nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01640-7

Inattentive responding can induce spurious associations between task behaviour and symptom measures

Received: 12 April 2021

Accepted: 23 May 2023

Published online: 6 July 2023

Check for updates

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

Although online samples have many advantages for psychiatric research, some potential pitfalls of this approach are not widely understood. Here we detail circumstances in which spurious correlations may arise between task behaviour and symptom scores. The problem arises because many psychiatric symptom surveys have asymmetric score distributions in the general population, meaning that careless responders on these surveys will show apparently elevated symptom levels. If these participants are similarly careless in their task performance, this may result in a spurious association between symptom scores and task behaviour. We demonstrate this pattern of results in two samples of participants recruited online (total $N = 779$) who performed one of two common cognitive tasks. False-positive rates for these spurious correlations increase with sample size, contrary to common assumptions. Excluding participants flagged for careless responding on surveys abolished the spurious correlations, but exclusion based on task performance alone was less effective.

In recent years, online labour markets (for example, Amazon Mechanical Turk (MTurk), Prolific and CloudResearch) have become increasingly popular as a source of research participants in the behavioural 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 multitasking during an experiment. They may also be more likely to use heuristic response strategies with the intention to minimize expenditure of time and cognitive effort (for example, responding randomly on self-report surveys or behavioural tasks). Here we refer to such inattentive or low-effort behaviours 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 behavioural research is to mitigate poor-quality data using the same screening methods that are typically used in in-person data collection (for example, excluding participants who perform at or below chance on behavioural tasks). However, these methods may be specifically inappropriate for online psychiatry studies, as we detail below.

In this Article, we wish to draw special attention to an underappreciated feature of psychiatric research using self-report symptom

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Rutgers-Princeton Center for Computational Cognitive Neuropsychiatry, Rutgers University, Newark, NJ, USA. [3]Department of Psychology, Princeton University, Princeton, NJ, USA. [4]School of Psychological Sciences, Monash University, Melbourne, Victoria, Australia. e-mail: szorowi1@gmail.com

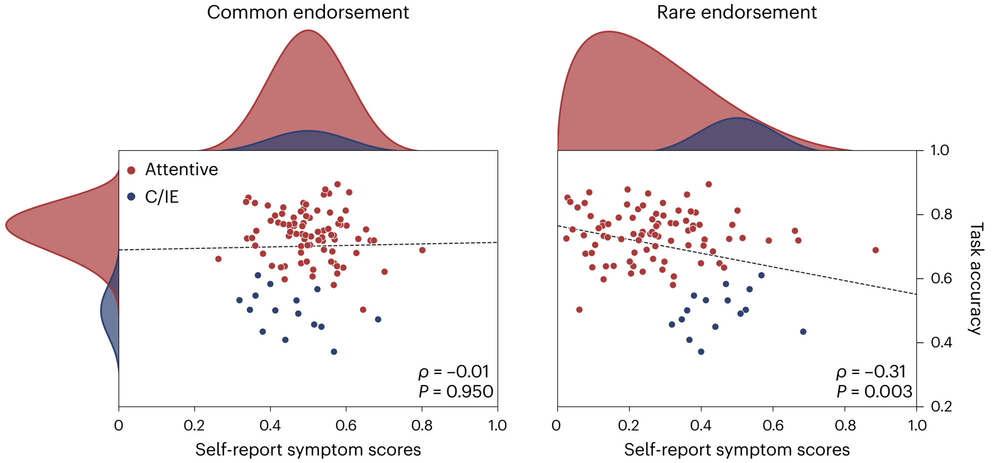

Fig. 1 | Simulated example of how spurious behaviour–symptom correlations can arise when symptom endorsement is rare. Left: when symptoms are moderately common in the general population, C/IE respondents (blue) are indistinguishable from attentive participants (red) in self-report measures (*x* axis,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

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 score 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 behavioural tasks—a common study design in computational psychiatry—this artefact has the potential to induce spurious correlations between symptom self-report scores and task behaviour. That is, while C/IE behaviour 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 behavioural tasks, this can cause experimenters to observe an entirely spurious correlation between greater symptom severity and worse task performance (Fig. 1). A similar effect has been well documented in personality psychology, where the presence of C/IE responding can induce correlations between questionnaires and can bias factor estimation 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 on the basis of 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 behavioural 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 markets, we found that screening based on task behaviour 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 behaviour and self-reported symptoms that are abolished when the data are screened more thoroughly.

Results

Narrative review of task and self-report screening practices

We first sought to what extent recent online studies screened 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 behavioural studies and evaluated whether and how each study performed task and self-report data screening (see Methods for details of the literature search).

Among the studies that we reviewed, approximately 80% (39/49) used at least one method to identify C/IE responding in task behaviour (Table 1). Of these, just over half relied on a single screening method, with considerable heterogeneity in behaviour screening methods across studies. Most common (46% of these studies) was identifying participants whose performance was statistically indistinguishable from chance level on some measure of accuracy. Almost as common (38% of these studies) was screening based on low response variability (that is, 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 the participants are explicitly told which response to select; for example, 'Please select "Strongly Agree"') and infrequency items (in which some responses are logically invalid or exceedingly improbable; for example,

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

Task screening		Self-report screening	
N **= 39 (80%)**		***N*** **= 19 (39%)**	
Measure	**Frequency**	**Measure**	**Frequency**
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%)

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 screening methods such as outlier detection or personal consistency.

In sum, whereas screening for C/IE responding in task behaviour was relatively common for online behavioural 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 behaviour 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-based 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 behavioural task. Specifically, we measured the empirical correspondence between common task-based and survey-based screening methods—as identified in our literature review—so that the results are informative with respect to typical study designs in online psychiatry research.

C/IE participants appear psychiatric when symptoms are rare

To measure the correspondence of screening measures estimated from task and self-report behaviour, we conducted an online behavioural experiment involving a simple decision-making task and a battery of commonly used self-report psychiatric symptom measures (Methods). A final sample of 386 participants from the MTurk (*N* = 186) and Prolific (*N* = 200) online labour markets completed a probabilistic reversal-learning task and five self-report symptom measures. The reversal-learning task required the participants to learn through trial and error which of three options yielded a reward most often; it was modelled after similar tasks used to probe reinforcement-learning deficits in psychiatric disorders[21,22]. The self-report measures were the Seven-Up (7-up), which measures symptoms of hypomania; the Seven-Down (7-down), which measures symptoms of depression; the Generalized Anxiety Disorder-7 (GAD-7), which measures generalized anxiety symptoms; the Behavioral Inhibition and Behavioral Activation Scales (BIS/BAS), which measure reward and punishment motivations; the Snaith–Hamilton Pleasure Scale (SHAPS), which measures anhedonia symptoms; and the Penn State Worry Questionnaire (PSWQ), which measures worry symptoms. These measures were chosen on the basis of previous literature to have a variety of expected response distributions (symmetric and asymmetric). In line with current best-practice recommendations[23], each self-report instrument included one 'infrequency' item that could be used to identify C/IE responses in the survey data (see Methods for a list of infrequency items). The entire experiment (surveys and task) was designed to require ten minutes on average to complete (observed mean, 10.28 minutes). To minimize any influence of fatigue on survey responding, the 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 behaviour 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 MTurk than on Prolific for both task data (MTurk: *N* = 18/186; Prolific: *N* = 8/200; two-tailed, two-sample proportions test: $z(384) = 2.224$; $P = 0.026$; $h = 0.230$; 95% confidence interval (CI), (0.006, 0.107)) and survey data (MTurk: 50/186; Prolific: 35/200; two-tailed, two-sample proportions test: $z(384) = 2.223$; $P = 0.026$; $h = 0.227$; 95% CI, (0.011, 0.176)).

We hypothesized that spurious behaviour–symptom correlations may emerge due to a mean-shift in the average level of symptom endorsement in participants engaging in C/IE responding relative to attentive participants. In turn, a mean-shift is expected to occur when the overall rate of symptom endorsement is low; that is, comparably higher scores are more likely for C/IE participants responding at random on a questionnaire with a right-skewed score distribution. 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 the most positively skewed (7-up, 7-down and GAD-7; Fig. 2). In contrast, where there were 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 the 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 (two-tailed, paired-samples *t*-test: $t(6) = -3.689$; $P = 0.021$; $d = 1.506$; 95% CI, (−0.048, −0.141)).

Second, we quantified the degree to which participants responded to self-report symptom surveys in a stereotyped fashion; that is, we determined whether 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 the self-report data (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 (for example, low on one symptom scale versus high on another). We observed a credible difference between the average values of this intercept for the two groups (Δintercept = −0.67; 95% highest density interval, (−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

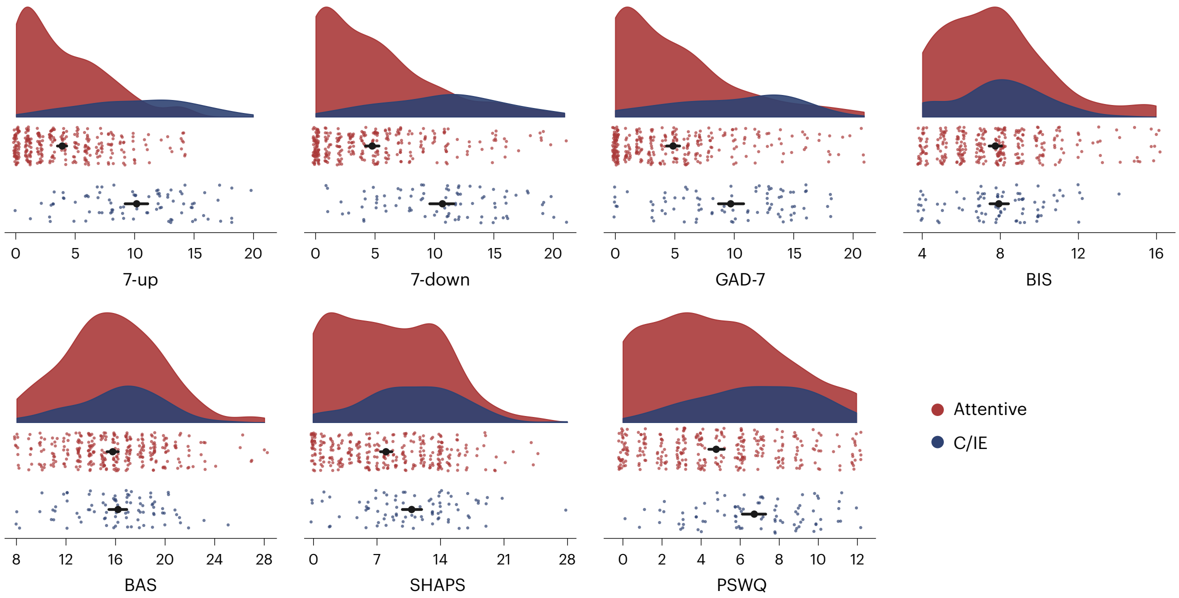

Fig. 2 | Raincloud plots of total symptom scores in attentive (*N* = 301; red) and C/IE (*N* = 85; blue) participants. Each coloured dot represents the symptom score for one participant. The black circles indicate the average score within each group (the error bars denote 95% bootstrap CIs). The shaded plots represent the distribution of scores for each group of participants. The scales are ordered approximately according to their estimated skew (Table 2) from top left (7-up) to bottom right (PSWQ). The average level of symptom endorsement is most markedly different between groups in symptom measures with the lowest overall rates of endorsement.

Table 2 | Descriptive statistics of the self-report symptom measures between attentive and C/IE participants

		Total score				Cronbach's *α*		Percentage at clinical cut-off (%)	
Subscale	Skew	Attentive	C/IE	*t*	*P*	Attentive	C/IE	Before	After
7-up	0.806	3.9	10.2	−13.312	<0.001	0.84	0.84	13.0	4.0
7-down	0.759	4.8	10.7	−9.987	<0.001	0.94	0.88	17.4	9.3
GAD-7	0.753	4.9	9.7	−7.881	<0.001	0.92	0.87	25.9	17.3
BIS	0.780	7.7	7.9	−0.542	0.612	0.83	0.62	–	–
BAS	0.171	15.7	16.2	−0.912	0.357	0.84	0.71	–	–
SHAPS	0.256	8.0	10.8	−4.043	<0.001	0.90	0.81	17.9	14.6
PSWQ	0.193	4.8	6.7	−4.784	<0.001	0.93	0.81	7.3	7.0

Skew is the empirical skewness of the distribution of total symptom scores. Total score is the average symptom score across attentive and C/IE participants. Scores were compared using a two-sample *t*-test (d.f. = 384; *α* = 0.05; two-tailed, not corrected for multiple comparisons). Cronbach's *α* is a measure of response consistency, where values closer to 1 indicate greater consistency in responses. Percentage at clinical cut-off is the percentage of participants reaching the threshold for clinical symptomology before and after screening based on the infrequency measure. The BIS/BAS scales do not have clinical thresholds.

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 (that is, depression and anhedonia, anxiety and behavioural inhibition).

Finally, we compared the proportion of participants meeting the cut-off for clinical levels of psychopathology before and after excluding participants on the basis of 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 the exclusion of C/IE respondents. We observed a similar pattern for both major depressive disorder (MDD) and anxiety (population prevalence estimates of 7% and 5%, respectively[11,30,31]). Interestingly, the proportion of participants meeting the threshold on the GAD-7 was elevated compared with previous literature. We suspect that this may reflect elevated rates of state anxiety during the COVID-19 pandemic[32], when these data were collected. 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].

Low agreement between task and self-report screening measures

We next evaluated the degree of correspondence between behavioural and self-report screening measures to determine whether screening on behaviour 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 behaviour and survey responses (see Methods for a description of the

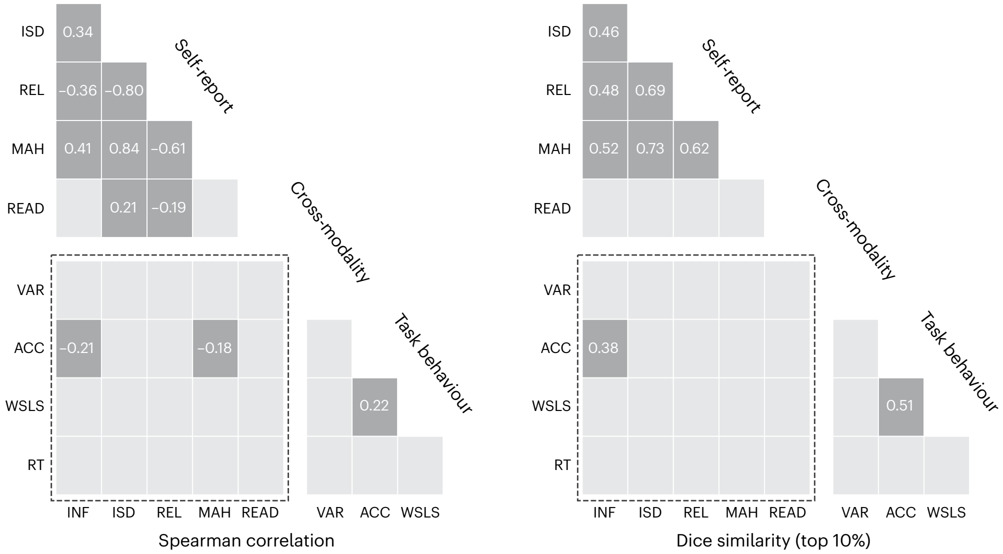

Fig. 3 | Similarity of task and self-report data screening measures. Each tile corresponds to the Spearman rank correlation (left) or Dice similarity coefficient (right) between two screening measures across participants (*N* = 386). The similarity indices are thresholded such that only the magnitudes of statistically significant associations (permutation test, *P* < 0.05, two-sided, corrected for multiple comparisons) are shown. (Unthresholded values are presented in Supplementary Tables 3–5.) Cross-modality correlations between task (*y* axis) and self-report screening measures (*x* axis) are in the dashed rectangle. INF, infrequency item; 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.

measures). To measure the degree of correspondence between these behavioural 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. The resulting pairwise similarity matrices are presented in Fig. 3 (left). After correcting for multiple comparisons, there were few significant correlations between the behavioural and self-report screening measures. Only choice accuracy showed significant associations with any self-report measure (specifically, the infrequency and Mahalanobis 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 (Fig. 3, right). This approach quantifies the degree of overlap between the sets of would-be excluded participants based on different screening measures under a common exclusion rate. Though some measures have relatively clear threshold cut-offs (for example, 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. (The 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 (Supplementary Table 5)). The 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 Dice similarity 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 excluding participants solely on the basis of poor behavioural performance—the most common approach in online studies—is unlikely to identify participants who engage in C/IE responding on self-report surveys.

C/IE responding yields spurious symptom–behaviour correlations

Here we examine the potential consequences of screening only on task behaviour 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 behaviour and self-reported psychiatric symptom severity.

For each participant, we computed both descriptive and computational-model-based measures of behaviour on the reversal-learning task (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 a measure of behaviour under four different conditions: no screening, screening only on task behaviour (that is, only participants whose choice accuracy was above chance), screening only on self-report responses (that is, only participants who responded correctly on all infrequency items) or both. The resulting pairwise behaviour–symptom correlations following each screening procedure are presented in Fig. 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 behavioural 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.

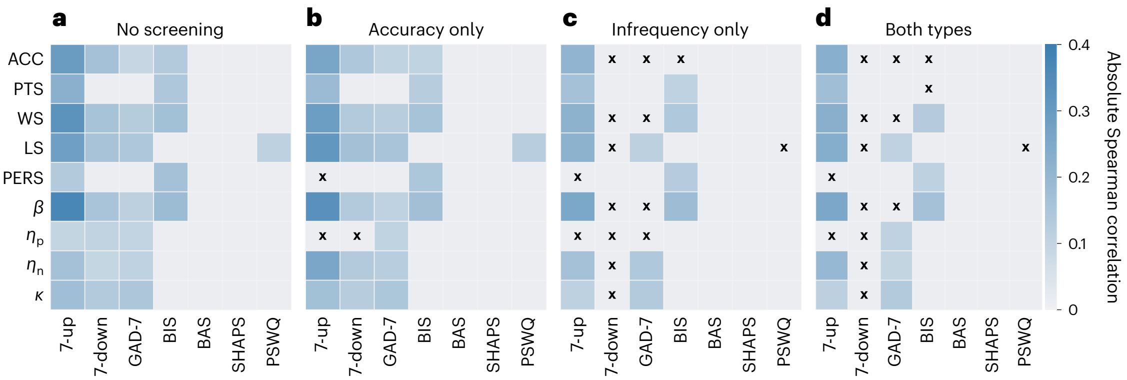

Fig. 4 | Absolute Spearman rank correlations between task behaviour (*y* axis) and symptom measures (*x* axis) under different regimes of data screening and participant exclusions. **a**, No screening—no exclusions ($N = 386$). **b**, Accuracy only—exclusions based on chance-level performance in the reversal-learning task ($N = 352$). **c**, Infrequency only—exclusions based on invalid or improbable responses to infrequency items ($N = 301$). **d**, Both types—exclusions based on the previous two measures ($N = 283$). Only statistically significant correlations are shown ($P < 0.05$, two-sided, not corrected for multiple comparisons; signed correlations are shown in Supplementary Fig. 1 and Supplementary Tables 6–9). The black Xs indicate significant correlations abolished under screening. PTS, total points earned; WS, win-stay rate; LS, lose–shift rate; PERS, perseveration errors; η_p, positive learning rate; η_n, negative learning rate.

When no rejections based on C/IE responding were applied (that is, all participants were included in the analysis; Fig. 4a), many significant correlations emerged between measures of task behaviour 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 behavioural inhibition). Consistent with our predictions, the majority of these correlations involved symptom measures with asymmetric score distributions. Attending to only the most skewed measures (that is, 7-up, 7-down and GAD-7), symptom endorsement was correlated with almost every behavioural measure. That is, significant correlations were not restricted only to general behavioural measures often used as proxies for participant effort (for example, accuracy and 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 (for example, anxiety as measured by the GAD-7 and worry as measured by the PSWQ and depression as measured by the 7-down and anhedonia as measured by the SHAPS).

Next, we excluded participants from analysis on the basis of task-behaviour screening (that is, choice accuracy, removing the 7% of participants exhibiting behaviour indistinguishable from chance; Fig. 4b). The pattern of correlations was largely unchanged: we again found many significant correlations between measures of behaviour and asymmetric symptom measures but almost no significant correlations involving symmetric symptom measures. This suggests that rejecting participants on the basis of the most common form of behavioural screening (that is, performance accuracy) had little effect on behaviour–symptom correlations compared with no screening.

In stark contrast, when we rejected participants on the basis of self-report screening (removing the 22% of participants who endorsed one or more invalid or improbable responses on the infrequency items; Fig. 4c), the number of significant correlations was markedly reduced, particularly for several of the most skewed symptom measures (7-down and GAD-7) and proxy measures of task attentiveness (for example, accuracy and inverse temperature). This pattern of correlations was largely similar when rejections were applied on the basis of both task and self-report screening measures (Fig. 4d). We also note that with stricter screening, the remaining significant correlations were mostly but not always weaker (Supplementary Tables 6–9).

These findings suggest that many of the significant behaviour–symptom correlations observed without strict participant screening may indeed be spurious correlations driven by C/IE responding. Importantly, screening based on task behaviour alone did not adequately protect against spurious symptom–behaviour 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 on the basis of 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–behaviour correlations was likely to be 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 the 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 (two-tailed, paired-samples *t*-test: $t(4{,}999) = 262.490$; $P < 0.001$; $d = 3.713$; 95% CI, (0.136, 0.138); Supplementary Fig. 2, compare with Fig. 4a). Thus, the reduction in significant correlations following screening was unlikely to be driven solely by a reduction in statistical power.

We next 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 (that is, 5%, 10%, 15% or 20%) and increased the total number of participants. Across all analyses, we measured the correlation between the 7-down depression scale and learning-rate asymmetry (κ), which we previously identified as probably exhibiting a spurious association. (The following results are not specific to learning-rate asymmetry and generalize to other pairs of variables (Supplementary Fig. 3)).

The outputs of the bootstrapping analysis are presented in Fig. 5. We found that, although estimated correlation magnitudes were independent of sample size (*x* axis, left), the absolute magnitude of the behaviour–symptom correlation increased with the proportion of C/IE participants (different-coloured circles, left). Crucially, we found that 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). This runs counter to a common assumption that larger sample

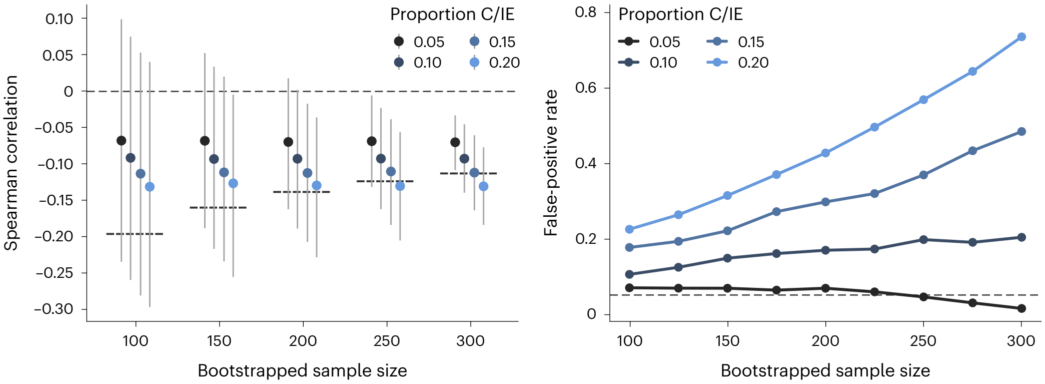

Fig. 5 | False-positive rates for spurious correlations increase with sample size. Left: Spearman rank correlations and 95% bootstrap CIs between *κ* and depression scores (7-down) as a function of sample size and the proportion of C/IE participants. The thick dashed lines indicate the threshold for statistical significance for the Spearman correlation at the corresponding sample size. The markers are jittered along the *x* axis for legibility. Right: false-positive rates for *κ* and depression scores (7-down) as a function of sample size and the proportion of C/IE participants. False-positive rate was calculated as the proportion of bootstrap samples in which the Spearman rank correlation between *κ* and 7-down was statistically significant ($P < 0.05$, two-sided). The horizontal dashed line denotes the expected false-positive rate at $\alpha = 0.05$.

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 behaviour–symptom correlations will become increasingly inflated as the sample size increases.

Findings replicate in a second study with alternative measures

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 whether spurious correlations would remain such a threat.

To evaluate the generalizability of our findings, we therefore conducted a conceptual replication 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 Supplementary Information section B for details). Importantly, the participants were recruited after CloudResearch and Prolific implemented new protocols to improve data quality on their respective platforms. As a final control measure, the participants completed not only self-report symptom measures as before but also personality measures with no hypothesized relationship to model-based planning behaviour on the two-step task.

For the sake of brevity, we report here only the main pattern of findings (all results are reported in Supplementary Information section B). In the replication sample, 55 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 fraction 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 (Supplementary Fig. 7 and Supplementary Table 11). Similarly, we again found that task-based screening and self-report screening measures showed low correspondence (Supplementary Fig. 8 and Supplementary Tables 12 and 13); that is, excluding participants on the basis of poor behavioural 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 (Supplementary Fig. 9). In contrast to our original findings, however, we found that excluding participants on the basis of either self-report or task screening measures was sufficient to abolish 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 behaviour 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.

Patients with depression do not fail attention checks more often

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 hypothesize 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 MDD (see Supplementary Information section C for details). Specifically, $N = 35$ psychiatric patients (confirmed to meet criteria for a diagnosis of MDD through a structured clinical interview) across 45 unique testing sessions and $N = 17$ healthy controls across 20 unique testing sessions, all recruited through the Rutgers-Princeton Center for Computational Cognitive Neuropsychiatry (that is, not via 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%) MDD patients 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 MDD patients (M1), failure rates are greater in MDD patients (M2), and failure rates are greater in healthy participants (M3). The model assuming equal rates of failure between healthy and MDD participants was 2.88 times more likely than the model assuming greater rates for MDD patients. In turn, the model assuming lower rates of failure for MDD patients was 1.27 times more likely than the model assuming equal rates. Finally, the model assuming lower rates of failure for MDD patients was 3.65 times more likely than the model assuming higher rates for MDD patients. Only the final comparison exceeds the cut-off 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 MDD patients are more likely to fail infrequency-item attention checks. These data suggest, therefore, that it is unlikely that individuals with high depression symptom severity were disproportionately flagged for C/IE responding in the main analyses. Accordingly, we tentatively 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, at least in the case of individuals with depression. It remains possible that other psychiatric symptoms might be associated with a different pattern of results.

Discussion

In this study, 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 behaviour 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. Careless responding on tasks used to measure cognitive constructs can then masquerade as a correlation between individual differences in these constructs and symptom dimensions. We found repeated evidence for this pernicious pattern in two samples of 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 behaviour alone was often insufficient to identify participants engaging in C/IE responding and prevent 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 on the basis of responding on surveys, a practice that abolished many spurious behaviour–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 probably include features of the experiment (for example, task difficulty and monetary incentives), facets of the participant (for example, subjective effort costs, intrinsic motivation and conscientiousness) and features of the online labour market itself (for example, opportunity costs and repercussions for careless responding).

Viewed from the perspective of effort expenditure, our results suggest that participants appraised the cost–benefit trade-off differently for behavioural tasks and self-report surveys. Specifically, we found that only 7% of participants in the first study were at chance-level performance in the task, whereas more than 22% of participants 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 behaviour and self-report responses. This suggests that the motivation for effortful responding was greater in the behavioural tasks, though precisely why is unclear. One possibility is that we gave participants a monetary incentive for attentive responding only during the tasks (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 behavioural tasks more engaging or the self-report inventory more tedious. Regardless of the reason, this discrepancy reinforces our observations concerning the inadequacy of behavioural-task screening as a stand-alone method for identifying C/IE responding. Since, in general, participants may appraise the costs and benefits of effortful responding differently for behavioural 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 behavioural tasks and self-report surveys. Indeed, there are more ways to respond carelessly than to respond attentively to a task or self-report inventory (for example, random response selection, straight-lining, zig-zagging and acquiescence bias)[9]. The specific response strategy a participant adopts is likely to reflect the idiosyncratic integration of multiple perceived benefits (for example, time saved and effort avoided) and costs (for example, loss of performance bonuses, risk of detection and forfeited pay). 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 (ISD) 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–behaviour 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 (for example, low motivation, effort avoidance and 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 MDD. Though our final sample was small, we did not find evidence that depressed patients were more likely to fail attention checks than healthy controls (if anything, healthy participants were more likely to be flagged by C/IE screening). These results provide preliminary evidence that rigorous C/IE screening is unlikely to result in overcontrol bias. However, further research with larger samples is necessary to validate attention checks in depressed and other patient populations.

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 (for example, 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, one could screen participants with respect to a multitude of measures and remove only the consistently flagged participants, 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 behaviour–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[11,30,32]. (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 behaviour that were consistent with effects found in previous literature[42–44]. For example, we found that 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 behaviour–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 (for example, logging each key/mouse interaction during survey administration to detect form-filling software or other forms of speeded responding[45]). We thus 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 behaviour–symptom correlations for only two tasks and two sets of self-report instruments. 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 behavioural 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 identify some of the issues that we have identified here.

This study highlights the need for more research on the prevalence of C/IE responding in online samples and its interactions with task–symptom correlations. Many open questions remain, including under what conditions task-screening and symptom-screening measures might better correspond, what screening measures are most effective and when, and under what conditions spurious correlations are more likely to arise. For example, we found that screening on task behaviour alone was insufficient to prevent putatively spurious correlations for one task (reversal learning) but was sufficient for another task (the two-step task). This discrepancy may reflect differences in the tasks (for example, the two-step task may be more challenging and thus more sensitive to C/IE responding) or differences in the screening measures (for example, choice accuracy across 90 trials may be a noisier measure than win–stay lose–shift (WSLS) choice behaviour across 200 trials).

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

We conclude with a list of concrete recommendations for future online studies involving correlations between task behaviour and self-report instruments. We note that these recommendations are not limited to computational psychiatry studies but are applicable to any online individual-differences cognitive science research involving similar methods (for example, behavioural economics and psycholinguistics).

Moving forward, we strongly recommend that experimenters employ some form of self-report screening method, preferably one recommended by the best-practices literature (for example, refs. 9,13,16,19,24). Our literature review found that, to date, the majority of online studies assessing behaviour–symptom correlations have not used self-report screening, and our results demonstrate that stand-alone task-behaviour screening is not necessarily sufficient to prevent spurious symptom–behaviour correlations induced by C/IE responding. We therefore 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 were flagged under each measure and what thresholds were used for rejection.

When collecting self-report questionnaire data, we encourage experimenters to use screening methods sensitive to multiple distinct patterns of C/IE responding (for example, random responding, straight-lining and side bias) and, if possible, to log all page interactions (for example, mouse clicks and keyboard presses). We specifically recommend that experimenters use infrequency-item attention checks rather than instructed-item checks, as multiple studies have now shown that online participants are habituated to and circumvent the latter[18–20] (Supplementary Information section B). Participants flagged by suspicious responses on attention-check items should either be excluded from further analysis or be 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, all else equal, symmetrically distributed measures of a given construct should be preferred to asymmetrically distributed measures (though this will often be infeasible given that the prevalence of many psychiatric symptoms in the general population is typically small). 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[46]. 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 (for example, at the time of writing, Reddit and 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.)

More broadly, we encourage experimenters in computational psychiatry to be mindful of the myriad reasons why participants may perform worse on a behavioural task. Wherever possible, researchers are encouraged to design experiments where the signature of some psychiatric syndrome could not also be explained by C/IE responding (for example,[47,48]). Experimenters should also carefully consider whether an online study is truly appropriate for the research question. In particular, if the project aims to study syndromes associated with considerable difficulty in task or survey engagement (for example, severe ADHD or acute mania), symptomatic participants are likely to produce responses that cannot be distinguished from C/IE responding. In such a case, correlational research in online samples is probably not the best approach for the research question. Finally, we conclude by noting that it is preferable to prevent C/IE responding than to account for it after the fact[49]. As such, we recommend that researchers take pains to ensure that their experiments promote engagement, minimize fatigue and confusion, and compensate participants fairly and ethically.

Methods

Experiment

Sample. A total of 409 participants were recruited to participate in an online behavioural experiment in late June through early July 2020. Specifically, 208 participants were recruited from MTurk, and 201 participants were recruited from Prolific. This study was approved by the Institutional Review Board of Princeton University, and all participants provided informed consent. The total study duration was approximately ten minutes per participant. The participants received monetary compensation for their time (rate US$12 per hour), plus an incentive-compatible bonus up to US$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[50]. (Note: this study was conducted prior to the introduction of CloudResearch's newest data-quality filters[51].) Following recent recommendations[52], MTurk workers were not excluded on the basis of 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 (for example, 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.

Data from several participants 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 are provided in Supplementary Table 1. Notably, the participants recruited from MTurk were older (mean difference, 7.7 yr; two-tailed, two-sample t-test: $t(384) = 6.567$; $P < 0.001$; $d = 0.669$; 95% CI, (5.4, 10.0)) and included fewer women (two-tailed, two-sample proportions test: $z(384) = 2.529$; $P = 0.011$; $h = 0.258$; 95% CI, (0.030, 0.228)).

Experimental task. The 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, the participants were presented with three choice options and were required to choose one. After their choice, the 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 a reward with 80% probability; the subordinate options yielded a reward with only 20% probability. The dominant option changed randomly to one of the two previously subordinate options every 15 trials. The 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 the participants as a fishing game in which each choice option was a beach scene made distinguishable by a coloured surfboard with unique symbol. The participants were told they were choosing which beach to fish at. Feedback was presented as either a fish (1 point) or trash (0 points). The participants were instructed to earn the most points possible by learning (through trial and error) and choosing the best choice option. The 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 Supplementary Information section A for the complete instructions). Prior to beginning the experiment, the 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[53] and distributed using custom web-application software. All experiment code is publicly available. A playable demo of the task is available at https://nivlab.github.io/jspsych-demos/tasks/3arm/experiment.html.

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

Seven-Up/Seven-Down. The 7-up/7-down[54] scale is a 14-item measure of lifetime propensity towards depressive and hypomanic symptoms. It is an abbreviation of the General Behavior Inventory[55], wherein only items that maximally discriminated between depression and mania were included. The items are scored on a four-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 GAD-7 (ref. 56) is a seven-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 two weeks. The items are scored on a four-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 BIS/BAS[57] are a measure of reward and punishment sensitivity. The original 42-item measure was recently abbreviated to a 14-item measure[58], which we use here. The items are scored on a four-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[59]. Importantly, 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 leftmost and rightmost anchors, respectively.

Snaith–Hamilton Pleasure Scale. The SHAPS is a 14-item measure of anhedonia[60]. The items are scored on a four-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[61,62], with only a 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 leftmost and rightmost anchors, respectively.

Penn State Worry Questionnaire. The PSWQ is a measure of worry symptoms[63]. The original 16-item was recently abbreviated to a 3-item measure[64], which we use here. The items are scored on a five-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.

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 (v.2.26)[65]. For all models, four separate chains with randomized start values each took 2,000 samples from the posterior. The first 1,500 samples from each chain were discarded. As a result, 2,000 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.

Validation analyses. To validate the infrequency items as a sensitive measure of C/IE responding, we performed three complementary 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* for each subscale and group. Next, we computed the average difference in Cronbach's *α* between the two groups. We then created a null distribution for this statistic by repeating the same analysis but permuting group membership (that is, randomly assigning participants to either group), holding fixed the sizes of both groups. This procedure was performed 5,000 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:

$$p(y_{ij} = k) = \begin{cases} 1 - \mathrm{logit}^{-1}(\mu_i + \mathbf{x}_j \cdot \boldsymbol{\theta}_i - \mathbf{c}_{j,1}) & \text{if } y = 1 \\ \mathrm{logit}^{-1}(\mu_i + \mathbf{x}_j \cdot \boldsymbol{\theta}_i - \mathbf{c}_{j,y-1}) - \mathrm{logit}^{-1}(\mu_i + \mathbf{x}_j \cdot \boldsymbol{\theta}_i - \mathbf{c}_{j,y}) & \text{if } 1 < y < K \\ \mathrm{logit}^{-1}(\mu_i + \mathbf{x}_j \cdot \boldsymbol{\theta}_i - \mathbf{c}_{j,K-1}) - 0 & \text{if } y = K \end{cases}$$

where μ_i is an intercept for participant i, $\boldsymbol{\theta}_i$ is a vector of latent factor scores for participant i, $\mathbf{x}_j$ is a vector of factor loadings for item j, $\mathbf{c}_j$ 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 non-zero 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 one-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 on others (GAD-7 and 7-up/7-down), we would not expect a consistent non-zero bias in this random intercept term for an attentive participant.

Clinical cut-offs. 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 cut-off score of 10 or higher[11,31]. Though the 7-up/7-down scales do not have firmly established clinical cut-offs, recent work has suggested a cut-off score of 12 or higher[66], which we use here. Finally, the original authors of the SHAPS recommended as a cut-off a score of 3 or more when the items are binarized (1, 'Strongly disagree' or 'Disagree'; 0, 'Strongly agree' or 'Agree'). We use this scoring approach in Table 2.

Correspondence of screening measures. To measure the correspondence of task-based and self-report-based screening measures, we estimated a number of standard measures of data quality from each participant's task behaviour (four in total) and self-report responses (five in total). Beginning 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, we piloted the infrequency items 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: ISD. The ISD is an estimate of a participant's response consistency on a self-report measure[67], defined as:

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

where y_i is a participant's response to item i, $\bar{y}$ 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 distance. The Mahalanobis distance is a multivariate outlier detection measure that estimates how dissimilar a participant is relative to all others. For a participant i, the Mahalanobis distance (D) is defined as:

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

where $(\mathbf{X}_i - \bar{\mathbf{X}})$ represents the vector of mean-centred item responses for participant i and $\Sigma_{\mathbf{XX}}^{-1}$ represents the inverted covariance matrix of all items. Greater Mahalanobis distance 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. Lower 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: WSLS. 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 or switch) was predicted by the previous trial's outcome (reward or 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 200 ms. 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% 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 (that is, Spearman's correlation or 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[68].

Correlations between behaviour and symptom measures. To quantify the effects of both task and self-report data screening on behaviour–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 behaviour 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 behaviour, the risk-sensitive temporal difference learning model[69]. In this model, the expected value of a choice option, $Q(s)$, is learned through a cycle of choice and reward feedback. Specifically, following a decision and reward feedback, the value of the chosen option is updated according to:

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

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

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

where r_t is the observed reward on trial t. 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:

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

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 behaviour as a function of three parameters (β, η_p and η_n).

We fit the reinforcement learning model to each participants' choice behaviour using Stan (the details are given above). Notably, 11 participants (3% of the sample) had parameter estimates with poor convergence (that is, $\hat{R} > 1.1$); their parameters were removed from the correlation analysis. Participants' parameters were fit individually (that is, not hierarchically) to prevent bias during parameter estimation from partial pooling between attentive and C/IE participants. Parameters were sampled using non-centred parameterizations (that is, 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 behaviour–symptom correlation analyses.

We confirmed that the model adequately fit the participants' choice behaviour through a series of posterior checks (Supplementary Fig. 5). In particular, we confirmed that the model recapitulated the group-average learning curves for each block of the experiment. Moreover, we confirmed that the model was able to recover the choice accuracy for each participant reasonably well.

The model-based measures included for analysis were choice sensitivity (β), positive learning rate (η_p), negative learning rate (η_n) and learning rate asymmetry ($\kappa = \frac{\eta_p - \eta_n}{\eta_p + \eta_n}$, the normalized difference between η_p and η_n). We chose these measures because they have been previously used to assess performance in clinical samples[22,42,70,71].

Correlation analysis. Behaviour–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[72] 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 behavioural 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.

Literature review

To characterize common data screening practices in online computational psychiatry studies, we performed a narrative literature review[73]. We identified studies for inclusion through searches on Google Scholar using permutations of query terms related to online labour platforms (for example, 'Mechanical Turk', 'Prolific' and 'online'), experimental paradigms (for example, 'experiment', 'cognitive control' and 'reinforcement learning') and symptom measures (for example, 'psychiatry', 'mental illness' and '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 to identify papers by recruitment method (for example, MTurk). We included in the review studies that (1) recruited participants online through a labour platform, (2) measured behaviour on at least one experimental task and (3) 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, as well as the search terms used to find them, are included in the GitHub repository for this study.

Two of the authors (S.Z. and 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 on the basis of previous taxonomies of screening methods (for example, ref. 9). In addition, we assessed the total number of screening measures each study used and whether monetary bonuses were paid to the participants. This review was not meant to be systematic but instead to provide a representative overview of common practices in online behavioural studies.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are openly available on GitHub at https://github.com/nivlab/sciops.

Code availability

All code for data cleaning and analysis associated with this study is available at https://github.com/nivlab/sciops. The experiment code is available at the same link. The custom web software for serving online experiments is available at https://github.com/nivlab/nivturk.

References

1. Stewart, N., Chandler, J. & Paolacci, G. Crowdsourcing samples in cognitive science. *Trends Cogn. Sci.* **21**, 736–748 (2017).
2. Chandler, J. & Shapiro, D. Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psycho.* **12**, 53–81 (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. *Curr. Opin. Neurobiol.* **55**, 152–159 (2019).
5. Strickland, J. C. & Stoops, W. W. The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Exp. Clin. Psychopharmacol.* **27**, 1–18 (2019).
6. Enkavi, A. Z. et al. Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proc. Natl Acad. Sci. USA* **116**, 5472–5477 (2019).
7. Kothe, E. & Ling, M. Retention of participants recruited to a one-year longitudinal study via Prolific. Preprint at *PsyArXiv* (2019).
8. Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M. & DeShon, R. P. Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* **27**, 99–114 (2012).
9. Curran, P. G. Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* **66**, 4–19 (2016).
10. Chandler, J., Sisso, I. & Shapiro, D. Participant carelessness and fraud: consequences for clinical research and potential solutions. *J. Abnorm. Psychol.* **129**, 49–55 (2020).
11. Lowe, B. et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Med. Care* **46**, 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**, 108 (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. *Clin. Psychol. Sci.* **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 Alcohol Depend.* **183**, 102–110 (2018).
15. Robinson-Cimpian, J. P. Inaccurate estimation of disparities due to mischievous responders: several suggestions to assess conclusions. *Educ. Res.* **43**, 171–185 (2014).
16. Huang, J. L., Liu, M. & Bowling, N. A. Insufficient effort responding: examining an insidious confound in survey data. *J. Appl. Psychol.* **100**, 828–845 (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. *Behav. Res. Methods* **52**, 2489–2505 (2020).
18. Barends, A. J. & de Vries, R. E. Noncompliant responding: comparing exclusion criteria in MTurk personality research to improve data quality. *Pers. Individ. Differ.* **143**, 84–89 (2019).
19. Thomas, K. A. & Clifford, S. Validity and Mechanical Turk: an assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* **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. *Behav. Res. Methods* **48**, 400–407 (2016).
21. Waltz, J. A. & Gold, J. M. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. *Schizophr. Res.* **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. *J. Abnorm. Psychol.* **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. *J. Bus. Psychol.* **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. *J. Bus. Psychol.* **33**, 559–577 (2018).
25. Maniaci, M. R. & Rogge, R. D. Caring about carelessness: participant inattention and its effects on research. *J. Res. Pers.* **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. *Appl. Psychol.* **67**, 309–338 (2018).
27. Maydeu-Olivares, A. & Coffman, D. L. Random intercept item factor analysis. *Psychol. Methods* **11**, 344–362 (2006).
28. Merikangas, K. R. et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **64**, 543–552 (2007).
29. Merikangas, K. R. & Lamers, F. The 'true' prevalence of bipolar II disorder. *Curr. Opin. Psychiatry* **25**, 19–23 (2012).
30. 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. *Int. J. Methods Psychiatr. Res.* **21**, 169–184 (2012).
31. Hinz, A. et al. Psychometric evaluation of the generalized anxiety disorder screener GAD-7, based on a large German general population sample. *J. Affect. Disord.* **210**, 338–344 (2017).
32. Yarrington, J. S. et al. Impact of the COVID-19 pandemic on mental health among 157,213 Americans. *J. Affect. Disord.* **286**, 64–70 (2021).
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. *Annu. Rev. Sociol.* **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. *Curr. Top. Behav. Neurosci.* **27**, 411–449 (2015).
36. Cohen, R., Lohr, I., Paul, R. & Boland, R. Impairments of attention and effort among patients with major affective disorders. *J. Neuropsychiatry Clin. Neurosci.* **13**, 385–395 (2001).
37. Culbreth, A., Westbrook, A. & Barch, D. Negative symptoms are associated with an increased subjective cost of cognitive effort. *J. Abnorm. Psychol.* **125**, 528–536 (2016).
38. Kane, M. J. et al. Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *J. Exp. Psychol. Gen.* **145**, 1017–1048 (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. *Q. J. Exp. Psychol.* **70**, 649–663 (2017).
40. Kool, W. & Botvinick, M. Mental labour. *Nat. Hum. Behav.* **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. *Behav. Res. Methods* **50**, 1960–1970 (2018).
42. Huang, H., Thompson, W. & Paulus, M. P. Computational dysfunctions in anxiety: failure to differentiate signal from noise. *Biol. 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. *J. Neurosci.* **38**, 7901–7911 (2018).
45. Buchanan, E. M. & Scofield, J. E. Methods to detect low quality data and its implication for psychological research. *Behav. Res. Methods* **50**, 2586–2596 (2018).
46. Emons, W. H. Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Appl. Psychol. Meas.* **33**, 599–619 (2009).
47. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* **6**, 6149 (2015).
48. Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M. & Daw, N. D. Increased and biased deliberation in social anxiety. *Nat. Hum. Behav.* **6**, 146–154 (2022).
49. Ward, M. & Meade, A. W. Applying social psychology to prevent careless responding during online surveys. *Appl. Psychol.* **67**, 231–263 (2018).
50. Litman, L., Robinson, J. & Abberbock, T. Turkprime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods* **49**, 433–442 (2017).
51. Litman, L. *New Solutions Dramatically Improve Research Data Quality on MTurk* (CloudResearch, 2020); https://www.cloudresearch.com/resources/blog/new-tools-improve-research-data-quality-mturk/
52. 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).

53. de Leeuw, J. R. jsPsych: a JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* **47**, 1–12 (2015).
54. 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. *Psychol. Assess.* **25**, 1377–1383 (2013).
55. Depue, R. A. et al. A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. *J. Abnorm. Psychol.* **90**, 381–437 (1981).
56. Spitzer, R. L., Kroenke, K., Williams, J. B. & Lowe, B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* **166**, 1092–1097 (2006).
57. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *J. Pers. Soc. Psychol.* **67**, 319–333 (1994).
58. 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. *Psychol. Assess.* **28**, 429–442 (2016).
59. 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. *Pers. Individ. Differ.* **43**, 295–305 (2007).
60. Snaith, R. et al. A scale for the assessment of hedonic tone: the Snaith–Hamilton Pleasure Scale. *Br. J. Psychiatry* **167**, 99–103 (1995).
61. 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). *J. Affect. Disord.* **99**, 83–89 (2007).
62. Leventhal, A. M. et al. Measuring anhedonia in adolescents: a psychometric analysis. *J. Pers. Assess.* **97**, 506–514 (2015).
63. Meyer, T. J., Miller, M. L., Metzger, R. L. & Borkovec, T. D. Development and validation of the Penn State Worry Questionnaire. *Behav. Res. Ther.* **28**, 487–495 (1990).
64. Kertz, S. J., Lee, J. & Bjorgvinsson, T. Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychol. Assess.* **26**, 1146–1154 (2014).
65. *Stan Modeling Language Users Guide and Reference Manual* (Stan Development Team, 2021); https://mc-stan.org
66. 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. *J. Clin. Child Adolesc. Psychol.* **50**, 579–595 (2020).
67. 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. *Pers. Individ. Differ.* **84**, 79–83 (2015).
68. 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).
69. 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. *J. Neurosci.* **32**, 551–562 (2012).
70. Brolsma, S. C. et al. Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. *Psychol. Med.* **52**, 303–313 (2020).
71. Ritschel, F. et al. Neural correlates of altered feedback learning in women recovered from anorexia nervosa. *Sci. Rep.* **7**, 5421 (2017).
72. Wilcox, R. R. & Rousselet, G. A. A guide to robust statistical methods in neuroscience. *Curr. Protoc. Neurosci.* **82**, 8–42 (2018).
73. Grant, M. J. & Booth, A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info. Libr. J.* **26**, 91–108 (2009).

Acknowledgements

We thank A. Norbury, A. Pike and O. Robinson for helpful discussion. The research reported in this article was supported in part by the National Institute of Mental Health (R01MH119511; Y.N.) and by the National Center for Advancing Translational Sciences (UL1TR003017; Y.N.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. S.Z. was supported by an NSF Graduate Research Fellowship. D.B. was supported by an Early Career Fellowship from the Australian National Health and Medical Research Council (no. 1165010). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

Author contributions

S.Z.: conceptualization (equal); software development (lead); data collection—online (lead); formal analysis (lead); writing—original draft (lead); writing—review and editing (supporting); visualization (lead). J.S.: software development (supporting); data collection—clinical (lead); writing—review and editing (supporting). Y.N.: writing—review and editing (equal); funding acquisition. D.B.: conceptualization (equal); software development (supporting); data collection—online (supporting); formal analysis (supporting); writing—review and editing (equal); visualization (supporting).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01640-7.

Correspondence and requests for materials should be addressed to Samuel Zorowitz.

Peer review information *Nature Human Behaviour* thanks Xiaosi Gu, Jonathan Roiser and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Samuel Zorowitz

Last updated by author(s): Apr 23, 2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The experimental task was programmed in jsPsych (http://jspsych.org/) 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.
Data analysis	Data were analyzed using the python programming language and the Stan probabilistic programming language (http://mc-stan.org/). All code is available at https://github.com/nivlab/sciops (including details about software package versions).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data are publicly available at https://github.com/nivlab/sciops

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	quantitative, experimental, cross-sectional design
Research sample	Participants were users of the Amazon Mechanical Turk and Prolific online labor platforms. Of the N=779 participants included for analysis, N=385 self-identified a man N=385 self-identified as a woman, and N=9 participants wished to withhold gender information. N=561 participants identified as Caucasian/white. (Full demographic information are included in the supplement of the manuscript.) Users of Amazon Mechanical Turk and Prolific are overall representative of the US population with respect to gender and ethnicity, but tend to be younger on average (https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/). Online participants were recruited so as to study C/IE responding in this increasingly popular convenience sample.
Sampling strategy	Participants were recruited online from the Amazon Mechanical Turk and Prolific labor platforms. Participants were required to be from the United States or Canada. We aimed for a final sample size of around N=400 per experiment in order to have 80% power to detect behavior-symptom correlations with effect size r=0.15. The sampling procedure was convenience sampling.
Data collection	Participants were recruited online from the Amazon Mechanical Turk and Prolific labor platforms. Participants completed self-report questionnaires and an experimental task via web browser on computers in their homes. The experimenters did not have direct communication with the participants during data collection. All payments were mediated by the Amazon Mechanical Turk and Prolific platforms. Experimenters were not present during data collection. Participants, but not experimenters, were blinded to study hypotheses.
Timing	All participants were recruited between late June through early July, 2020 (original study) and in February, 2022 (replication study).
Data exclusions	Of the N=809 participants recruited, a total of N=30 participants were excluded. N=3 participants were excluded for missing data. N=27 participants were excluded for participating in the study twice (once via Amazon Mechical Turk and then Prolific).
Non-participation	no participants dropped out or withdrew participation
Randomization	There was no experimental randomization in this study. All participants completed the same self-report questionnaires and experimental task.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Participants were recruited online from the Amazon Mechanical Turk and Prolific labor platforms. Participants were required to be from the United States or Canada. Participants were not required to have any history of mental illness to participate, and payment was not conditioned on endorsing symptomatology on the self-report symptom inventories. As such, we do not anticipate issues of selection bias.

Ethics oversight	Institutional Review Board of Princeton University (#5291)

Note that full information on the approval of the study protocol must also be provided in the manuscript.