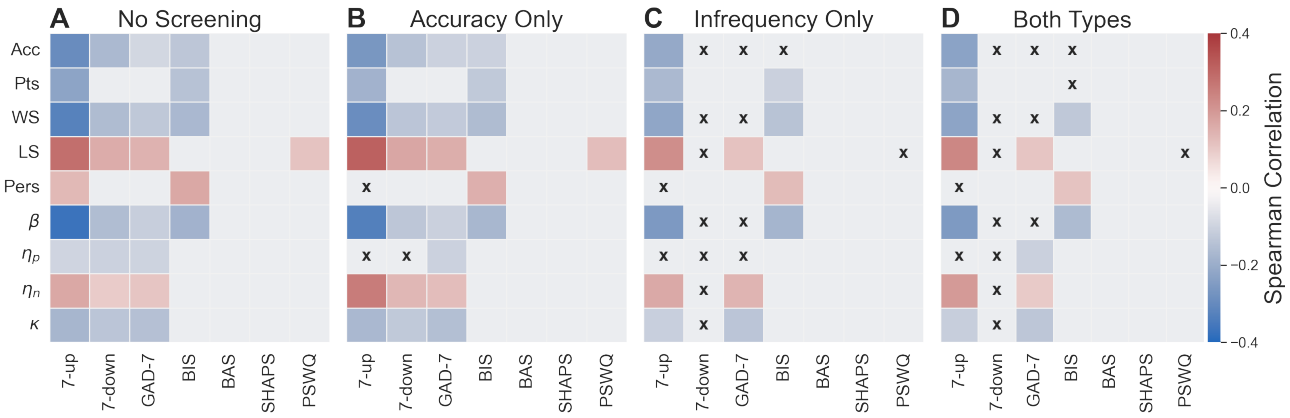

Figure S1: Signed Spearman rank correlations between task behavior (y-axis) and symptom measures (x-axis) under different regimes of data screening and participant exclusions. Only statistically significant correlations are shown ($p < 0.05$ not corrected for multiple comparisons). Black Xs indicate significant correlations ablated under screening. No Screening = no exclusions; Accuracy Only = exclusions based on chance-level performance in the reversal-learning task; Infrequency Only = exclusions based on invalid or improbable responses to infrequency items; Both Types = exclusions based on the previous two measures. Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry.

Bootstrapping Analysis

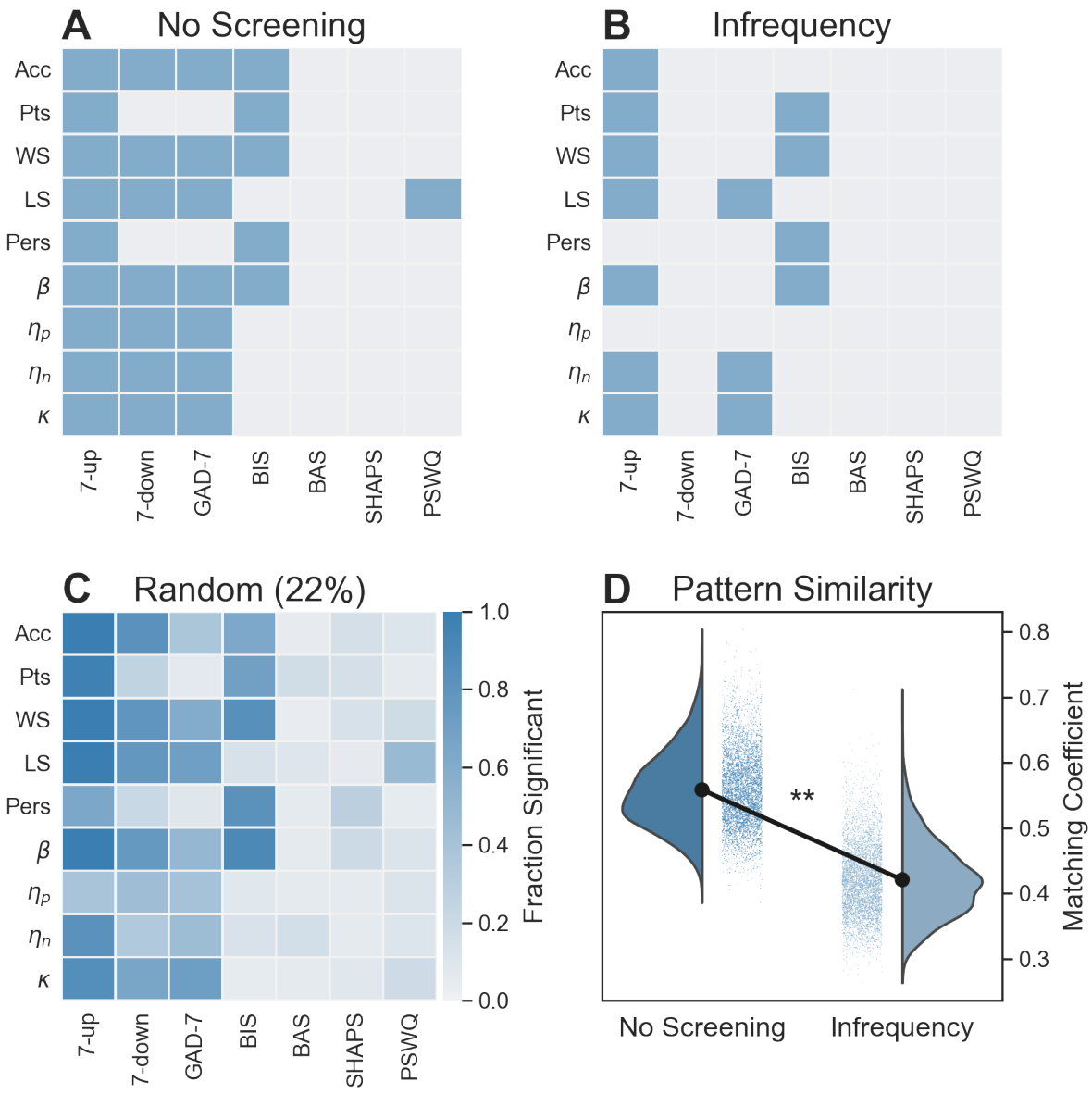

Figure S2: The pattern of significant behavior-symptom correlations before (panel A) and after (panel B) screening compared to the resulting pattern when random subsets of participants (22% of total, matched to screening using the infrequency measure) are removed (panel C). Panels A and B are reproduced from Figure 4 for convenience. Panel C: the fraction of significant correlations in 5000 bootstrapped samples. Panel D: The similarity of the pattern of correlations after removal of random subsets to that before and after screening using the infrequency measure. Similarity was calculated using the simple matching coefficient. Random removal subsets were significantly more similar to the "No Screening" than to the "Infrequency" screening datasets ($t = 262.49$, $p < 0.001$).

The relationship between sample size and false positive rates generalize to other sets of variables

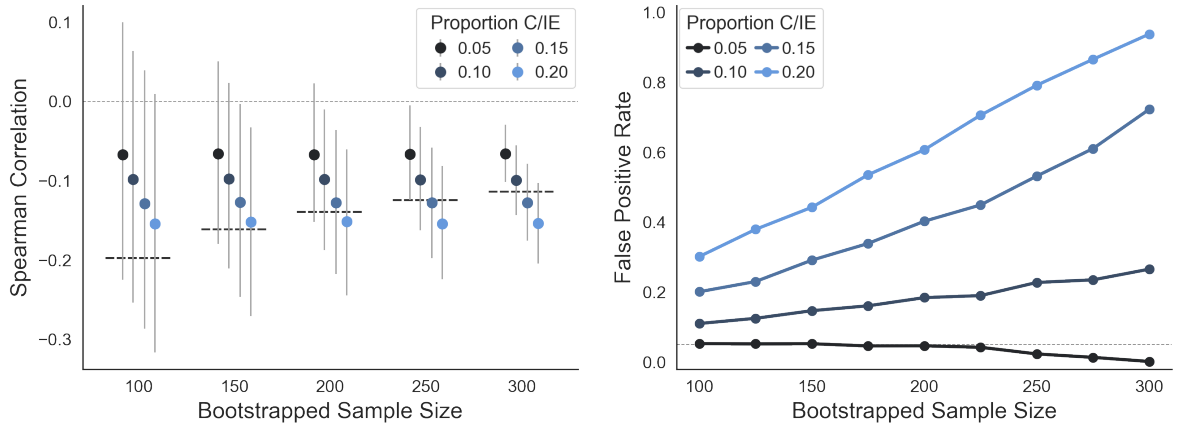

Figure S3: False positive rates for spurious correlations *increase* with sample size. *Left:* Bootstrapped Spearman correlations between inverse temperature (β) and depression scores (7-down) as a function of sample size and proportion of C/IE participants. The grey error bars indicate 95% bootstrap confidence intervals. The thick dashed lines indicate the threshold for statistical significance for the Spearman correlation at the corresponding sample size. Markers are jittered along the x-axis for legibility. *Right:* False positive rates for inverse temperature (β) and depression scores (7-down) as a function of sample size and proportion of C/IE participants. False positive rate was calculated as the proportion of bootstrap samples in which the correlation between β and 7-down was statistically significant. The horizontal dotted line denotes the expected false positive rate at $\alpha = .05$.

The relationship between sample size and true positive rates for true correlations

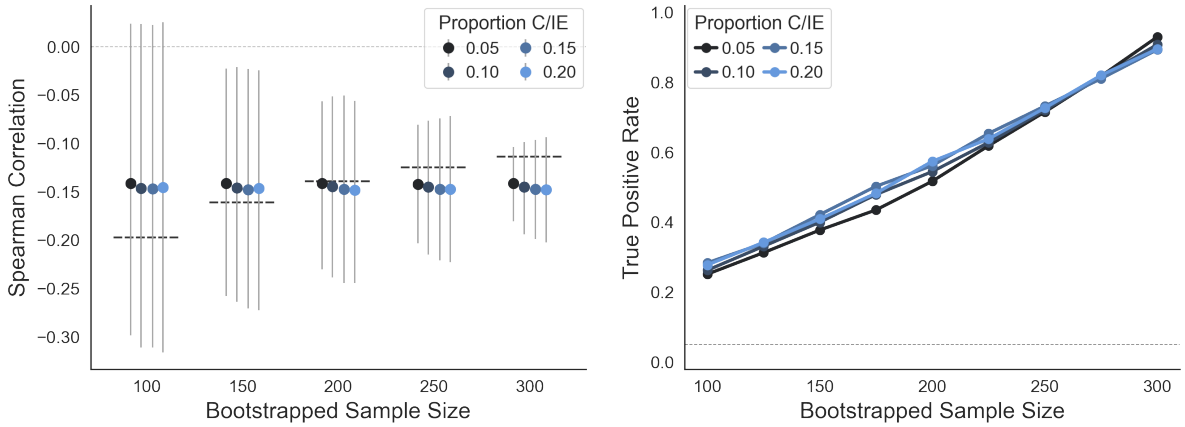

Figure S4: True correlations are independent of the proportion of C/IE participants in the sample. *Left:* Bootstrapped Spearman correlations between learning rate asymmetry (κ) and anxiety scores (GAD-7) as a function of sample size and proportion of C/IE participants. The grey error bars indicate 95% bootstrap confidence intervals. The thick dashed lines indicate the threshold for statistical significance for the Spearman correlation at the corresponding sample size. Markers are jittered along the x-axis for legibility. *Right:* True positive rates for learning rate asymmetry (κ) and anxiety scores (GAD-7) as a function of sample size and proportion of C/IE participants. True positive rate was calculated as the proportion of bootstrap samples in which the correlation between κ and GAD-7 was statistically significant. The horizontal dotted line denotes the expected false positive rate at $\alpha = .05$.

Posterior predictive checks

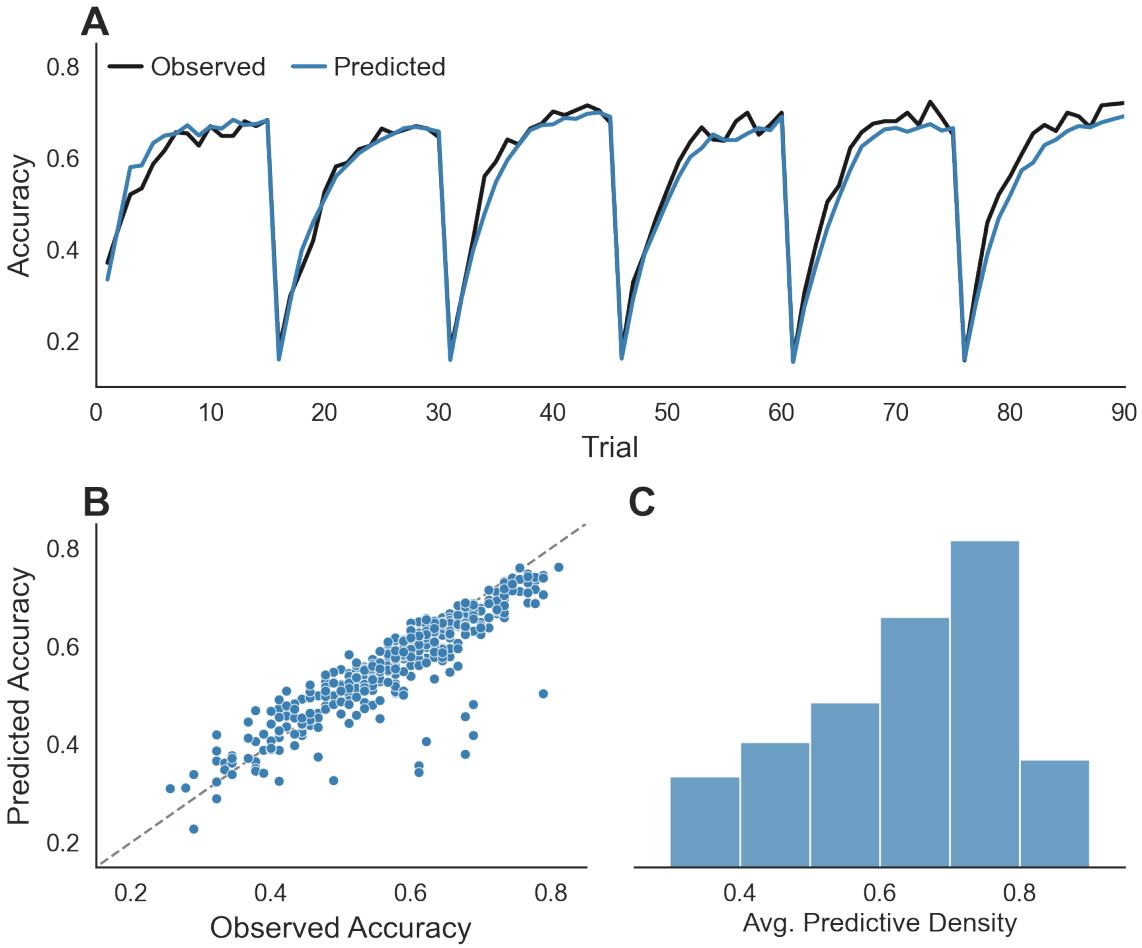

Figure S5: Posterior predictive checks for the risk-sensitive temporal difference learning model. *Panel A:* Observed (black) and predicted (blue) learning curves averaged across the group. *Panel B:* Observed versus predicted choice accuracy for each participant. *Panel C:* Distribution of average predictive density across participants.

Participant demographics

	MTurk		Prolific	
Total	N=186		N=200	
Age	N	%	N	%
18-25	16	8.6	78	39.0
26-35	76	40.9	69	34.5
36-45	46	24.7	31	15.5
46-55	22	11.8	13	6.5
55+	26	14.0	9	4.5
Gender	N	%	N	%
Female	83	44.6	112	56.0
Male	103	55.4	85	42.5
Other	0	0.0	3	1.5
Ethnicity	N	%	N	%
Hispanic or Latino	15	8.1	10	5.0
Not Hispanic or Latino	168	90.3	183	91.5
Rather not say	2	1.1	7	3.5
Unknown	1	0.5	0	0.0
Race	N	%	N	%
African American	21	11.3	7	3.5
Asian	5	2.7	53	26.5
White	151	81.2	121	60.5
Multiracial	6	3.2	4	2.0
Rather not say	1	0.5	12	6.0
Use other platform	N	%	N	%
Yes	71	38.2	28	14.0
No	115	61.8	172	86.0

Table S1: The demographics of each sample by online labour market. On average, the samples were similar though the sample from Mechanical Turk was older ($t = 6.567$, $p < 0.001$) and comprised of fewer women ($z = 6.567$, $p = 0.011$). Note: the demographics do not include 20 participants excluded for participating in the study twice, once per platform.

Careless participants show different behaviors on the reversal learning task

	Attentive	C/IE	t-value
Acc	0.587	0.532	4.008*
Pts	50.163	47.729	2.376*
WS	0.898	0.776	5.387*
LS	0.609	0.751	−5.335*
Pers	0.245	0.259	−1.505
β	6.754	4.082	5.404*
η_p	0.643	0.551	2.846*
η_n	0.738	0.784	−1.516
κ	-0.069	-0.218	3.729*

Table S2: Measures of task behavior compared between attentive and C/IE participants. Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry. * Denotes statistical significance ($\alpha = 0.05$, not corrected for multiple comparisons).

Correspondence of screening measures

The following are the unthresholded results of the screening measure correspondence analyses.

	INF	ISD	REL	MAH	READ	VAR	ACC	WSLS	RT
INF	-								
ISD	0.372*	-							
REL	-0.408*	-0.811*	-						
MAH	0.406*	0.843*	-0.643*	-					
READ	-0.111	0.193*	-0.168*	0.138	-				
VAR	-0.061	-0.029	0.058	-0.024	-0.026	-			
ACC	-0.206*	-0.154	0.099	-0.182*	-0.074	0.027	-		
WSLS	0.060	0.102	-0.089	0.115	0.103	-0.060	0.221*	-	
RT	0.040	-0.019	0.014	0.013	-0.158	-0.007	-0.094	-0.05	-

Table S3: Spearman's rank correlations of task and self-report data screening measures. Each entry corresponds to the Spearman correlation between two screening measures. Acronyms: INF = infrequency item; ISD = inter-item standard deviation; REL = personal reliability; MAH = Mahalanobis distance; READ = reading time; VAR = choice variability; ACC = choice accuracy; WSLS = win-stay lose-shift rate; RT = suspicious response times. *Denotes statistical significance after correcting for multiple comparisons.

	INF	ISD	REL	MAH	READ	VAR	ACC	WSLS	RT
INF	-								
ISD	0.462*	-							
REL	0.484*	0.691*	-						
MAH	0.516*	0.732*	0.619*	-					
READ	0.319	0.165	0.165	0.216	-				
VAR	0.208	0.216	0.238	0.259	0.292	-			
ACC	0.379*	0.312	0.258	0.344	0.237	0.282	-		
WSLS	0.253	0.247	0.227	0.258	0.299	0.303	0.505*	-	
RT	0.267	0.219	0.271	0.271	0.333	0.251	0.239	0.26	-

Table S4: Dice similarity coefficients (top 10%) for task and self-report data screening measures. Each entry corresponds to the Dice coefficient between two screening measures for the 10% most suspicious participants. Acronyms: INF = infrequency item; ISD = inter-item standard deviation; REL = personal reliability; MAH = Mahalanobis distance; READ = reading time; VAR = choice variability; ACC = choice accuracy; WSLS = win-stay lose-shift rate; RT = suspicious response times. *Denotes statistical significance after correcting for multiple comparisons.

	INF	ISD	REL	MAH	READ	VAR	ACC	WSLS	RT
INF	-								
ISD	0.355*	-							
REL	0.355*	0.564*	-						
MAH	0.355*	0.667*	0.359*	-					
READ	0.290*	0.231	0.154	0.231	-				
VAR	0.116	0.080	0.080	0.160	0.133	-			
ACC	0.269*	0.137	0.164	0.192	0.247	0.171	-		
WSLS	0.242	0.103	0.154	0.205	0.231	0.240	0.630*	-	
RT	0.164	0.105	0.132	0.184	0.289*	0.164	0.225	0.237	-

Table S5: Dice similarity coefficients (top 25%) for task and self-report data screening measures. Each entry corresponds to the Dice coefficient between two screening measures for the 25% most suspicious participants. Acronyms: INF = infrequency item; ISD = inter-item standard deviation; REL = personal reliability; MAH = Mahalanobis distance; READ = reading time; VAR = choice variability; ACC = choice accuracy; WSLS = win-stay lose-shift rate; RT = suspicious response times. *Denotes statistical significance after correcting for multiple comparisons.

Correlations between behavior and symptom measures

The following are the unthresholded results of the correlation analyses between task behavior and self-reported symptoms.

	7u	7d	GAD-7	BIS	BAS	SHAPS	PSWQ
Acc	-0.295*	-0.166*	-0.093*	-0.134*	-0.020	-0.051	-0.037
Pts	-0.225*	-0.076	-0.023	-0.144*	-0.061	-0.051	0.024
WS	-0.327*	-0.160*	-0.129*	-0.171*	-0.006	-0.048	-0.062
LS	0.285*	0.158*	0.146*	0.050	-0.037	0.000	0.110*
Pers	0.134*	0.066	0.032	0.166*	0.018	0.080	-0.004
β	-0.370*	-0.157*	-0.114*	-0.185*	0.017	-0.063	-0.043
η_p	-0.097*	-0.105*	-0.101*	-0.033	-0.020	-0.015	-0.041
η_n	0.168*	0.094*	0.108*	-0.050	-0.056	-0.020	0.042
κ	-0.175*	-0.137*	-0.147*	-0.008	-0.020	-0.028	-0.061

Table S6: Spearman correlations between task behavior and self-report symptom measures when no screening and rejections have been applied. Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry. *Denotes statistical significance ($\alpha = 0.05$, not corrected for multiple comparisons).

	7u	7d	GAD-7	BIS	BAS	SHAPS	PSWQ
Acc	-0.263*	-0.144*	-0.105*	-0.106*	-0.009	-0.020	-0.033
Pts	-0.187*	-0.042	-0.020	-0.126*	-0.055	-0.028	0.036
WS	-0.291*	-0.137*	-0.123*	-0.161*	-0.019	0.006	-0.044
LS	0.314*	0.170*	0.156*	0.034	-0.036	-0.037	0.124*
Pers	0.083	0.022	0.011	0.151*	0.010	0.076	-0.021
β	-0.332*	-0.134*	-0.109*	-0.173*	0.010	-0.040	-0.023
η_p	-0.056	-0.089	-0.105*	-0.013	-0.029	0.034	-0.035
η_n	0.259*	0.134*	0.122*	-0.021	-0.063	0.019	0.053
κ	-0.171*	-0.125*	-0.150*	-0.016	-0.032	-0.033	-0.060

Table S7: Spearman correlations between task behavior and self-report symptom measures after applying rejections based on choice accuracy (7% of sample). Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry. *Denotes statistical significance ($\alpha = 0.05$, not corrected for multiple comparisons).

	7u	7d	GAD-7	BIS	BAS	SHAPS	PSWQ
Acc	-0.210*	-0.048	0.009	-0.082	-0.020	-0.005	0.009
Pts	-0.167*	0.040	0.070	-0.107*	-0.046	-0.011	0.071
WS	-0.220*	-0.045	-0.008	-0.143*	-0.025	0.001	0.016
LS	0.219*	0.084	0.113*	0.019	-0.021	-0.050	0.082
Pers	0.088	-0.028	-0.024	0.127*	0.006	0.048	-0.052
β	-0.257*	-0.038	0.047	-0.180*	-0.011	-0.032	0.056
η_p	-0.052	-0.064	-0.079	-0.015	-0.008	0.004	-0.055
η_n	0.165*	0.067	0.141*	-0.037	-0.089	-0.030	0.054
κ	-0.111*	-0.046	-0.137*	0.011	0.015	0.002	-0.054

Table S8: Spearman correlations between task behavior and self-report symptom measures after applying rejections based on infrequency items (22% of sample). Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry. *Denotes statistical significance ($\alpha = 0.05$, not corrected for multiple comparisons).

	7u	7d	GAD-7	BIS	BAS	SHAPS	PSWQ
Acc	-0.229*	-0.090	-0.072	-0.053	0.016	0.034	-0.040
Pts	-0.177*	0.016	0.011	-0.086	-0.013	0.025	0.038
WS	-0.227*	-0.085	-0.057	-0.131*	-0.025	0.049	-0.007
LS	0.236*	0.094	0.107*	0.008	-0.016	-0.075	0.079
Pers	0.095	-0.007	0.014	0.110*	-0.015	0.037	-0.025
β	-0.255*	-0.072	-0.008	-0.165*	-0.011	-0.015	0.028
η_p	-0.045	-0.088	-0.108*	0.004	-0.000	0.049	-0.073
η_n	0.196*	0.046	0.098*	-0.015	-0.085	0.008	0.024
κ	-0.114*	-0.046	-0.134*	0.007	0.012	-0.004	-0.051

Table S9: Spearman correlations between task behavior and self-report symptom measures after applying rejections based on both choice accuracy and infrequency items. Acronyms: Acc = choice accuracy; Pts = total points earned; WS = win-stay rate; LS = lose-shift rate; Pers = perseveration errors; β = inverse temperature; η_p = positive learning rate; η_n = negative learning rate; κ = learning rate asymmetry. *Denotes statistical significance ($\alpha = 0.05$, not corrected for multiple comparisons).

Task Instructions

The following are the instructions given to participants for the probabilistic reversal learning task. As a reminder, the task was given a fishing-themed cover story. Each paragraph below denotes one screen of instructions.

Welcome to the fishing game! We will now give you some instructions on how to play the game. Use the buttons below (or the arrow keys) to navigate the instructions.

In the fishing game, there are three beaches you can fish at. Each beach has its own unique surfboard. (The colors and pictures on the surfboards are there just to help you tell the beaches apart – they don't have any special meaning other than that.)

On each turn you will be shown three beaches, and you will choose which one you want to fish at. You can make your choice using the left, up, and right arrow keys.

When you fish at a beach, you will either catch a fish or you will catch trash. Try to catch fish, and try not to catch trash!

Some beaches are better than others. You are more likely to catch fish at some beaches (though you will still sometimes catch trash), and you are more likely to catch trash at other beaches (though you still sometimes catch fish).

The beaches will change over time. As times goes by, you may be less likely to catch fish at a beach where you were previously catching many fish.

Your goal is to catch as many fish as you can. You will receive a performance bonus up

to $0.25 that depends on how many fish you catch.

Now we will ask you some questions about the game. You must answer all questions correctly to proceed. Feel free to read back through the instructions if there is anything you are not certain about.

Following the instructions, participants completed a brief comprehension check where they were asked the following questions about the task:

1. True or False: Your goal is to catch as many fish as you can. (True)
2. True or False: You are more likely to catch fish at some beaches than others. (True)
3. True or False: You will always catch fish at the best beach. (False)
4. True or False: How likely you are to catch a fish at a beach stays the same over time. (False)
5. True or False: The number of fish you catch will affect your final performance bonus. (True)

Participants were required to answer all of the items correctly before they could proceed to the task. If they failed to do so, they restarted the instructions. There was no upper limit as to how many times a participant could loop through the instructions (the large majority of participants passed the comprehension check on their first try).

C/IE responding manifests as qualitatively distinct behavioral strategy

In an exploratory analysis, we employed a theory-agnostic modeling approach to investigate how C/IE participants on the probabilistic reversal-learning task compared to attentive participants. The motivation for this analysis was to better understand why C/IE responding was inconsistently predicted by chance-level performance, and also correlated with asymmetric learning rates.

To characterize participants' choice behavior, we adapted the softmax regression model from [1]. This model estimates, for each participant, how much their choice depends on the recent history of trial events (rewarding outcomes, non-rewarding outcomes, and choices from the preceding 5 trials). Specifically, the influence of the history of particular type of event is defined as:

$$\sum_{i}^{K} w = x_{t-1} \cdot w_{t-1} + x_{t-2} \cdot w_{t-2} + \ldots + x_{t-k} \cdot w_{t-k}$$

where x_{t-i} is a binary indicator [0,1] denoting if an event (i.e. reward, non-reward, previous choice) occurred on trial $t - i$ and w_{t-1} is the associated decision weight. These

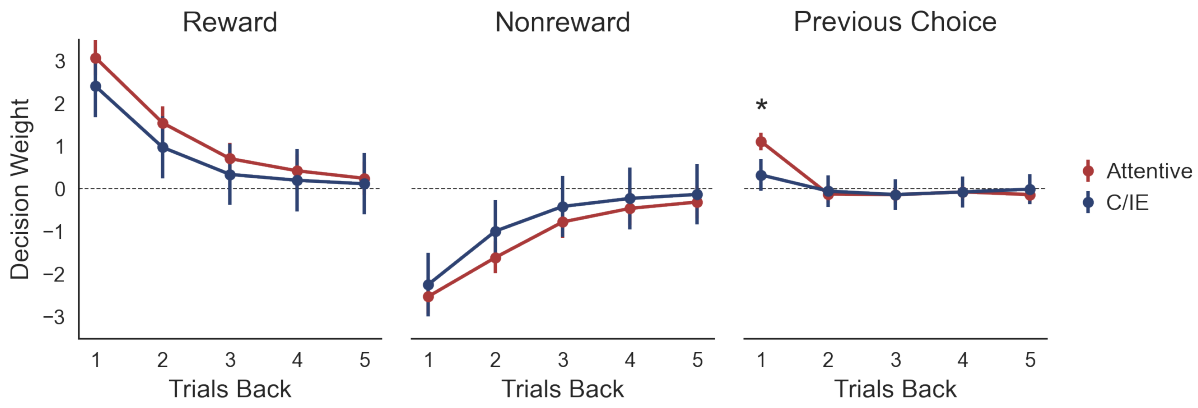

Figure S6: Softmax regression decision weights in attentive (red) and C/IE (blue) participants. The weights dictate the extent to which the recent history of rewards, nonrewards, or previous choices influence current choice. Error bars indicate the 95% highest density interval of each weight. * denotes where the 95% highest density interval of the difference in weights excluded zero.

weights were estimated for rewards, non-rewards, and previous choices up to five trials in the past. The overall tendency to choose a particular choice option is dictated by a softmax choice rule:

$$p(y_t = i) = \frac{\exp\left(\sum w_i^{\text{reward}} + \sum w_i^{\text{nonreward}} + \sum w_i^{\text{choice}}\right)}{\sum_i \exp\left(\sum w_i^{\text{reward}} + \sum w_i^{\text{nonreward}} + \sum w_i^{\text{choice}}\right)}$$

Note that these weights were fit independently; that is, we did not employ an exponential kernel to parameterize the decay of the weights at successively distant trial lags. In sum then, the theory-agnostic model describes a participant's choice behavior as a function of 15 parameters.

We fit the softmax regression model using Stan following the same procedure as for the theory-based analyses. Participant parameters were fit individually (i.e. not hierarchically) so as to prevent bias during parameter estimation from partial-pooling between attentive and C/IE participants. Parameters were sampled with Gaussian priors with $\mu = 0$ and $\sigma = 5$.

The regression weights for each event, averaged within attentive and C/IE participants, are presented in Figure S6. Comparing attentive to C/IE participants, we observed a credible difference (i.e. 95% highest density intervals excluded zero) only for the $T - 1$ weight for previous choice. That is, attentive participants were more likely to repeat their previous choice (i.e. greater choice hysteresis) than were C/IE participants.

This result may be initially surprising, since one might expect choice hysteresis to result in more perseveration errors following contingency reversals. However, choice hysteresis is adaptive in this probabilistic reversal-learning task. Because rewards in the task are probabilistic, once the reward-maximizing response option has been identified ignoring an occasional unrewarding outcome and instead performing the same response is optimal

(until the next reversal occurs and is identified). Interestingly, participants engaging in C/IE responding were also numerically (though not significantly) less affected by previous outcomes, suggesting that their behavior was not more adaptive, but rather just more random.

This pattern of results also helps explain the pre-screening correlations with asymmetric learning rates. Previous work has established that, when choice hysteresis is not accounted for in reinforcement learning models, it can manifest as positive learning rate asymmetries [2, 3]. Since C/IE participants showed decreased hysteresis, which our reinforcement learning model did not explicitly account for, we should expect to find a negative correlation between learning-rate asymmetries and symptoms before C/IE participants are excluded. Indeed, this is what we observed above.

In sum, the theory-agnostic analysis of task behavior revealed that C/IE participants exhibited a qualitatively distinct behavioral strategy on the probabilistic reversal-learning task. C/IE participants showed less adaptive choice hysteresis. Moreover, they were numerically (but not significantly) less sensitive to outcomes. The latter finding helps clarify in part why we observed low correspondence between task and self-report screening measures (that is, C/IE participants were not significantly more likely to respond randomly during the task). These results also present another hidden danger of C/IE responding: qualitatively distinct patterns of behavior under C/IE responding can bias the estimation of parameters of theoretical interest if not properly accounted for.

Appendix B: Replication study

Background & motivation

Here we report a conceptual replication of our original study. The motivations for conducting a replication study were threefold. First, we wanted to examine the generalizability of our findings under new labour-platform conditions. Since summer 2020, when the original data were collected, online labour markets like Mechanical Turk/CloudResearch and Prolific Academic have undergone important changes. To address rampant data quality issues on Mechanical Turk, CloudResearch introduced their "Approved Participants" filter. When selected, only Mechanical Turk participants with a prior history of attentive and careful work are invited to participate in experiments [4, 5]. Similarly, a deluge of new, lower-quality users signed up to participate in studies on Prolific in summer, 2021 [6]. In response, Prolific introduced new controls and filters to improve data quality on the platform [7]. In the wake of these changes, we wanted to explore the relevance of our original findings. Specifically, with these new safeguards, we wanted to examine whether the chance of spurious correlations has been considerably reduced.

Second, we wanted to examine the generalizability of our findings to other behavioral measures. In the original study, we used a short, straightforward, and relatively easy reversal-learning task. Our motivation then was to demonstrate the risk of spurious correlations even for experiments where the possibility of participant fatigue (and consequently C/IE responding) had been minimized. One consequence of this design, however, was that the majority of participants performed reasonably well on the task (only 26 participants, or 7% of the sample, exhibited choice accuracy at or below chance levels). This could in part explain why we observed such low correspondence between self-report and task-based screening measures. As such, we wanted to repeat our experiment and analyses using a more difficult task. Therefore, in the replication study we used the two-step task [8], which is both more challenging and takes longer than our original reversal-learning task.

Finally, we wanted to examine the generalizability of our findings to other self-report measures. In the original study, we found that self-report symptom measures with low rates of endorsement were more likely to yield spurious correlations with behavior in the presence of C/IE responding, as C/IE participants were more likely to endorse symptoms, as well as to perform poorly on the task. In principle, this effect should not be limited to symptom measures, and should extend to any self-report measure with an expected skewed or asymmetric score distribution. Therefore, in the current replication study, we used two sets of self-report scales: one set of psychiatric symptom measures and one set of personality measures. As before, each set includes scales whose score distributions are expected to be symmetric as well as scales with asymmetric (skewed) score distributions. Crucially, the two personality scales we chose, artistic interests and greed avoidance, measure constructs that should have no meaningful relationship with model-based choice behavior on the two-step task. Therefore, the personality measures serve as a stronger test

of our hypothesis that spurious correlations between self-report and behavioral measures are more likely for skewed score distributions.

Methods

Participants

400 total participants were recruited to participate in an online behavioral experiment in February, 2022. Specifically, 200 participants were recruited from Amazon Mechanical Turk (MTurk) and 200 participants were recruited from Prolific. The study was approved by the Institutional Review Board of Princeton University (#11968), and all participants provided informed consent. Total study duration was approximately 20 minutes. Participants received monetary compensation for their time (rate USD $12/hr), plus an incentive-compatible bonus up to $1.00 based on task performance.

Participants were eligible if they resided in the United States or Canada. Participants from MTurk were recruited with the aid of CloudResearch services [9] using their "Approved participants" data quality filters [4]. As in the original study, MTurk workers were not excluded based on work approval rate or number of previous jobs approved [10]. No other exclusion criteria were applied during recruitment.

Data from N=7 participants who completed the experiment were excluded prior to analysis because these participants (all from MTurk) disclosed that they had also completed the same experiment on the other platform. This left a final sample of N=393 participants (MTurk: N=193, Prolific: N=200) for analysis. The demographics of the sample split by labour market is provided in Table S10. Participants recruited from MTurk were older on average ($\Delta M = 4.9$ yrs, $t = 4.248, p < 0.001$) and comprised of fewer women (35.2% versus 61%, $z = 4.248, p < 0.001$).

Experiment

Participants completed a gamified version of the two-step task [8] designed to dissociate "model-free" and "model-based" decision-making. On every trial of the task, participants' goal is to collect as much "space treasure" as possible by traveling to one of two different planets and "trading" with one of two aliens who live on that planet. Participants first chose between two different-colored rocket ships (first-stage choice). Each rocket ship had a 70% chance of traveling to one particular planet (e.g., the green rocket ship to the blue planet and the purple rocket ship to the red planet; common transitions) and a 30% chance of traveling to the other planet (uncommon transition). The rocket ship and planet colors were randomized across participants, as were the mappings between rocket ships and planets. On each planet, participants chose which of two aliens to "trade" with (second-stage choice). If chosen, an alien would give the participant "treasure" with some slowly-changing probability, otherwise it would give "junk". The reward probabilities for

	MTurk		Prolific	
Total	N=193		N=200	
Age	N	%	N	%
18-25	11	5.7	47	23.5
26-35	71	36.8	76	38.0
36-45	60	31.1	41	20.5
46-55	29	15.0	22	11.0
55+	22	11.4	14	7.0
Gender	N	%	N	%
Female	68	35.2	122	61.0
Male	124	64.2	73	36.5
Other	1	0.5	5	2.5
Ethnicity	N	%	N	%
Hispanic or Latino	16	8.3	12	6.0
Not Hispanic or Latino	177	91.7	180	90.0
Rather not say	0	0.0	8	4.0
Race	N	%	N	%
American Indian/Alaska Native	2	1.0	0	0.0
Asian	15	7.8	41	20.5
Black or African American	13	6.7	12	6.0
White	156	80.8	133	66.5
Multiracial	6	3.1	10	5.0
Rather not say	1	0.5	4	2.0

Table S10: The demographics of each sample by online labour market.

each alien and trial were generated according to independent Gaussian random walks. Participants completed 201 trials of the task, with an optional break after the first half.

Prior to the beginning of the experiment, participants had to read a set of instructions in which they were told that the rocket ships mostly traveled to one planet, but sometimes went to the other, and that the chance an alien would give them treasure would change slowly over the course of the task. Before they could start the task, participants had to correctly answer three sets of comprehension questions about the instructions. Failing to correctly answer all items forced the participant to reread a section of the instructions. Participants were permitted up to ten retries of the comprehension questions before they were removed from the experiment; however, no participant exceeded this limit.

The task was programmed in jsPsych [11] and distributed using custom web-application software. The experiment code is available at https://github.com/nivlab/sciops, and the web-software is available at https://github.com/nivlab/nivturk. A playable demo of the task is available at https://nivlab.github.io/jspsych-demos/tasks/

two-step/experiment.html.

Self-report measures

Prior to the start of the two-step task, participants completed four self-report measures in a randomized order. One was the 14-item seven-up/seven-down scale (7u/7d; [12]), which measures lifetime incidence of depressive and (hypo)mania symptoms. This scale is expected to elicit lower rates of symptom endorsement, thereby resulting in asymmetric (right-skewed) score distributions. Participants also completed a 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.

In addition, participants completed a 6-item measure of artistic interests (e.g. "I believe in the importance of art"; [14]). Based on previous studies, this scale is expected to elicit high rates of endorsement, thereby resulting in an asymmetric score distribution. 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 .

Correspondence of screening measures

As in the original study, we measured the correspondence of screening measures based on the task and self-report behavior. We calculated a number of standard measures of data quality from each participant's task behavior (four in total) and self-report responses (five in total). The self-report screening measures were identical to those used in the original study, except that we used (mostly) new infrequency items. We describe each of the new screening measures below.

Self-report screening measure: Infrequency items. Infrequency items are questions for which all (or virtually all) attentive participants should provide the same response. We embedded four infrequency items across the self-report measures. Specifically, we used the following questions:

1. Have there been times in your life where you blinked your eyes at least once per day? (Expected response: *Very often*)
2. Have there been times of a couple days or more when you were able to breathe underwater (without an oxygen tank)? (Expected response: *Never or hardly ever*)

3. I was worried about the canine World Cup. (Expected response: *Not at all*)
4. I have used a computer. (Expected response: *Slightly Agree*, *Agree*, or *Strongly agree*)

Prior to conducting the study, the infrequency items were piloted on an independent sample of participants to ensure that they elicited one dominant response. We also included one instructed item ("Please select 'Neutral' as your response") to compare to the infrequency items. We measured the number of 'suspicious' (i.e., incorrect) responses made by each participant to these questions. For thresholded analyses, participants were flagged if they responded incorrectly to one or more of these items.

Task-based screening variable: Side variability. Side variability was defined as the fraction of trials a participant chose the left option (by pressing the left arrow key) across first stage choices during the two-step task. Side variability could range from 0.00 (only right arrow key used) to 1.00 (only left arrow key used). Extreme values (i.e. closer to zero or one) are indicative of more careless responding during the task, as the sides for which each choice option was displayed was determined randomly on each trial.

Task-based screening variable: Choice variability. Choice variability was defined as the fraction of trials a participant chose the same first-stage choice option (randomized to the right or left side of the screen) during the two-step task. Choice variability could range from 0.00 (selected the green rocket ship exclusively) to 1.00 (selected the purple rocket ship exclusively). Extreme values (i.e. closer to zero or one) are indicative of more careless responding during the task as the most rewarding option changed throughout the task.

Task-based screening variable: Win-Stay Lose-Shift. Win-stay lose-shift (WSLS) measures a participant's tendency to stay with a first-stage choice option following a second-stage reward versus shifting to a the other choice option following a non-reward. WSLS thus measures a participant's sensitivity to reward feedback. WSLS was estimated per participant via regression, predicting each first-stage choice (stay, switch) by the previous trial's outcome (reward, non-reward) and an intercept. We used the first (slope) term to represent a participant's WSLS tendency. Lower values of this term indicate less sensitivity to reward feedback and are thus indicative of more careless responding during the task. Thresholds for chance-level WSLS performance were determined by fitting the same regression model to 5000 randomly-generated datasets of first-stage choice (datasets were generated by matching the probability of staying with the previous trial's choice to the distribution observed empirically, but choices were otherwise independent across trials; that is, independent of previous outcome). The threshold for above-chance WSLS was defined as the 95th percentile of the distribution of slope estimates for the random data, corresponding to a one-tailed hypothesis test ($\alpha = 0.05$) that the slope coefficient is greater than zero.

Task-based screening variable: Response times. "Suspicious response time" was defined as the proportion of first-stage choices with a response faster than 200ms. Greater proportions of outlier response times are indicative of more careless responding during the task.

Correspondence Analysis. We measured the correspondence of the above screening measures via two complementary approaches. First, we computed pairwise correlations on the unthresholded (continuous) measures using Spearman's rank correlation. Second, we estimated the pairwise rate of agreement on the binarized measures using the Dice similarity coefficient (looking at the top 10% most suspicious respondents for each measure). The former approach estimates two measures' monotonic association, whereas the latter approach estimates their agreement as to which participants were most likely engaging in C/IE responding. For significance testing, we used permutation testing wherein a null distribution of similarity scores (Spearman's correlations or Dice coefficients) was generated for each pair of screening measures by iteratively permuting participants' identities within measures and re-estimating the similarity. P-values were computed by comparing the observed score to its respective null distribution. We corrected for multiple comparisons using family-wise error rates [15].

Correlations between behavior and symptom measures

To quantify the effects of both task and self-report data screening on behavior-symptom correlations, we estimated the pairwise correlations between the scale scores of each self-report measure and several measures of model-agnostic performance on the two-step task [16]. Logistic regression analyses were conducted with the *statsmodels* package [17] in the python programming language. The model tested if participants' first-stage choice behavior (coded as Stay = 1, Switch = 0) was influenced by the previous trial's reward (coded as Rewarded = 1, Unrewarded = 0), previous trial's transition (coded as Common = 1, Uncommon = 0), and their interaction. Importantly, the interaction term between previous reward and transition is a proxy for the contribution of model-based learning to choice behavior [16].

Correlations between the behavioral measures (i.e. logistic regression coefficients) and self-report measures were calculated using Spearman's rank correlation, after various forms of screening and exclusion. Significance testing was performed using the percentile bootstrap method [18] so as to avoid making any parametric assumptions. These correlation analyses were not corrected for multiple comparisons, since our overarching purpose was to demonstrate the extent of this issue across multiple behavioral measures and self-report symptoms. Any one of these correlations considered individually can be thought of as emulating a conventional analysis where fewer statistical tests would be performed.

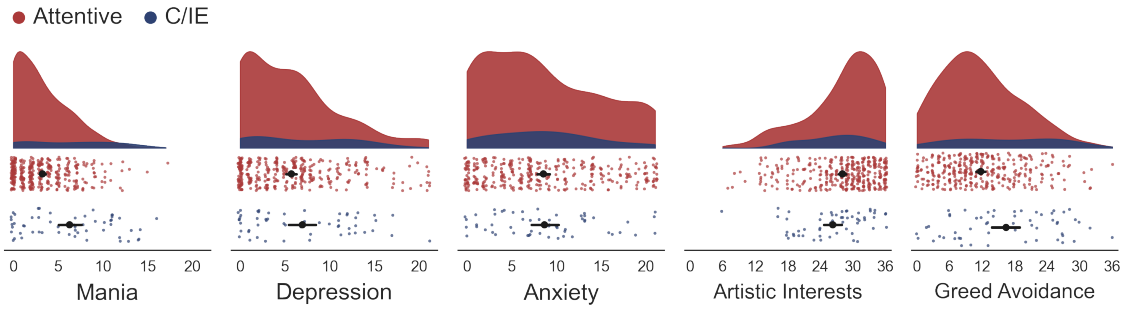

Figure S7: Raincloud plots of total symptom scores in attentive (red) and C/IE (blue) participants. Each colored dot represents the symptom score for one participant. Black circles: average score within each group (error bars denote 95% bootstrap confidence interval). Shaded plots: estimated distribution of responses for each group of participants.

Results

Careless participants (often) appear symptomatic when the overall level of symptom endorsement is low

To begin our analysis of the replication dataset, we examined the number of participants flagged by the WSLS and infrequency item screening measures. Only 31 participants (8%) were flagged as exhibiting choice behavior at or below statistically chance levels in the two-step task. In contrast, 55 participants (14%) endorsed a logically invalid or improbable response on one or more of the infrequency items when completing the self-report measures. The proportion of participants flagged for C/IE responding was significantly greater on Mechanical Turk compared to Prolific for both task (MTurk: N=23/193; Prolific: N=8/200; $z = 2.91, p = 0.004$) and survey data (MTurk: 34/193; Prolific: 21/200; $z = 2.03, p = 0.042$). Across the four infrequency items, the average failure rate was 5.3% (range: 2.5% – 8.9%). In contrast, no participant failed the instructed item. This discrepancy in the proportion of participants flagged by each type of attention check is consistent with previous research, which found that instructed items are insensitive measures of C/IE responding [19–21].

Previously, we observed a mean-shift in the average level of symptom endorsement for participants suspected of engaging in C/IE responding relative to attentive participants on measures for which the overall rate of symptom endorsement is low. This result was (mostly) replicated in the current dataset. Total scores were noticeably exaggerated in participants suspected of C/IE responding for the symptom measures where overall rates of symptom endorsement were the lowest (e.g. mania; Figure S7, leftmost plot). Where there were higher rates of symptom endorsement overall (e.g. anxiety), the distributions of symptom scores between the two groups of participants were more similar (Figure S7, middle plots). Permutation testing confirmed that observed mean-shifts in symptom scores for C/IE participants were statistically significant for the most skewed symptom measures (Table S11).

For the personality measures, however, an interesting pattern emerged (Figure S7, right-

		Total Score		
Subscale	Skew	Attentive	C/IE	t-value
Mania	1.07	3.25	6.27	-6.05^*
Depression	0.89	5.72	6.95	-1.57
Anxiety	0.44	8.54	8.65	-0.13
Artistic interests	-0.92	28.07	26.29	1.87
Greed avoidance	0.55	11.77	16.44	-4.17^*

Table S11: Descriptive statistics of the self-report measures between attentive and C/IE participants. Skew: the empirical skewness of the distribution of total symptom scores. Total score: the average symptom score across attentive and C/IE participants. Stars indicate statistical significance at $p < 0.05$.

most plots). We did not observe a statistically significant mean-shift in total scores for the subscale with the most skewed score distribution (i.e. artistic interests; $t = 1.87$, $p = 0.061$). However, an unexpected and statistically significant mean-shift in the average level of endorsement was observed for the more symmetrically-distributed greed-avoidance subscale ($t = 16.44$, $p < 0.001$). The reason for this finding is unclear. Regardless, this finding presents an opportunity to test for spurious correlations between behavioral and self-report measures in the presence of a mean-shift in scores in the absence of (substantially) skewed score distributions.

Low correspondence between task and self-report measures of C/IE responding

Next, we evaluated the degree of correspondence between behavioral and self-report screening measures to determine whether screening on behavior alone was sufficient to identify and remove careless participants. To measure the degree of correspondence between these behavioral and self-report screening measures, we performed two complementary analyses. First, we computed pairwise correlations on the unthresholded (continuous) measures using Spearman's rank correlation (Figure S8, left panel). After correcting for multiple comparisons, there were a handful of significant correlations between the behavioral and self-report screening measures. Abnormal response times emerged as the metric most correlated with the self-report screening measure. Crucially, as in the original study, the sizes of these observed correlations were roughly half those observed for the correlations between the self-report measures.

Second, we used the Dice similarity coefficient to quantify agreement between different screening methods in the set of participants flagged for exclusion (Figure S8, right panel). This approach quantifies the degree of overlap between the set of would-be excluded participants based on different screening measures under a common exclusion rate. Results were largely consistent with the correlation analysis: only a handful of task and self-report screening measures achieved levels of agreement greater than what would be expected by

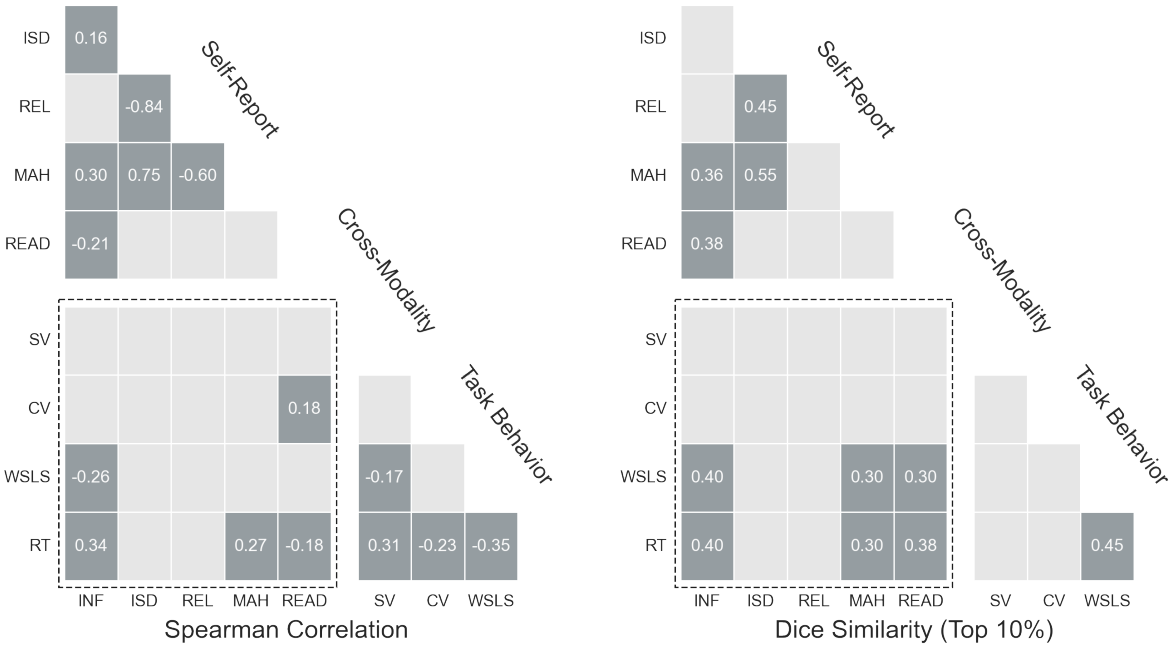

Figure S8: Similarity of task and self-report data screening measures. Each tile corresponds to the Spearman rank correlation (left) and Dice similarity coefficient (right) between two screening measures. Acronyms: INF = infrequency item; ISD = inter-item standard deviation; REL = personal reliability; MAH = Mahalanobis distance; READ = reading time; SV = side variability; CV = choice variability; WSLS = win-stay lose-shift rate; RT = suspicious response times. Similarity scores have been thresholded after correcting for multiple comparisons. Numbers denote the strength of statistically significant correlations. Cross-modality correlations between task-behavior (left) and infrequency-item self-report measures (bottom) are in the dashed rectangle.

chance. Of the significant cross-modality pairs, the average similarly coefficient was less than 0.4. In other words, when any of these sets of two measures are used to identify the top 10% of participants most strongly suspected of C/IE responding, they agree on only two out of every five participants. Screening on task behavior alone would fail to identify the majority of participants most likely engaging in C/IE responding.

Taken together, these findings corroborate the results of the original study: measures of C/IE responding in task and self-report data do not identify the same set of participants. This means that solely excluding participants on the basis of poor behavioral performance—the most common approach in online studies—is unlikely to identify participants who engage in C/IE responding on self-report surveys.

Spurious symptom-behavior correlations produced by C/IE responding

To understand the effects of applying different forms of screening, we estimated the correlations between each unique pairing of a self-report measure and measure of behavior under four different conditions: no screening, screening only on task behavior (i.e. removing participants whose win-stay lose-shift behavior was not above chance), screening

only on self-report responses (i.e. removing only participants who responded incorrectly on one or more infrequency items), or both. The resulting pairwise behavior-symptom correlations following each screening procedure are presented in Figure S9. We note that we did not correct these correlation analyses for multiple comparisons, since our purpose was to demonstrate the extent of this issue across multiple behavioral measures and self-report symptoms. Any one of these correlations considered individually can be thought of as emulating a conventional analysis where fewer statistical tests would be performed.

When no rejections were applied (i.e. all participants were included; Figure S9A), we observed multiple significant correlations between measures of task behavior and symptom scores for hypomania and depression. Consistent with our predictions, these correlations involved measures with low overall endorsement rates and mean-shifts in score distributions between attentive and C/IE participants. Conversely, we found no significant correlations with the symmetrically-distributed anxiety scores. This is despite the fact this scale measures symptoms that are comorbid with depression and mania. Crucially, of the two personality measures, we observed significant correlations only for the measure found to exhibit a mean-shift in scores between attentive and C/IE participants (i.e. greed avoidance). These included a significant correlation with the interaction term, which is used as a proxy measure for model-based choice behavior. That is, significant correlations were not restricted only to general behavioral measures but also to measures of specific theoretical interest.

Next, we excluded participants from the analysis based on task-behavior screening (i.e. lack of win-stay lose-shift behavior, removing the 8% of participants exhibiting behavior indistinguishable from chance; Figure S9B). In contrast to the findings of the original study, the pattern of correlations was meaningfully changed: the putatively spurious correlations between greed avoidance and performance on the two-step task were ablated. Two previously significant correlations between hypomania and two-step performance were also rendered non-significant. A similar pattern of results was observed when we rejected participants based on self-report screening (removing 14% of participants who endorsed one or more invalid or improbable responses on the infrequency items; Figure S9C) and when rejections were applied based on both task and self-report screening measures (removing 18% of participants; Figure S9D).

These findings suggest that some of the significant behavior-symptom correlations observed without strict participant screening may indeed be spurious correlations driven by C/IE responding. Interestingly, in contrast to the original study, with a more demanding behavioral task, screening based on either task behavior or self-report behavior alone was sufficient to protect against spurious symptom-behavior correlations in the presence of mean-shifts in scores between attentive and C/IE participants. For example, both forms of screening ablated the would-be significant correlation between model-based behavior and greed avoidance. The discrepancy in results between the original and replication studies may reflect the smaller numbers of participants failing attention checks in the replication study (14%) compared to the original study (22%), as well as differences in the behavioral tasks. Regardless, we replicate the findings of the original study in that

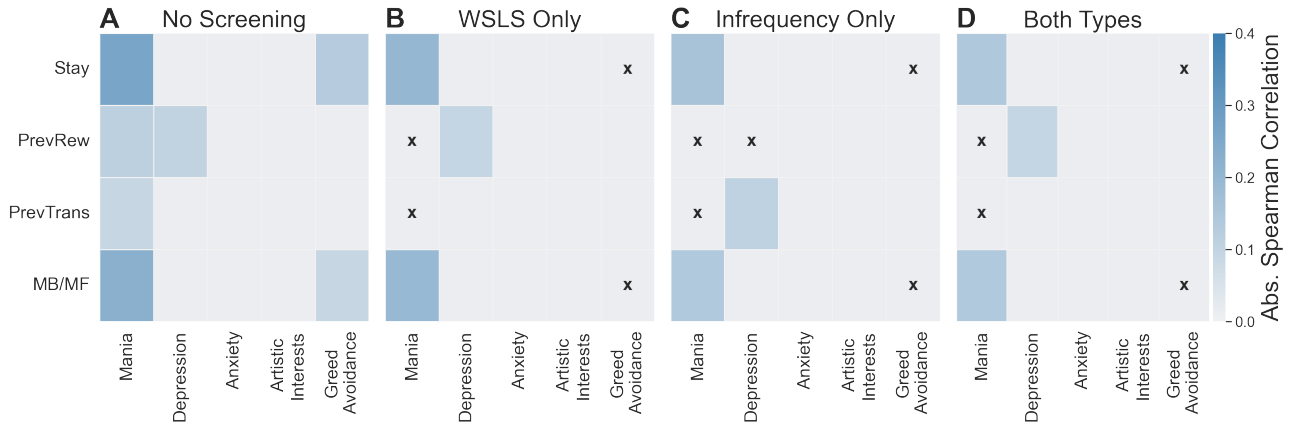

Figure S9: Absolute Spearman rank correlations between task behavior (y-axis) and symptom measures (x-axis) under different regimes of data screening and participant exclusions. Only statistically significant correlations are shown ($p < 0.05$, not corrected for multiple comparisons). Black Xs indicate significant correlations ablated under screening. No Screening = no exclusions; WSLS Only = exclusions based on chance-level performance in the two-step task; Infrequency Only = exclusions based on invalid or improbable responses to infrequency items; Both Types = exclusions based on the previous two measures. Acronyms: PrevRew = sensitivity to reward on the previous trial; PrevTrans = sensitivity to transition type on previous trial; MB/MF = index of model-based/model-free behavior (interaction between *PrevRew* & *PrevTrans*).

screening on self-report data allowed us to identify symptom-behavior correlations most likely to be spurious.

Discussion

Here we reported findings from a replication study whose purpose was to examine the generalizability of our original findings under new labour market conditions, different behavioral measures, and different self-report measures. To this end, we recruited an independent sample of almost 400 participants, using CloudResearch's and Prolific's latest data-quality filters, to complete the two-step task and a novel set of self-report measures. As evidence of the efficacy of the new data-quality filters, the proportion of participants flagged for C/IE responding in the self-report measures was noticeably smaller in the replication sample (14%) compared to original sample (22%). This decrease in the number of participants suspected of C/IE responding was observed for both MTurk and Prolific (though, as in the original study, the proportion of low-quality participants was significantly, albeit marginally, smaller for Prolific than MTurk). Regardless, although the new online labour platform quality-control measures seem to be effective, they did not completely solve the problem; indeed, the proportion of participants engaging in C/IE responding was reduced only by one-third.

Next, we compared the distribution of self-report scale scores for attentive participants and participants suspected of engaging in C/IE responding. Replicating our previous

result, we observed a mean-shift in the average level of symptom endorsement for participants suspected of C/IE responding relative to attentive participants, only when the overall rate of symptom endorsement was low. Specifically, flagged participants showed significantly elevated scores on the hypomania scale (with its right-skewed score distribution) but not so on the anxiety scale (with its symmetric score distribution). Interestingly, we found the opposite pattern for the personality measures: scores for participants engaging in C/IE responding were not significantly different than those of their attentive counterparts on the artistic interests scale (with its left-skewed score distribution), but was so for the greed avoidance scale (with its more symmetric score distribution). One possible explanation for the discrepancy in findings between the symptom and personality measures is the direction of the skew for the artistic interests scale. Previously, using random-intercept item factor analysis, we observed that participants engaging in C/IE responding were more likely to use the right-half of the response scale. As such, such a pattern of responding is more likely to produce a mean-shift in scale scores, compared to attentive participants, on a scale with a right-skewed distribution (e.g. mania, depression scales) than a scale with a left-skewed distribution (e.g. artistic interests scale). Further research is still needed to characterize patterns of C/IE responding.

The results in the replication study did corroborate the findings of the original study in terms of the degree of correspondence between behavioral and self-report screening measures, suggesting that measures of C/IE responding in task and self-report data do not identify the same set of participants. Even with a more difficult task (i.e., the two-step task), we observed relatively low correspondence between self-report and task-based screening measures. This supports our suggestion that both forms of screening are necessary to identify participants providing low-quality responses. Finally, we examined the consequences of various types of screening methods for correlations between behavioral and self-report measures. As in the original study, we detected significant spurious correlations when no screening was applied. This included correlations between model-based planning on the two-step task and scores on the greed avoidance scale, for which there is no theoretical reason to predict a correlation. Instead, this correlation almost certainly reflects the mean-shift in scores between attentive participants and participants flagged for C/IE responding on the greed avoidance scale. As evidence of this, excluding participants who failed one or more infrequency items ablated this correlation. In contrast to the original study, excluding participants based on poor performance on the two-step task was also sufficient to ablate this correlation. Thus, there may be instances where screening based on poor behavioral performance is sufficient to prevent spurious correlations. However, in the absence of perfect information as to when those situations should arise, we conclude that it is simply safer to screen participants on both dimensions of performance.

In summary, we conclude that the results of the original study are not limited to the task and self-report measures used in that study, or to online platforms at a particular point in time. Despite legitimate advances in data quality controls, online labour platforms still suffer from participants engaging in C/IE responding. Given *a priori* uncertainty regarding the ability of task measures alone to screen such participants, we recommend

also using infrequency items to detect inattentive responding on self-report measures. Finally, and most importantly, this second study strengthened the finding that C/IE responding is likely to result in mean-shifts in scores for symptom scales with overall low rates of endorsement, which are in turn capable of yielding spurious correlations between self-report and behavioral measures. The best safeguard against such spurious correlations continues to be screening in both domains.

Appendix C: Attention checks in healthy and psychiatric patients

Background & motivation

One concern with using attention checks for screening and exclusion of participants is that we might inadvertently introduce an overcontrol bias [22]. That is, to the extent that C/IE responding reflects symptoms of psychopathology such as lack of motivation [23], avoidance of effort [24, 25], or more frequent lapses of attention [26, 27], it is plausible that rigorous screening of C/IE responding might lead to the differential exclusion of truly symptomatic participants. As a result, true associations between poor or different task performance and psychopathology symptoms could go undetected (type II error). The purpose of the below preliminary and exploratory study was to examine whether individuals with a confirmed psychiatric disorder were indeed more likely than healthy controls to fail attention checks embedded in self-report symptom scales.

Methods

Participants

Participants were recruited as part of two independent studies to investigate reward processing in individuals with and without a history of major depression and bipolar disorder (results of those studies not reported here). Both studies required participants to complete (1) a structured clinical interview (the SCID-5) with a trained interviewer to verify that they met the criteria for one or more psychiatric disorders, and (2) a series of computerized self-report surveys and cognitive tasks.

In both studies, participants volunteered for a multi-session study conducted online via video conferencing. Specifically, participants completed each session from their homes while on Zoom with a study coordinator. During each session, while participants were completing self-report questionnaires or behavioral tasks, the study coordinator turned off their camera and microphone, but was available if the participant had any questions.

Participants were recruited through clinician referral and online ads (i.e., Google ads, Craigslist) targeting individuals with a history of depression, anhedonia, apathy, and/or (hypo)mania symptoms. Participants were eligible for participation if they (1) had no history of head injury resulting in loss of consciousness for more than 20 minutes; (2) had not been diagnosed with intellectual disability; (3) had not been diagnosed with any neurological condition; (4) did not meet criteria for substance dependence (excluding nicotine) in the past 6 months; (5) had not received electroconvulsive therapy in the past 8 weeks; and (6) were aged between 18-65. For one of the two studies, participants were also required to score 6 or higher on the Wechsler Test of Adult Reading to be included in the study. Furthermore, clinical participants were eligible if they met criteria for a diagnosis

of major depressive disorder and, if they were on medication, they had been on on stable treatment with this medication for at least the past 4 weeks. Non-clinical (control group) participants were eligible if they did not meet criteria for any psychiatric diagnosis and were not currently taking any medication used to treat psychiatric disorders.

Infrequency items

To measure and compare C/IE responding between healthy and psychiatric participants, we used two sets of attention checks. Each set was composed of three infrequency items and one instructed item. Participants were assigned one or the other set. The six infrequency items were:

1. Worrying too much about the 1977 Olympics. (Expected response: *Not at all*)
2. I have never used a computer. (Expected response: *Completely untrue* or *Quite untrue*)
3. I would be able to lift a small (1 lb) weight. (Expected response: *Extremely characteristic of me* or *Somewhat characteristic of me*)
4. Have there been times of a couple days or more when you were able to stop breathing entirely (without the aid of medical equipment)? (Expected response: *Never*)
5. Over the past year, how often did you have days where you were able to blink your eyes without difficulty? (Expected response: *Often* or *Very often*)
6. I am generally able to remember my own name. (Expected response: *True*)

Analysis

Of primary interest here was whether psychiatric patients fail infrequency item attention checks at equal or at greater rates than healthy participants. To test this, we conducted Bayesian hypothesis testing using Bayes factors [28]. Specifically, we defined three competing models:

- $M_0 : \text{Binom}(N_1, p), \text{Binom}(N_2, p)$
- $M_1 : \text{Binom}(N_1, p - \delta), \text{Binom}(N_2, p + \delta)$
- $M_2 : \text{Binom}(N_1, p + \delta), \text{Binom}(N_2, p - \delta)$

where N_1 and N_2 are the observed number of healthy and psychiatric participants, respectively; p is the latent probability of a participant failing one or more attention checks; and δ is an offset parameter specifying the hypothesized difference in failure rates between groups. Thus, M_0 assumes equal rates of failures between healthy and psychiatric participants, whereas M_1 and M_2 assume greater and lower rates, respectively, for psychiatric participants compared to healthy participants. The common latent probability, p, was set to the observed average rate across the two groups. The offset, δ, was set to 0.05. This

value was selected because it signifies a difference in portions of $\Delta p = 0.1$, corresponding to a small effect for a difference in proportions test ($h = 0.2$; [29]).

Results

In total, 16 of 65 (24.6%) participants failed one or more attention checks. Interestingly, the overall proportion of flagged participants was similar to that observed for the original study. Subdivided by group, 6 of 20 healthy participants (30%) and 10 of 45 psychiatric participants (22%) were flagged for C/IE responding. Unsurprisingly, given the modest sample size, the difference between the two proportions was not significantly different from zero ($z = 0.672$, $p = 0.502$). Across the six infrequency items, the average failure rate was 8.5% (range: 0.0% – 19.7%). In contrast, no participant failed either instructed item. This result further corroborates previous research, which has found that instructed items are poor measures of C/IE responding [19–21].

Next, we computed the Bayes factor for each pair of candidate models. A model assuming equal rates of failure between healthy and psychiatric participants was 2.88 times more likely than the model assuming greater rates for psychiatric patients. In turn, a model assuming lower rates of failure for psychiatric patients was 1.27 times more likely than the model assuming equal rates. Finally, a model assuming lower rates of failure for psychiatric patients was 3.65 times more likely than the model assuming higher rates for psychiatric patients. Only the final comparison exceeds the cutoff value of 3, which is conventionally treated as the minimal amount of evidence required to treat a model comparison as meaningful.

Discussion

Here we sought to examine whether actual psychiatric patients were equally or more likely to fail infrequency items than healthy controls in settings similar to those experienced by online participants (i.e., completing an experiment online, on a computer in one's home or otherwise chosen environment). Although the small sample precludes any definitive conclusion, it is noteworthy that the model least consistent with the data was the one where psychiatric patients were more likely to fail infrequency-item attention checks. Indeed, a model in which healthy controls failed attention checks at a greater rate than patients was credibly preferred to the alternative. This preliminary finding may reflect differences in motivation between patients and controls for participating in psychiatric research. 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 therefore may be motivated to provide higher-quality responses. Regardless, further research is required to examine whether this preliminary finding holds in larger samples and other testing contexts.

Supplementary references

1. Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P. & Dolan, R. Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience* **32,** 5833–5842 (2012).
2. Katahira, K. The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87,** 31–45 (Dec. 2018).
3. Sugawara, M. & Katahira, K. Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific reports* **11,** 1–13 (2021).
4. Litman, L. *New Solutions Dramatically Improve Research Data Quality on MTurk* https://www.cloudresearch.com/resources/blog/new-tools-improve-research-data-quality-mturk/. (Accessed: 2021-02-23).
5. Hauser, D. *et al.* Evaluating CloudResearch's Approved Group as a Solution for Problematic Data Quality on MTurk (2021).
6. Letzter, R. *A teenager on TikTok disrupted thousands of scientific studies with a single video* https://www.theverge.com/2021/9/24/22688278/tiktok-science-study-survey-prolific. Accessed: 2022-10-17. Sept. 2021.
7. *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: 2022-10-17.
8. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69,** 1204–1215 (2011).
9. Litman, L., Robinson, J. & Abberbock, T. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* **49,** 433–442 (2017).
10. 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).
11. De Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods* **47,** 1–12 (2015).
12. Youngstrom, E. A., Murray, G., Johnson, S. L. & Findling, R. L. The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment* **25,** 1377–1383 (2013).
13. Watson, D. *et al.* The development of preliminary HiTOP internalizing spectrum scales. *Assessment* **29,** 17–33 (2022).
14. Ashton, M. C. & Lee, K. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review* **11,** 150–166 (2007).
15. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *Neuroimage* **92,** 381–397 (2014).

16. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *elife* **5,** e11305 (2016).

17. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python* in *Proceedings of the 9th Python in Science Conference* **57** (2010), 10–25080.

18. Wilcox, R. R. & Rousselet, G. A. A guide to robust statistical methods in neuroscience. *Current protocols in neuroscience* **82,** 8–42 (2018).

19. 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).

20. Thomas, K. A. & Clifford, S. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* **77,** 184–197 (2017).

21. 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).

22. Elwert, F. & Winship, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology* **40,** 31–53 (2014).

23. Barch, D. M., Pagliaccio, D. & Luking, K. Mechanisms underlying motivational deficits in psychopathology: similarities and differences in depression and schizophrenia. *Behavioral neuroscience of motivation,* 411–449 (2015).

24. Cohen, R., Lohr, I., Paul, R. & Boland, R. Impairments of attention and effort among patients with major affective disorders. *The Journal of neuropsychiatry and clinical neurosciences* **13,** 385–395 (2001).

25. Culbreth, A., Westbrook, A. & Barch, D. Negative symptoms are associated with an increased subjective cost of cognitive effort. *Journal of abnormal psychology* **125,** 528 (2016).

26. Kane, M. J. *et al.* Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General* **145,** 1017 (2016).

27. Robison, M. K., Gath, K. I. & Unsworth, N. The neurotic wandering mind: An individual differences investigation of neuroticism, mind-wandering, and executive control. *The Quarterly Journal of Experimental Psychology* **70,** 649–663 (2017).

28. Harms, C. & Lakens, D. Making'null effects' informative: statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research* **3,** 382 (2018).

29. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 2013).