Running head: Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Noam Sarna[1], Reuven Dar[1], & Matan Mazor[2]

[1] School of Psychological Sciences, Tel Aviv University

[2] All Souls College and Department of Experimental Psychology, University of Oxford

Author note

Correspondence concerning this article should be addressed to Noam Sarna, Tel Aviv, Israel 69978. E-mail: noamsarna@mail.tau.ac.il

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Abstract

Large-scale online studies with healthy adults have documented consistent associations between transdiagnostic psychiatric traits and metacognitive biases. Here, analysis of existing and new large-scale datasets reveals that such correlations are largely driven by surface-level dimensions of questionnaire-filling behaviour: systematic rating biases and inattentive responding. Specifically, a bias to report positive or negative values in self-report scales generalizes to confidence ratings, producing spurious correlations between the two. Additionally, systematic over-confidence among inattentive responders produces spurious positive correlations between confidence and the endorsement of rare symptoms. We show that previously identified transdiagnostic dimensions of “anxiety-depression” and “compulsivity and intrusive thought,” both shown to correlate with decision confidence, map neatly onto these two biases of questionnaire-filling behaviour. In a pre-registered experiment, we further show that decision confidence and self-reported obsessive-compulsive tendencies are correlated with independent measures of inattentive and biased responding. Taken together, we find an alarming degree of influence of inattentive and biased responding over both self-report psychiatric measures and confidence ratings. When not accounted for, these factors produce a mirage of apparent metacognitive alterations in mental health. We discuss concrete precautionary measures that are needed to control for these biases.

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Introduction

The last decade in computational psychiatry can be broadly characterized by two prominent trends: the transition to transdiagnostic phenotyping and the proliferation of large online samples (Boldt, Fox, Gillan, & Gilbert, 2024; Fox et al., 2024; Gillan & Daw, 2016; Gillan & Whelan, 2017; Huys, Maia, & Frank, 2016; Rouault, Seow, Gillan, & Fleming, 2018; Seow & Gillan, 2020; Seow, Rouault, Gillan, & Fleming, 2021; Wise & Dolan, 2020; Wise, Robinson, & Gillan, 2023). These trends are linked: transdiagnostic phenotyping strives to define and classify impaired mechanisms across disorders, replacing the traditional focus on disorders as unified, though highly heterogeneous, entities (Insel et al., 2010). In practice, this is often done by having participants complete a large pool of self-report inventories and then using factor analysis to identify a low-dimensional manifold structure in the space of inventory items. Such analysis requires data from large samples, which is made possible by relying on online experimentation (Gillan & Daw, 2016).

The original and most widely used factor analysis of this type, aiming to find a specific psychiatric dimension associated with deficits in goal-directed control, was published by Gillan, Kosinski, Whelan, Phelps, and Daw (2016). In their analysis, three factors emerged from a pool of ten psychiatric questionnaires and were termed Anxious-Depression (AD), Compulsive Behavior and Intrusive Thought (CIT), and Social Withdrawal (SW). These factor labels were derived from the individual items with the highest and most consistent loadings on each factor. In the AD factor, the highest loading items were from questionnaires assessing trait anxiety, apathy and depression; in the CIT factor, from measures of obsessive-compulsive disorder (OCD), eating disorder and alcohol addiction; and in the SW factor, from a social anxiety inventory (Gillan et al., 2016).

In a typical transdiagnostic computational study, once these factors are derived, their relationships with various tasks are assessed. Research looking into metacognition in mental health documented reliable associations between transdiagnostic psychiatric dimensions and confidence biases (that is, biases to be over- or under-confident in one's decisions; Hoven et al., 2023; Rouault et al., 2018; Seow et al., 2021; Seow and Gillan, 2020; Seow et al., 2025; Katyal et al., 2025). A prominent finding in this literature, originally documented in a perceptual discrimination task (deciding which of two briefly presented squares contained more dots), is that higher CIT factor scores were associated with higher decision confidence, and that higher AD factor scores were associated with lower decision confidence (Rouault et al., 2018). This finding has since been replicated in other, independent samples (Benwell et al., 2022; Hoven et al., 2023; Hoven, Rouault, et al., 2023; Katyal et al., 2025; Seow et al., 2025) and extended to a variety of cognitive tasks (e.g., a predictive inference task, Seow & Gillan, 2020; a gamified version of the perceptual-decision making task, Fox et al., 2024; an external reminder-usage task, Boldt et al., 2024).

Confidence abnormalities in psychopathology have attracted much attention as a promising model for interpreting and understanding mental health symptoms, with the transdiagnostic dimensions approach serving as an alternative for the traditional unitary diagnostic framework (for review and discussion see Hoven et al., 2019; Seow et al., 2021; Wise et al., 2023). Here, we suggest that the well-documented associations between metacognition and transdiagnostic dimensions are driven in large by factors related not to mental health, but to the psychometric properties of self-report questionnaires and to response biases. In particular, we propose that the scores and derived factors that make up the widely used psychiatric dimensions reflect not only the substantive phenomena they are meant to measure (i.e., mental health) but also surface-level individual differences in questionnaire-filling behavior. We submit that the same individual differences manifest also in self-reported confidence ratings, ultimately resulting in incorrect conclusions about the relationship between psychiatric dimensions and confidence.

We consider two properties of questionnaire-filling behaviour that can lead to spurious correlations between psychiatric questionnaire scores and decision confidence: *acquiescence* and *inattentive responding*. Both properties can be described in the context of a process model of self-reports (Fig. 1A, leftmost panel). In this model, a questionnaire item (P) induces in a respondent an “internal variable” that corresponds to their level of agreement with the content of the item. This variable is then translated, using a response selection process, to a point on a scale.

Acquiescence is a property of the response selection process, reflecting the tendency of respondents to agree or disagree with self-report items irrespective of its content (Podsakoff et al., 2003; Fig. 1A, upper panels). In this paper we refer to acquiescence as the general tendency to have a rating bias, be it positive or negative. Acquiescence effects have been thoroughly documented, with various methods employed to detect and model them (for review see McGrath, et al., 2010; Weijters et al., 2013). Critically, acquiescence is likely to affect both questionnaire responses and subjective confidence ratings, producing an appearance of a link between decision confidence and symptom severity (Fig 1B, left panel).

An inattentive, or careless, responding style is a feature of the first part of the process model, broadly defined as responding while paying little attention to the content of questionnaire items, thereby failing to consistently generate an internal variable (Meade & Craig, 2012). Inattentive respondents are thought to sample their responses semi-randomly from a nearly uniform distribution (Chandler et al., 2020; Zorowitz et al., 2023; Fig. 1A, rightmost panel). This uniformity leads to a relative increase in the endorsement of symptoms that have lower prevalence in the general population, effectively making inattentive responders appear highly symptomatic (Figure 1B; middle panel). The effect of inattentive responding on correlations with confidence rests on an empirical observation: inattentive responders tend to be overly confident in their responses. In the Results section we provide direct support for this effect, which produces spurious correlations between the endorsement of rare items and decision confidence (Figure 1B, middle panel).

We elaborate on these two independent factors in the Methods section and demonstrate their respective contributions to the reported associations between mental health and metacognition in two large datasets (Rouault et al., 2018; Seow & Gillan, 2020) in the Results section. Finally, analysis of a new dataset with direct measures of inattentive responding and acquiescence reveals that variability in these surface-level properties of questionnaire-filling behaviour may explain a puzzling finding, obtained only in online studies: over-confidence, rather than the well- documented under-confidence, among participants with compulsive or obsessive-compulsive tendencies.

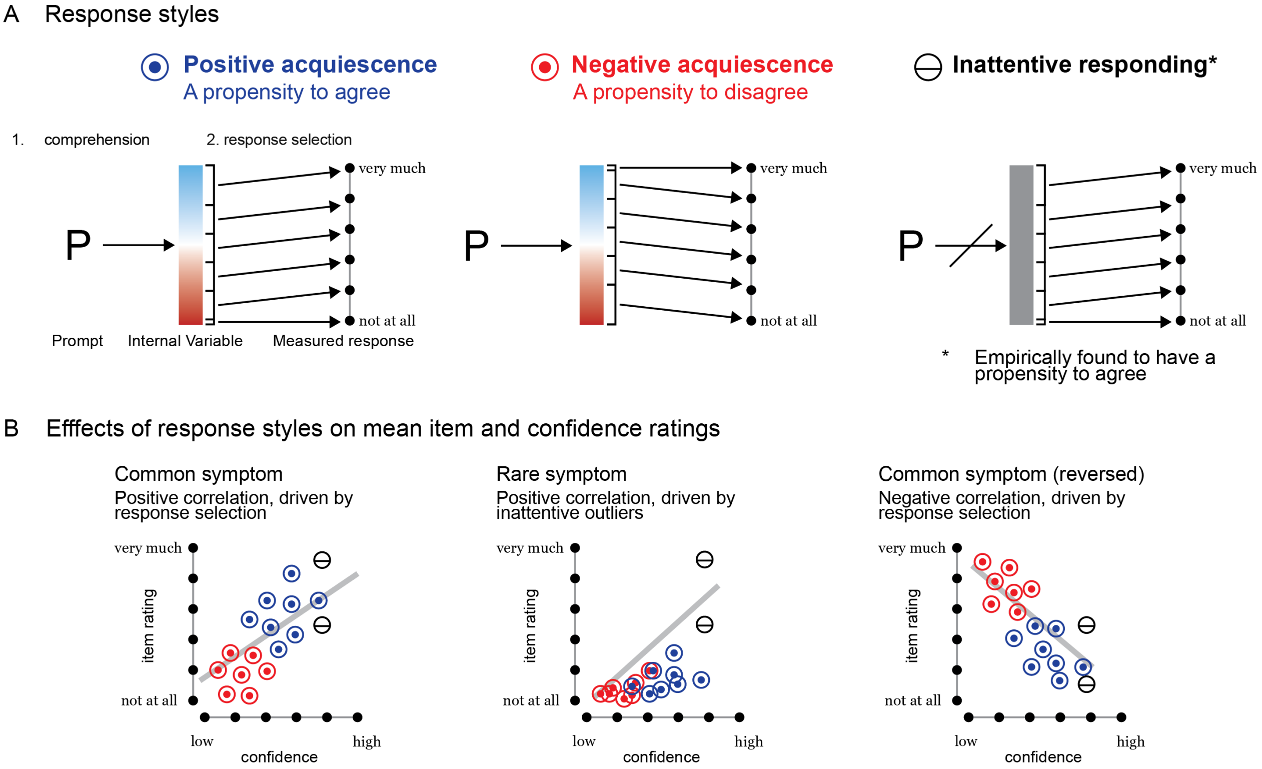

Figure 1 *A schematic illustration of the effects of acquiescence and inattentive responding on item-confidence correlations.* A) We describe the production of a self-report as a two-step process. First, a prompt is read, generating an internal variable that represents a subjective level of agreement. Then, the internal variable is translated to a rating via a response selection process (arrows). We distinguish three prototypical response styles. Positive and negative acquiescence, a feature of the response selection process, correspond respectively to general tendency to agree or disagree with the prompt regardless of item content. Inattentive responding affects both steps of the process: no internal variable is generated, and there is a general tendency to agree. B) The effect of response style on both self-report items and confidence ratings. 'Common symptom' refers to self-report items asking about symptoms with high prevalence in the population (e.g., "I get tired for no reason" SDS, item 10); 'Rare symptom' refers to self-report items asking about symptoms with low prevalence in the population (e.g., "I have the impulse to vomit after meals." EAT item 26); 'Common (reversed)' refers to symptoms with high prevalence in the population which are articulated in a reversed tense (e.g., "I'm mostly happy" STAI item 10).

Results

We start by reporting the reanalysis of two large-scale online metacognition studies, assessing both transdiagnostic dimensions and trial-by-trial confidence ratings. In these studies, participants completed questionnaires for alcohol use (Alcohol Use Disorder Identification Test [AUDIT; Saunders et al., 1993]), apathy (Apathy Evaluation Scale [AES; Marin et al., 1991]), depression (Self-Rating Depression Scale [SDS; Zung, 1965]), eating attitudes (Eating Attitudes Test [EAT-26; Garner et al., 1982]), impulsivity (Barratt Impulsivity Scale [BIS-11; Patton et al., 1995]), obsessive-compulsive tendencies (Obsessive-Compulsive Inventory – Revised [OCI-R; Foa et al., 2002]), schizotypy (Short scales for measuring schizotypy [Mason et al., 2005]), social anxiety (Liebowitz Social Anxiety Scale [LSAS; Liebowitz, 1987]) and anxiety (State-Trait Anxiety [STAI; Spielberger, 1970]). The same participants also rated their confidence in perceptual decisions. In Rouault et al. (2018, Exp. 2; Fig. 2 left panel), participants decided which of two briefly presented boxes had more dots in it and rated their subjective confidence on a 6-point scale. In Seow and Gillan (2020; Fig. 2 right panel), participants positioned a bucket to catch a flying particle and rated their subjective confidence on a 100-point scale. For more detail and an explanation of the study selection rationale, see the Methods section.

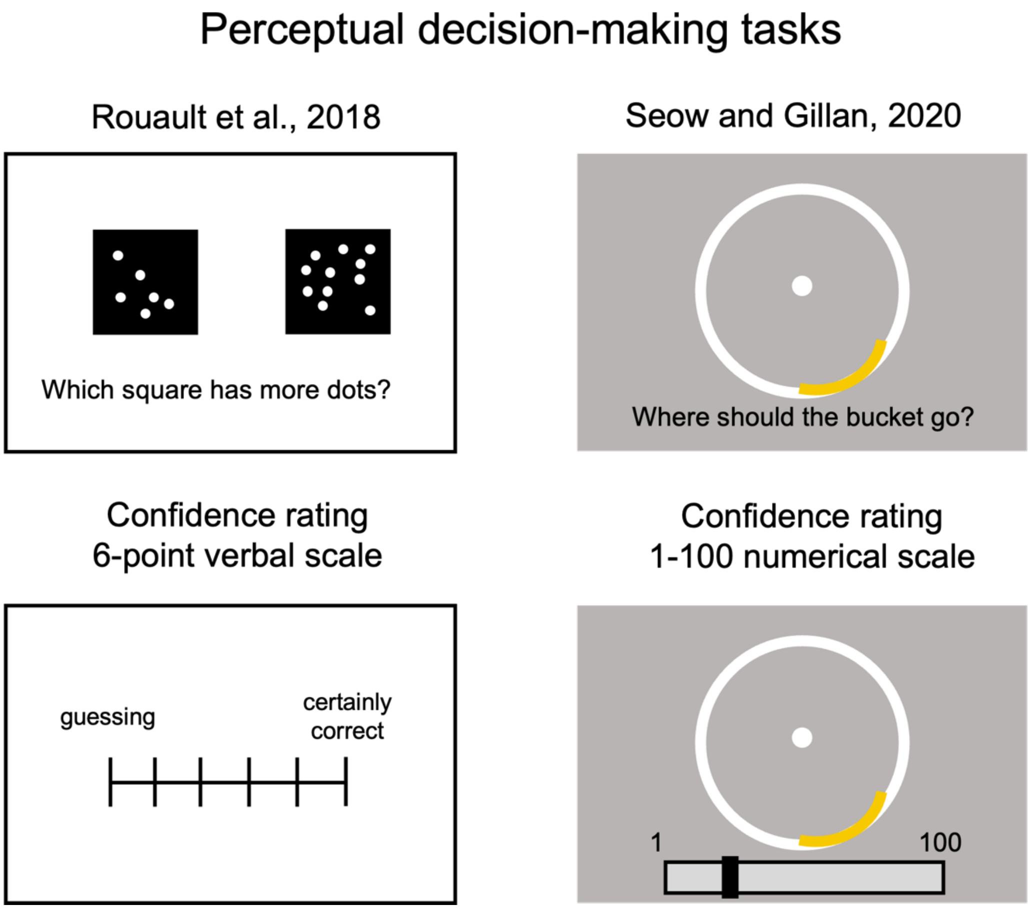

Figure 2 *Illustration of perceptual decision-making tasks re-analyzed in this report.* Left: in the dot comparison task used by Rouault et al. (2018), participants decided which of two squares contains more dots and then rated their decision confidence on a 6-point scale. Right: in the predictive inference task used by Seow and Gillan (2020), participants positioned a bucket (yellow arc on the circle edge) to catch a flying particle and then rated their confidence in that they would catch the particle on a 100-point sliding scale.

Analysis 1.1: Testing the effect of acquiescence on confidence rating.

Confidence ratings are similar to psychiatric questionnaire items in that they require participants to translate an internal representation to a number, or a point on a scale. As such, they may be subject to similar biases. For example, participants showing positive acquiescence in their rating of psychiatric items (a tendency to produce high ratings) would also tend to show positive acquiescence in their confidence rating (a tendency to report high confidence). This will affect both their apparent mental health profile and, crucially, their mean self-reported confidence level, producing a spurious correlation between the two (Figure 1B, leftmost panel). To test whether acquiescence plays a role in the association between confidence and psychiatric dimensions, we calculated for each participant their mean confidence rating over all trials in the perceptual decision-making task, and their mean rating over all self-report items across all psychiatric inventories (with reversed items coded using the original unreversed scale; for more details see Methods). As these studies did not include items with neutral content that could be used to independently assess acquiescence (Weijters et al., 2013), we used this mean rating as a proxy for acquiescence.

In Seow and Gillan (2018), there was a positive correlation between mean item rating and mean confidence rating ($r = .33$, 95% CI $[.24, .41]$, $t(435) = 7.18, p < .001$; Fig. 3, right panel), such that higher mean ratings across items were associated with higher mean confidence ratings. A positive correlation was also found in Rouault et al., 2018 ($r = .18$, 95% CI $[.10, .27]$, $t(495) = 4.11, p < .001$). These results can mean at least one of two things: either that psychiatric symptoms, as measured with these questionnaires, are truly associated with higher levels of decision confidence, or that acquiescence in self-report rating scales affects both responses to questionnaire items and confidence ratings, producing a spurious correlation between the two. Our next analysis provides direct support for the second alternative.

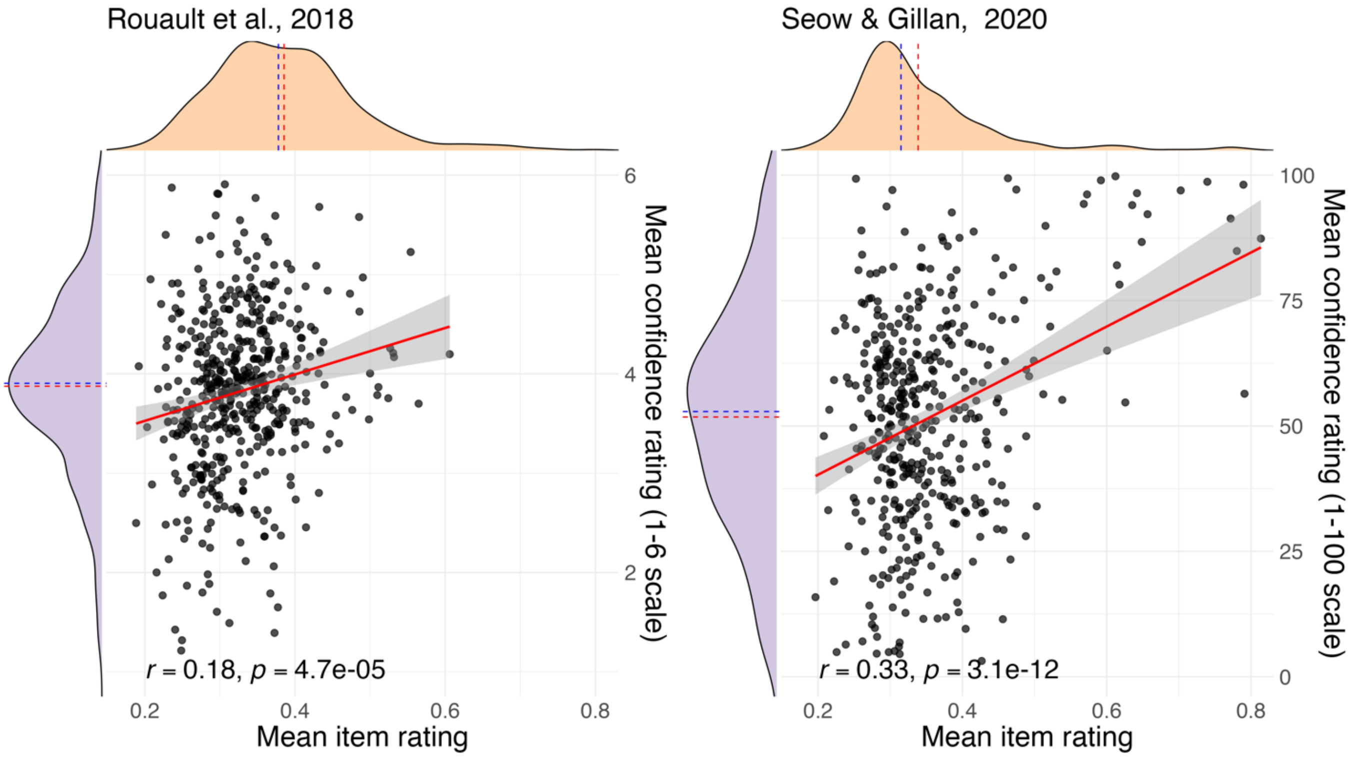

Figure 3 *Correlation between mean item rating and mean confidence rating* in Rouault et al., 2018 (left panel) and Seow & Gillan, 2020 (right panel). Each point represents a single participant's mean item rating across all inventories and mean confidence rating across all trials. Item-ratings for reversed items were recoded to a left-to-right space (as they were shown to the participant). Item ratings were scaled to a 0–1 range to maintain consistency across inventories with different scales. The red line represents a linear regression fit, and the shaded gray area represents the standard error of the fit. Density plots shown on the y- and x-axis with red dashed lines present the mean and blue dashed lines present the median.

Analysis 1.2: The effect of acquiescence reflected in reversed coded items.

To further assess the magnitude and impact of acquiescence on confidence, we made use of the fact that some questionnaires measuring anxiety (State-Trait Anxiety Inventory, STAI; Spielberger, 1970), impulsivity (Barratt Impulsivity Scale, BIS-11; Patton, Stanford, and Barratt, 1995), depression (Self-Rating Depression Scale, SDS; Zung, 1965) and apathy (Apathy Evaluation Scale, AES; Marin, Biedrzycki, and Firinciogullari, 1991) include reverse-coded items: items that tap into the same cognitive constructs but phrased in opposite ways. For example, items 1 and 2 in the STAI read “I feel pleasant” and “I feel nervous and restless,” respectively (possible answers: Almost never, Sometimes, Often, and Almost always). Item 1 is a reversed item. An answer of “Almost always” to this item is coded as 1, and an answer of “Almost never” is coded as 4. The opposite is true for item 2. Crucially, valid responses to these two items should show opposite trends — low endorsement of pleasantness should be associated with high endorsement of restlessness and vice versa. Conversely, acquiescence is expected to result in an inconsistency between the anxiety scores derived from regular and reversed items, namely high or low endorsement of both pleasantness and restlessness.

Following this rationale, we tested the effect of coding direction (standard or reversed) on the correlation between questionnaire responses and confidence. For each item in the STAI, BIS, SDS and AES, we assessed the correlation between participants’ ratings and their mean confidence level in the decision-making task. In this analysis, items were scored based on their semantic meaning, i.e. reversed items are coded using a reversed scale, as explained above. A true association between the measured construct (in the example above, anxiety) and confidence should produce a similar correlation between item and confidence ratings when considering standard and reversed items. In contrast, acquiescence is expected to produce opposite correlations of confidence with standard compared to reverse coded items. This latter pattern is exactly what we found. In the Seow and Gillan (2020) dataset, standard items were on average more positively correlated with mean confidence ratings (mean *r* across the 43 standard items =

0.13) than were reversed items (mean *r* across the 45 reversed items = -0.04). A t-test comparing the mean of the Pearson correlation coefficients between the two samples was statistically significant ($\Delta M = 0.17$, 95% CI $[0.14, 0.20]$, $t(85.12) = 11.28$, $p < .001$; Figure 4, right panel), with a large effect size (Cohen's d = 2.41, 95% CI [1.85, 2.96]). A similar pattern was observed in the Rouault et al. (2018) dataset, where on average, standard items showed a more positive correlation with mean confidence ratings (mean *r* across the 43 standard items = 0.03) than reversed items (mean *r* across the 45 reversed items = -0.09, $\Delta M = 0.11$, 95% CI $[0.09, 0.14]$, $t(78.29) = 8.98$, $p < .001$; Figure 4, left panel), with a large effect size (Cohen's d = 1.93, 95% CI [1.41, 2.43]). This effect remained significant when accounting for the main (intercept) effect of questionnaire in a mixed-effect model (Seow and Gillan, 2020: t(84.24) = 13.45, p < .001; Rouault et al, 2018: t(85.97) = 8.92, p < .001; see Appendix).

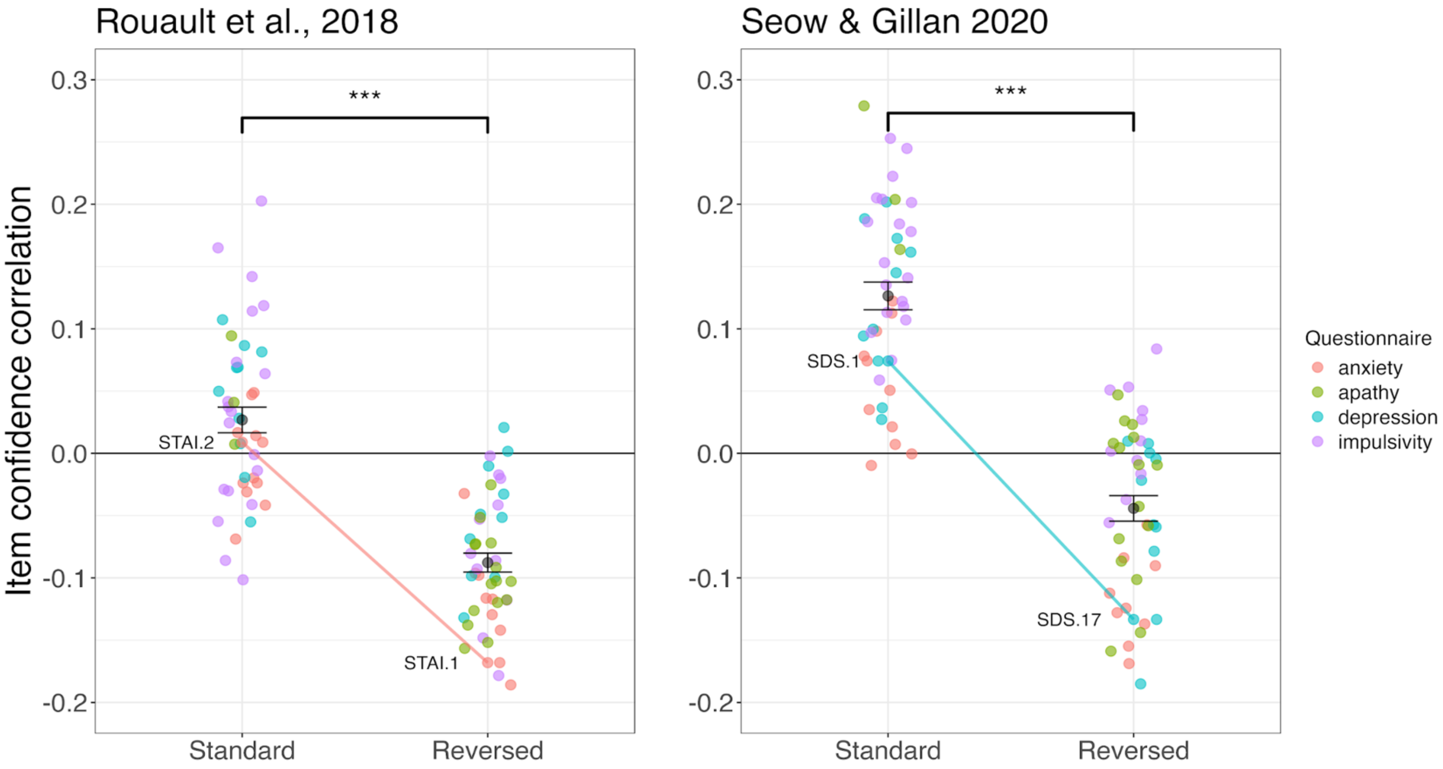

Figure 4 *Relationship between reversed-coded items and item confidence correlation.* Questionnaires: Anxiety - State-Trait Anxiety Inventory (STAI; Spielberger, 1970). Apathy - Apathy Evaluation Scale (AES; Marin et al., 1991). Depression - Self-Rating Depression Scale (SDS; Zung, 1965). Impulsivity - Barratt Impulsivity Scale (BIS-11; Patton et al., 1995). Reference line at y=0 indicates zero item confidence correlation. STAI.1 ('I feel pleasant.') and STAI.2 ('I feel nervous and restless.') are items from the STAI inventory. SDS.1 ('I feel down-hearted and blue.') and SDS.17 ('I feel that I am useful and needed.') are items from the SDS inventory. Upper line and asterisks denote significance (p<0.001) of t-test.

Analysis 2: Testing the effect of inattentive responding with item-level skewness.

The semi-random responses of inattentive responders make them symptomatic on items that are rarely endorsed by attentive responders, that is, on items with a right-skewed response distribution. As a result, a participant who endorses a right-skewed item is more likely to be inattentive than a participant who does not (Chandler et al., 2020). For example, consider an item that describes a rare symptom that is experienced by only 10% of the population. If rated on a five-point scale (ranging from 0 = ‘Almost never’ to 4 = ‘Almost always’), it will receive non-zero ratings from only 10% of attentive participants, but from 80% of inattentive participants who sample their responses uniformly, irrespective of content. Therefore, participants who provide non-zero ratings will be more likely than those who provide zero-ratings to be inattentive responders (see Figure 1B, middle panel).

Given that inattentive participants were previously found to be biased towards using the positive (‘agree’) half of a survey rating scale (Zorowitz et al., 2023), we reasoned that inattentive responders may rate their confidence as higher on average, leading to a spurious positive correlation between the endorsement of rare (right skewed) psychiatric symptoms and decision confidence. Supporting this conjecture, we found a positive correlation between item skewness and its correlation with confidence in both datasets (Seow & Gillan, 2020: $r_s = .71$, $p < .001$, Rouault et al., 2018: $r_s = .55, p < .001$; figure 5, top panels), such that as items are more right-skewed—that is, less frequently endorsed—the correlation between item endorsement and mean confidence increases. The positive relationship between skewness and item-confidence correlations was present in most individual questionnaires (5 out of 8) in Rouault et al. (2018) and in all questionnaires in Seow & Gillan (2020), indicating that this association exists independently of specific questionnaire characteristics (see Appendix).

To elucidate the relationship between item skewness and item-confidence correlations, consider item 2 from the Eating Attitudes Test (EAT-2): “I avoid eating when I am hungry;”

Figure 5, bottom left). Endorsement of this item is significantly correlated with mean confidence ($t(495) = 2.35, p = .019$), but visual inspection suggests that this correlation is largely driven by a small minority of participants who reported “usually” or “always” and also had a high mean confidence rating. This pattern is more suggestive of a spurious correlation due to inattentive responding—participants reporting high agreement with items without paying attention to their content—than of a substantive psychological relationship between self-starvation and confidence.

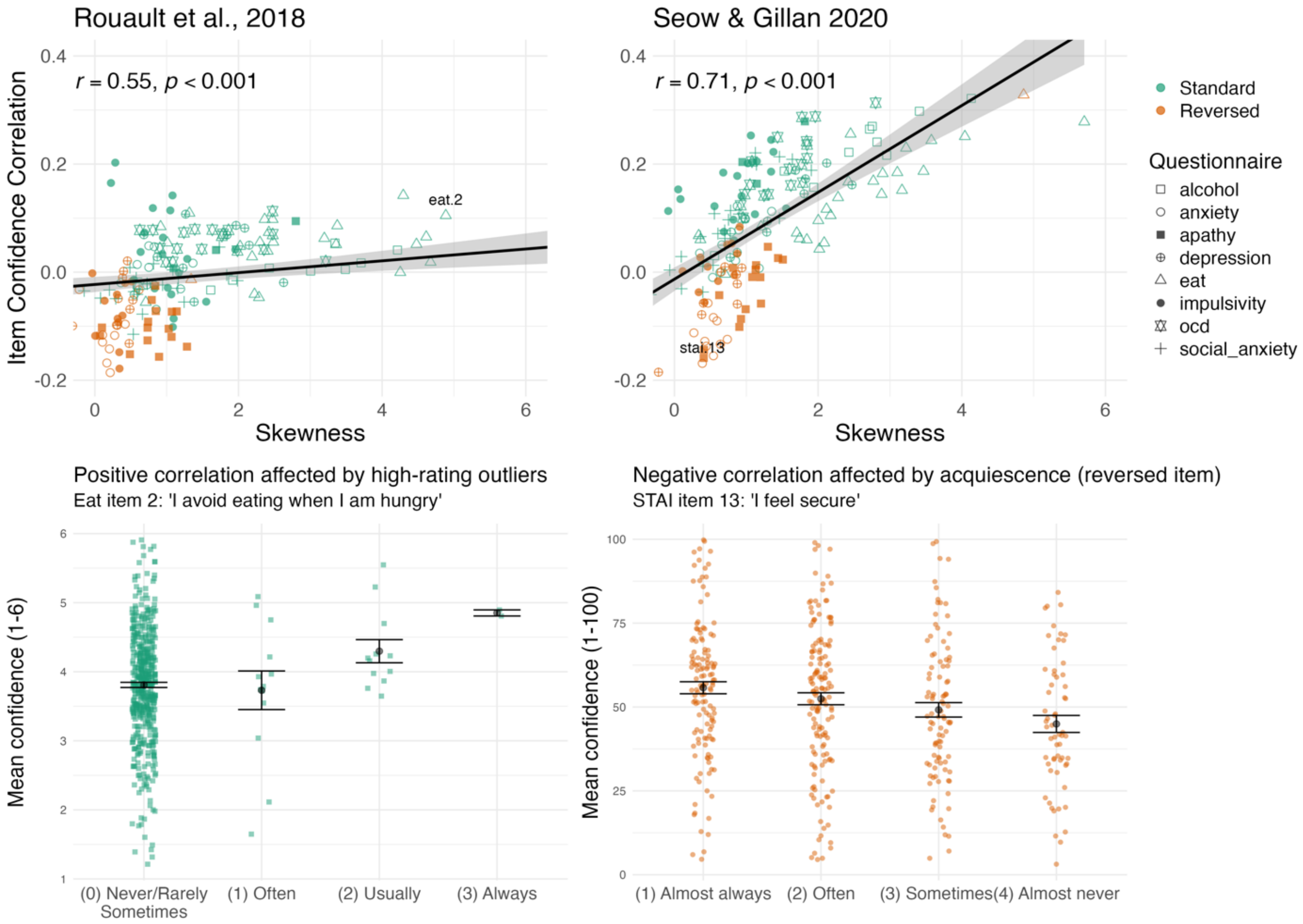
Rouault et al., 2018
$r = 0.55$, $p < 0.001$
Item Confidence Correlation
eat.2
Skewness
Seow & Gillan 2020
$r = 0.71$, $p < 0.001$
stai.13
Skewness
0.4
0.2
0.0
-0.2
0
2
4
6
● Standard
● Reversed
Questionnaire
□ alcohol
○ anxiety
■ apathy
⊕ depression
△ eat
● impulsivity
✡ ocd
+ social_anxiety
Positive correlation affected by high-rating outliers
Eat item 2: 'I avoid eating when I am hungry'
Mean confidence (1-6)
1
2
3
4
5
6
(0) Never/Rarely Sometimes
(1) Often
(2) Usually
(3) Always
Negative correlation affected by acquiescence (reversed item)
STAI item 13: 'I feel secure'
Mean confidence (1-100)
0
25
50
75
100
(1) Almost always
(2) Often
(3) Sometimes
(4) Almost never

Figure 5 *Correlation between item-level skewness and item confidence correlation.* Top panel: each point represents an item from the self-report questionnaires, with the shapes indicating different questionnaires. The x-axis represents the skewness score of each item, which was limited to a range of 0–6 for an easier visualization. The y-axis represents the Pearson correlation coefficient between the item's rating and mean confidence, across individuals. The black line represents a linear regression fit, and the shaded gray area represents the standard error of the fit. Bottom Panel: mean confidence ratings by questionnaire item responses. Left: EAT item 2, "I avoid eating when I am hungry," from Rouault et al. (2018), showing a positive correlation affected by high-rating outliers in the "Usually" and "Always" response category. Right: STAI item 13, "I feel secure," from Seow & Gillan (2020), demonstrating a negative correlation influenced by acquiescence to reversed items. Error bars represent standard errors of the mean. Questionnaires: Alcohol - Alcohol Use Disorder Identification Test (AUDIT; Saunders et al., 1993); Apathy - Apathy Evaluation Scale (AES; Marin et al., 1991); Depression- the Self-Rating Depression Scale (SDS; Zung, 1965); EAT - the Eating Attitudes Test (EAT-26; Garner et al., 1982); Impulsivity- the Barratt Impulsivity Scale (BIS-11; Patton et al., 1995); OCD - the Obsessive-Compulsive Inventory – Revised (OCI-R; Foa et al., 2002); Anxiety - the State-Trait Anxiety Inventory (STAI; Spielberger, 1970); Social anxiety- Liebowitz Social Anxiety Scale (LSAS; Liebowitz, 1987).

Analysis 3: Associations between transdiagnostic dimensional weights with skewness and coding direction.

As discussed in the introduction, the CIT (compulsive behavior and intrusive thought) dimension has been associated with increased mean confidence, whereas the AD (anxious-depression) dimension has been linked to reduced mean confidence (Hoven, Luigjes, et al., 2023; Rouault et al., 2018; Seow & Gillan, 2020). An alternative explanation for these associations is that both psychiatric dimensions and reported confidence are subject to similar response biases, simultaneously influencing their observed relationships. To examine this explanation, we tested the contribution of acquiescence and inattentive responding to the transdiagnostic factor structure itself, irrespective of confidence ratings. For this, we obtained the factor weights of individual items as originally computed by Gillan et al. (2016) as both studies (Rouault et al., 2018; Seow & Gillan, 2020) rely on these weights in their analysis, and since these factors were computed based on a large sample (N=1413). In Figure 6, we plot these item weights against the item skewness with visual coding for reversed items (color-coded in orange) for the CIT and AD dimensions in both datasets.

Two prominent trends emerge from these plots. First, in the CIT dimension (Figure 6, left column) there is a positive correlation between item weight and its skewness, such that more skewed items contribute more to the CIT factor. This association was large and significant in both datasets: in Rouault et al. (2018) ($r_s = .67, p < .001$), and in Seow and Gillan (2020) ($r_s = .68, p < .001$). This finding is consistent with the conjecture that the positive correlation between the compulsivity dimension and confidence was driven, at least in part, by high confidence ratings among inattentive responders. As items become more skewed, the proportion of inattentive participants is expected to exceed that of truly attentive symptomatic participants (Figure 1B, Rare Symptom; see also Chandler et al., 2020). In the appendix we present the results of a simulation, showing that a positive correlation between item skewness and factor weights is unexpected under reasonable assumptions about the link between skewness and diagnostic content. In addition, the highest-loading items were standard rather than reversed items, further indicating that item weights in the CIT factor are driven by item-specific properties, over and above any psychopathology-related characteristics. Second, in the AD dimension (Figure 6, right column), a pattern indicating the contribution of both skewness and reversed items is observed, but in the opposite direction to the case of CIT: reversed items with low skewness load heavily on the AD factor. This pattern is not surprising given the high negative correlation between the weights of these two factors across items ($r = -.58$, 95% CI $[-.67, -.47]$, $t(164) = -9.09$, $p < .001$).

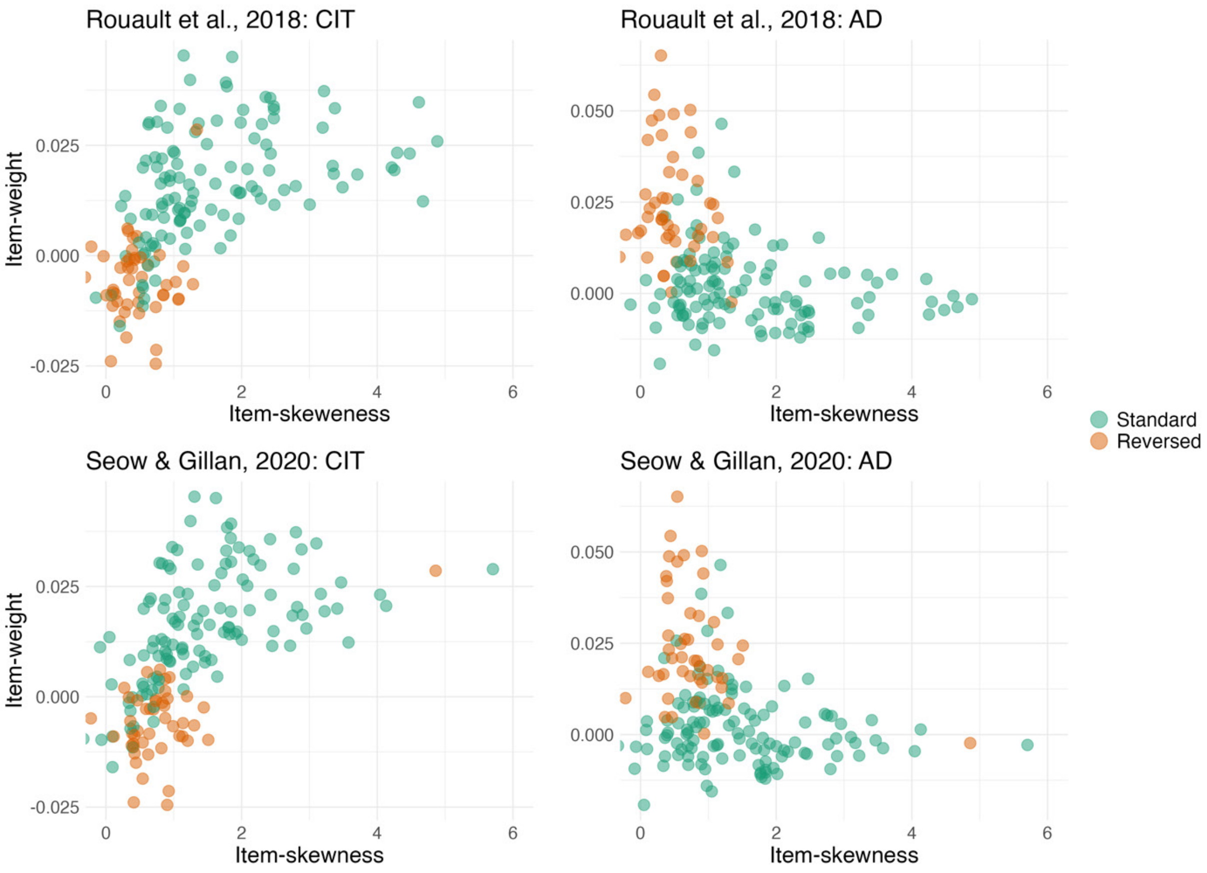

Figure 6 *The effect of reversed items and skewness on the CIT and AD dimensions.* Top row: Rouault et al. (2018) dataset; Bottom row: Seow and Gillan (2020) dataset. Columns: CIT - Compulsive behavior and intrusive thought; AD - Anxious-depression. Item weights were taken from Gillan et al. (2016). The higher the item's weight, the larger its contribution to the factor. Each point represents an item from one of the inventories shown in Figure 4. Standard and reversed items are color-coded. The X-axis was set to the range 0-6.

So far, we relied on post hoc measures—item skewness and coding direction (standard vs. reversed)— to assess the impact of inattentiveness and acquiescence on confidence ratings. In the next section, we measure inattentive responding and acquiescence directly, to detect these biases and empirically test their influence on confidence ratings.

Inattentiveness and acquiescence are associated with shifts in confidence ratings

We conducted an online experiment involving a perceptual decision task with self-report measures designed to detect inattentive responding and acquiescence. Our sample comprised 195 participants recruited from Prolific, of whom 50 were classified as inattentive, in line with our preregistered sample size (https://osf.io/jdquy). Using the same perceptual decision task as in Rouault et al., (2018), participants were asked to decide which of two squares contained more black dots and rate their confidence in this decision. Five participants were excluded for an average accuracy below 60%. Subsequently, participants completed questionnaires for OCD (OCI-R; Foa et al., 2002) and depression (SDS; Zung, 1965). Inattentive participants were detected using catch 'infrequency items' such as "I often rearrange the furniture in my home to prepare for the arrival of magical beans" (expected answer: 'not at all'), as suggested by Zorowitz and colleagues (2023).

In addition, we included an inventory of 14 "content-neutral" items that were curated by us to quantify participants' tendency to produce high or low ratings irrespective or content. We used a subset of items from the validated Extreme Response Style inventory (e.g. "I like to visit places that are totally different from my home", Greenleaf, 1992), and added items of our own (e.g., "I believe there are relatively few different breeds of cats"). Crucially, items were chosen such that psychopathology-relevant content should be balanced out at the inventory level. For example, the novelty-avoiding item "When I go shopping, I find myself spending very little time checking out new products and brands" was mirrored by the novelty-seeking item "I like to visit places that are totally different from my home." As a result, the mean rating to these items should primarily reflect participants tendency to agree with statements regardless of their content, that is, their acquiescence.

The experiment aimed to test two main hypotheses: that inattentive participants provide higher confidence ratings compared to attentive participants, and that acquiescence is positively correlated with confidence ratings. Consistent with our first hypothesis, inattentive participants gave significantly higher confidence ratings (M = 0.65, SD = 0.17) compared to attentive participants (M = 0.54, SD = 0.17; $\mathrm{t(73.51) = 3.56, p < .001}$), Figure 7, panel A), with a medium-to-large effect size (Cohen's d = 0.62, 95% CI [0.28, 0.95]). In addition, consistent with our second hypothesis, acquiescence was moderately correlated with mean confidence across the entire sample. As preregistered, a Spearman correlation showed a significant association ($r_S = .28, p < .001$). To maintain consistency with the other analyses, we also report Pearson's correlation, which yielded a similar effect ($r = .30$, 95% CI $[.17, .43]$, $t(188) = 4.38, p < .001$; Figure 7, panel D). The effects of inattentiveness and acquiescence remained significant when controlling for age and sex (see Appendix).

Next, we performed a series of exploratory analyses to measure the contribution of inattentiveness and acquiescence to the correlations between decision confidence and psychiatric questionnaire scores. First, we examined the association between obsessive-compulsive tendencies and mean confidence and found that the two were positively correlated (r = .28, 95% CI [.14, .41], t(188) = 3.98, p < .001; Figure 7, panel E). This finding aligns with previous reports of overconfidence in participants with high OCI-R scores (Hoven, Rouault, et al., 2023; Seow and Gillan, 2020) and those with high CIT factor scores (Rouault et al., 2018; Hoven, Luigjes, et al., 2023). As noted above, this positive correlation contrasts with the clinical presentation of doubt, indecisiveness and heightened uncertainty among individuals with OCD (Dar, 2004; Dar, Lazarov, & Liberman, 2021; Rasmussen & Eisen, 1989; Sarig, Dar, & Liberman, 2012) and with experimental findings of under-confidence in OCD from lab-based experiments (Cougle et al., 2007; Dar, 2004; Karadag et al., 2005; Marton et al., 2019; McNally & Kohlbeck, 1993; Zitterl et al., 2001; for review see Dar et al., 2022). We were therefore especially interested to see how this correlation would relate to the two surface-level properties of questionnaire-filling behaviour, namely acquiescence and inattentive responding.

Two aspects of the OCI-R questionnaire make it particularly vulnerable to inattentive and biased responding. First, several OCI-R items represent rare behaviours and cognitions (e.g., OCI-R item 10: “I feel I have to repeat certain numbers.”), with a mean skewness of 0.82 across all items. As a result, inattentive participants, who sample their responses semi-randomly, would appear highly symptomatic on this inventory. And second, the OCI-R contains no reversed items, which opens the door to acquiescence effects. Supporting this conjecture, inattentive participants, identified based on infrequency items, had much higher OCI-R scores (mean OCI-R= 30.23) than attentive participants (mean OCI-R= 17.59, t(66.15) = -5.04, p < .001), with a large effect size (Cohen’s d = 0.94, 95% CI [0.59, 1.28]). Furthermore, OCI-R was significantly correlated with acquiescence, measured as the mean rating over content-neutral items across participants (r = .27, 95% CI [.14, .40], t(188) = 3.91, p < .001). When we excluded inattentive participants, the correlation between OCI-R and confidence weakened ($r = .16$, 95% CI $[.00, .32]$, $t(142) = 1.97$, $p = .051$) and became non-significant when further controlling for acquiescence by regressing out the mean response to content-neutral items ($r = .09$, 95% CI $[-.08, .25]$, $t(142) = 1.03, p = .304$; Figure 7, panel H). In the appendix we provide additional analysis to show that this reduction in the correlation coefficient cannot be explained by the reduction in the sample size (excluding inattentive participants) nor by the correction procedure (regressing out the mean response to neutral items).

Turning to the SDS depression questionnaire, we observed the expected negative correlation between total scores and decision confidence (although this correlation was not statistically significant, r = -.11, CI [-.25, .03], t(188) = -1.58, p = .117). Low confidence among depressed individuals is in line with the clinical picture of low self-esteem and low self-efficacy that characterize depression (Fu et al., 2005; Hancock, 1996; Richards, 2011; Szu-Ting Fu et al., 2012). Responses to SDS items were descriptively less skewed than to OCI-R items (mean skewness across items, with reversed items reversed= 0.66), as most SDS items pertain to thought patterns that are more common in the general population (with the notable exception of two highly skewed items: “I have trouble with constipation,” skewness = 1.86, and “I feel that others would be better off if I were dead,” skewness = 2.28). Presumably for that reason, there was no difference in SDS scores between attentive (mean total SDS = 40.61) and inattentive responders (mean total SDS = 41.11; $t(96.57) = -0.32, p = .747$; Cohen’s d = 0.05, 95% CI [-0.28, 0.38]). Furthermore, with half of the items being reverse-coded, SDS is robust to effects of acquiescence, and indeed, SDS scores were uncorrelated with acquiescence (r = -.02, CI [-.16, .12], t(188) = -0.28, p = .779). This null effect was due to opposing effects of acquiescence on SDS standard items (r = .16, CI [.02, .30], t(188) = 2.28, p = .024) and reversed items (r = -.16, 95% CI [-.29, -.02], t(188) = -2.20, p = .029), which cancelled each other out. Consequently, unlike the OCI-R, the correlation between SDS and confidence was unaffected by the removal of inattentive responders (r = -.13, 95% CI [-.29, .04], t(142) = -1.55, p = .124) and when controlling for acquiescence (r = -.13, 95% CI [-.28, .04], t(142) = -1.52, p = .131).

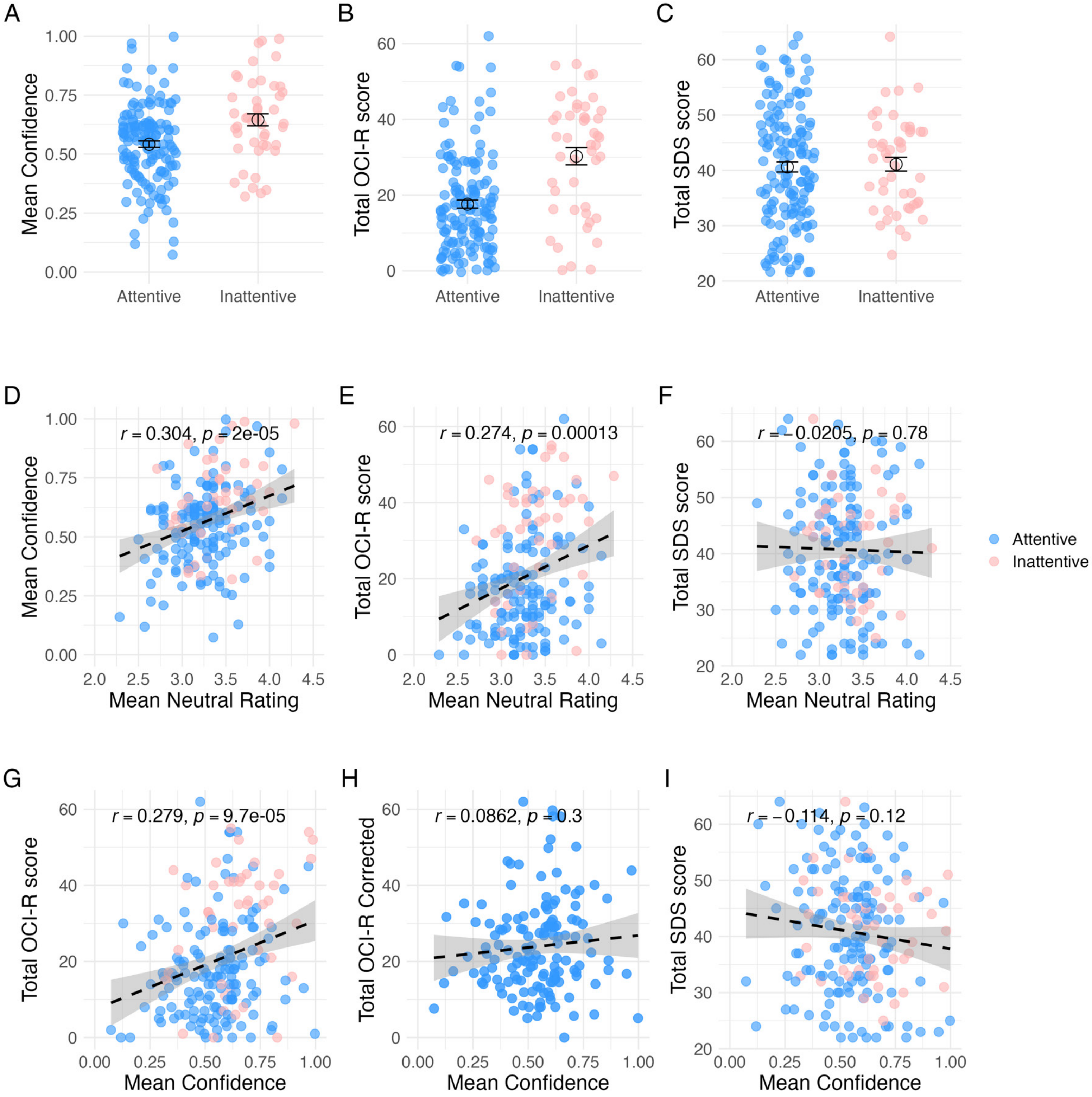
A
Mean Confidence
1.00
0.75
0.50
0.25
0.00
Attentive
Inattentive
B
Total OCI-R score
60
40
20
0
Attentive
Inattentive
C
Total SDS score
60
50
40
30
20
Attentive
Inattentive
D
$r = 0.304$, $p = 2e\text{-}05$
Mean Confidence
Mean Neutral Rating
2.0
2.5
3.0
3.5
4.0
4.5
E
$r = 0.274$, $p = 0.00013$
Total OCI-R score
Mean Neutral Rating
F
$r = -0.0205$, $p = 0.78$
Total SDS score
Mean Neutral Rating
Attentive
Inattentive
G
$r = 0.279$, $p = 9.7e\text{-}05$
Total OCI-R score
Mean Confidence
0.00
0.25
0.50
0.75
1.00
H
$r = 0.0862$, $p = 0.3$
Total OCI-R Corrected
Mean Confidence
I
$r = -0.114$, $p = 0.12$
Total SDS score
Mean Confidence

Figure 7 *Effects of inattentiveness and acquiescence on confidence ratings, OCI-R and SDS scores.* Panels A-C show comparisons between attentive and inattentive participants for: (A) mean confidence, (B) total OCI-R scores, and (C) total SDS scores. Each point represents scores for one participant. The hollow black circles represent the group mean and error bars show standard error of the mean (SEM). Inattentive participants are marked in pink. Panels D-F show the Pearson correlation between acquiescence and: (D) mean confidence, (E) total OCI-R scores, and (F) total SDS scores across participants. The dashed line indicates a linear fit, with the shaded area showing the 95% confidence interval. Panels G-I show the Pearson correlations between mean confidence and: (G) total OCI-R scores, (H) total OCI-R corrected for acquiescence and excluding inattentive participants and (I) total SDS scores.

Discussion

Decades of psychological research have identified limitations in the use of self-reports to measure psychological traits and mental health (Bagozzi & Yi, 1990; Baumgartner & Steenkamp, 2001; Campbell & Fiske, 1959.; Chandler et al., 2020; Curran, 2016; Greenleaf, 1992; Huang et al., 2012; Meade & Craig, 2012; Nichols, Greene, & Schmolck, 1989; Ophir et al., 2020; Podsakoff et al., 2003; Spector, 1987; Weijters et al., 2013), devising partial solutions and best-practice recommendations (see Huang, Bowling, Liu, & Li, 2015; Weijters et al., 2013). This accumulated wisdom has been largely left behind with the recent transition to massive-scale online testing and reliance on multiple questionnaires as a basis for the extraction of transdiagnostic psychiatric dispositions. Recently, concerns about the use of self-reports have resurfaced in the context of online testing (Chandler et al., 2020; Ophir et al., 2020), with evidence that inattentive responding leads to spurious negative correlations between the endorsement of rare items and task performance (Zorowitz et al., 2023). Our findings expand on and amplify these concerns. They show that biases common in responses to self-report inventories generalize to confidence ratings – a form of self-reports in themselves – thus giving rise to spurious correlations between psychiatric dimensions and metacognitive biases.

We believe that the finding of a positive correlation between compulsivity and decision confidence may be an important example of this effect. This finding, which has been obtained repeatedly in large-scale online studies, appears inconsistent with the clinical presentation of OCD as a "doubt disease" (Berrios, 1989; Janet & Raymond, 1903) and with the negative association between OCD symptoms and decision confidence observed in lab-based experiments. Importantly, this negative association was found both in individuals with high OCD tendencies (Hoven, Rouault, et al., 2023; Seow and Gillan, 2020, Lazarov, Dar, Liberman, & Oded, 2012; Zhang et al., 2017) and in clinical OCD samples (Cougle et al., 2007; Foa et al., 1997; Hermans et al., 2008; Karadag et al., 2005; McNally and Kohlbeck, 1993; Moritz et al., 2007; see Dar et al., 2022 and Hoven et al., 2019 for a review), even when controlling for anxiety and depression (Dar, 2004; Dar et al., 2000; Tolin et al., 2001). A direct comparison between participants with clinical OCD and compulsive individuals (measured using the OCI-R) within the same study found opposing trends: over-confidence among highly compulsive individuals and under-confidence in OCD participants, leaving this discrepancy unresolved (Hoven, Rouault, et al., 2023). Notwithstanding the differences between the CIT factor, obsessive-compulsive tendencies and clinical OCD, it is hard to think of a theoretical account that would explain the observed sign flip of the correlations of confidence ratings with the different measures.

Our findings suggest that this observed reversal is not due to substantive psychological differences between the three constructs but instead is accounted for by surface level questionnaire-filling behaviours. We show that a transdiagnostic dimension of compulsivity and intrusive thought partially captures surface-level behaviours, which may drive the positive correlation between this dimension and decision confidence. We suggest that previously reported positive correlations between decision confidence and compulsivity are likely to reflect the higher prevalence of inattentive responders (which are, as we show, also highly confident) in online samples.

By introducing a direct measure of inattentive responding to an online task we were able to show that inattentive participants are not only more likely to endorse rare symptoms but are also more confident in their decisions relative to attentive participants. This finding extends the results of Zorowitz and colleagues (2023) where inattentive responders were biased to give higher ratings in questionnaires regardless of their content. We suggest two factors that may contribute to this effect. First, roughly two thirds of inattentive responders in our study were self-declared males, compared to roughly half of all attentive responders ($\chi^2(1, n = 188) = 3.21$, $p = .073$ Figure A4). Given that male participants were, on average, more confident than female participants ($t(184.87) = 3.65, p < .001$; Cohen's d = 0.53, 95% CI [0.24, 0.82]; Figure A5), it is possible that part of the association between inattentiveness and high confidence is related to these gender differences. Second, we found that inattentive participants performed on average an easier task than attentive participants (Appendix, Figure A6, left panel). This effect was due to the staircase procedure, which is commonly employed in studies of population variability in metacognition, whereby poorer performance leads to incremental decreases in task difficulty (Hauser et al., 2017; Rouault et al., 2018; Wise et al., 2023). Task difficulty was in turn negatively associated with mean confidence, such that as the task became easier, mean confidence increased (Figure A6, right panel). As inattentive participants were on average facing an easier task than attentive participants, it is not surprising that they were more confident in their performance.

Looking forward, we would like to make several practical recommendations. First, researchers using self-report measures to probe psychiatric dimensions should adopt sensitive measures of inattentive and careless responding. Of note, comprehension checks have been included in both studies re-analyzed here (Seow & Gillan, 2020; Rouault et al., 2018), and more recent studies incorporated infrequency items as well (Fox et al., 2024). As the field moves forward, however, researchers should create novel infrequency items rather than relying on existing ones, as online participants often discuss unusual items in online forums, which undermines their efficacy (Zorowitz et al., 2023). When devising new infrequency items, it is advisable to use a similar language to the one used in other questionnaire items to avoid the item standing out, even to inattentive participants. Not only the content, but also the number of infrequency items can make a big difference. In our sample, 11.2% of all participants were identified as inattentive when using one infrequency item to identify careless responding, 17.5% when using two, 21.4% when using three and 24.2% when using four. A model that assumes that 28% of all participants are inattentive, and that the probability of an inattentive responder to fail an infrequency item is 39%, fitted our data well (see Appendix, Figure A9). This means that even with four infrequency items, 15.5% of all inattentive participants in our sample were not classified as such. In practice, then, it may be impossible to exclude all inattentive responders. Our recommendation is therefore to use infrequency items not only for participant exclusion, but also as a tool for researchers to quantify and report the potential effects of undetected inattentive responders on the observed patterns in the data.

Second, we recommend including a content-neutral self-report measure to assess participants' tendency for acquiescence. Such a measure should comprise items that have minimal association with psychiatric tendencies. This is especially important when testing correlations with questionnaires that do not include reversed items, such as the OCI-R.

Third, if using a staircasing procedure in studies of individual differences in metacognition, any effects of individual variability in task difficulty should be reported and discussed. As we show in the Appendix, staircasing renders performance similar across participants, but at the same time makes the task encountered by inattentive responders (or other groups that show poor performance) objectively easier. This can produce differences both in mean confidence and in more nuanced measures of metacognitive monitoring such as the difference in confidence between correct and incorrect decisions (e.g., metacognitive sensitivity; Maniscalco & Lau, 2012).

A more general recommendation is to broaden the scope of metacognition research beyond confidence ratings. Metacognitive knowledge and monitoring can be probed in ways that do not involve verbal or numerical self-reports, such as post-decisional wagering ("am I confident enough to bet on this decision?", Ben Shachar et al., 2013; Hembacher & Ghetti, 2017; Persaud et al., 2007) and information seeking ("do I require more evidence before committing a decision?", Siegel et al., 2021; Schulz, Fleming & Dayan, 2023; Selmeczy et al., 2013). Similarly, it has been suggested that decisions about absence, experimentally measured as decisions about missing targets and non-learned words, open a window into metacognitive knowledge about perception ("I would have seen the target if it was present," Mazor & Fleming, 2022; Mazor, Moran & Press, 2024; Sarna, Mazor & Dar, 2024) and memory ("I would have remembered this face if I had seen it before;" Ghetti, 2003; Mazor, 2021).

Finally, whenever theory-based predictions about interactions between metacognition and test conditions are possible, prioritizing such interactions over analyses of overall confidence levels is recommended. For example, theoretical accounts of metacognitive deficiencies in OCD make specific predictions about a metacognitive failure to separate thoughts from actions ('thought-action-fusion'; Rachman and Shafran, 1999), a difficulty to generate a feeling of knowing ('yedasentience'; Szechtman and Woody, 2004), or attenuated access to one's internal states, including memory (Liberman, Lazarov & Dar, 2023). Predictions from such theoretical models are often more specific than global effects on confidence ratings, making them more robust to pattern mimicry from surface-level questionnaire-filling behaviours. Causal interventions can provide an additional support for a true link between metacognition and mental health. For example, Fox et al., 2024 found that a decrease in AD scores following treatment was associated with a corresponding increase in mean confidence ratings. It should be noted, however, that treatment might affect surface-level questionnaire-filling behaviour such as acquiescence, which could simultaneously influence both reported confidence and dimensional mental health scores. To disentangle genuine treatment effects from changes in response styles, intervention studies should also implement sensitive measures of acquiescence and inattentive responding.

As a final note, we strongly believe that the marriage between computational modelling of behaviour and mental health research is a promising one. Given the centrality of metacognition to many psychiatric conditions, recent developments in our understanding of the computational underpinning of subjective confidence may have important implications for how we identify and treat mental health problems. Furthermore, the move away from theory-driven psychiatric classifications to a data-driven, dimensional approach, may open up fresh theoretical perspectives and avenues for personalized treatment. At the same time, conflicts between traditional, disorder-based research and more novel, dimension-based research are all but inevitable. Such conflicts should be welcome; by forcing the field to address them, they have great potential to advance our science. In particular, they are invaluable for promoting the integration of paradigmatic innovation with clinical theorizing and experience, which will be key to fostering research with clinical translational value.

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Methods

Analysis of existing datasets

Study selection rationale

We performed a literature review and found two published articles that include both raw scores of psychiatric inventories and data from a cognitive task with confidence rating. We report a reanalysis of data from two articles that publicly shared their raw data. Both make use of the original inventories from the factor analysis by Gillan et al. (2016) and importantly, both report an association between heightened mean confidence and CIT, and lowered mean confidence with AD.

1. Rouault et al. (2018), in which participants performed a perceptual discrimination task (decide which of two boxes has more dots in it) and rated their subjective confidence on a 6-point scale after each perceptual decision. We focused on experiment 2, which included the full pool of psychiatric questionnaires and shared the original analysis.
2. Seow and Gillan (2020), in which participants performed a predictive inference task (position a bucket to catch a flying particle) and rated their subjective confidence on a 100-point scale after making each prediction.

The following nine questionnaires were administered in both studies to assess various psychiatric symptoms: the Alcohol Use Disorder Identification Test (AUDIT) to measure alcohol addiction (Saunders, Aasland, Babor, De La Fuente, & Grant, 1993), the Apathy Evaluation Scale (AES) to assess apathy (Marin et al., 1991), the Self-Rating Depression Scale (SDS) to evaluate depression (Zung, 1965), the Eating Attitudes Test (EAT-26) for eating disorders (Garner, Olmsted, Bohr, & Garfinkel, 1982), the Barratt Impulsivity Scale (BIS-11) to measure impulsivity (Patton et al., 1995), the Obsessive-Compulsive Inventory – Revised (OCI-R) to assess obsessive-compulsive disorder (Foa et al., 2002), the State-Trait Anxiety Inventory (STAI) for trait anxiety (Spielberger, 1970), the Short Scales for Measuring Schizotypy (SSMS) to assess schizotypy (Mason, Linney, & Claridge, 2005), and the Liebowitz Social Anxiety Scale (LSAS) for social anxiety (Liebowitz, 1987).

In this analysis, we excluded the SSMS questionnaire because its binary scoring renders the measure of skewness irrelevant.

We used R (Version 4.3.2; R Core Team, 2023) and the R-packages *cowplot* (Version 1.1.3; Wilke, 2024), *ggpubr* (Version 0.6.0; Kassambara, 2023), *polycor* (Version 0.8.1; J. Fox, 2022), *psych* (Version 2.4.3; William Revelle, 2024), *gridExtra* (Version 2.3; Auguie, 2017), *groundhog* (Version 3.1.2; Simonsohn & Gruson, 2023), *lme4* (Version 1.1.35.3; Bates, Mächler, Bolker, & Walker, 2015), *moments* (Version 0.14.1; Komsta & Novomestky, 2022), *papaja* (Version 0.1.2; Aust & Barth, 2022), *patchwork* (Version 1.2.0; Pedersen, 2024), and *tidyverse* (Version 2.0.0; Wickham et al., 2019) for all our analyses.

Assessing acquiescence

1.1 Mean rating across items

To measure acquiescence, we calculated participants' mean responses to self-report items across all inventories. We used the mean response to self-report items as a *proxy* for acquiescence, as these studies did not include neutral items. This method allowed us to identify consistent agreement/disagreement patterns across diverse content. However, it represents a variation of Weijters et al. (2013) approach, as all items potentially share a core P factor related psychopathology (Caspi et al., 2014), dictated by the absence of neutral items in the original studies. The mean rating score was extracted after transforming all questionnaire response scales to the same range of 0-1. This was necessary because different questionnaires use different response scales, which can affect the mean rating (i.e., scales with higher values could become more influential). In this range, 0 represents the leftmost side of the scale and 1 represents the rightmost side of the scale. Additionally, we recoded reversed items to their original left-to-right position, as the acquiescence analysis focuses on preference for left-right position on the scale, regardless of the semantic meaning of each item. Item 25 from the EAT questionnaire ('I enjoy trying new reach food.') was excluded from this analysis because it could not be reversed to its original left-to-right rating. Its many-to-few coding (e.g., 1: 'Always,' 2: 'Usually,' and 3: 'Often' all coded as 0) made reversal impossible.

1.2 Reversed items inconsistency

Another marker of acquiescence is an inconsistency between responses to reversed and regular items (Weijters et al., 2013). For example, a participant who has a tendency to agree with self-report items independent of their content will show an inconsistency between reversed and standard items (agreement with an item and its opposite item, for instance both with 'I feel relaxed' and with 'I feel restless'). In our reanalysis section, we used the difference in item-confidence correlations between standard and reversed items inconsistency as a proxy for acquiescence.

Assessing Careless/Inattentive Responding Effects

There are various documented methods to detect inattentive responders in self-report inventories. Some methods rely on a priori inclusion of bogus or infrequency items (e.g., “I am paid biweekly by leprechauns”), while others rely on response patterns, such as identical consecutive responses, or inconsistency between responses to reversed and standard items (see Meade & Craig, 2012 for a review).

Here, we were particularly interested in a specific phenomenon discussed by (Chandler et al., 2020; King, Kim, & McCabe, 2018; Zorowitz et al., 2023) whereby inattentive participants appear symptomatic when symptoms frequencies are rare (see figure 1B ‘rare symptom’). Specifically, Zorowitz et al. (2023) found that when a self-report inventory probes for symptom with low base-rate frequency in the population (for example, an inventory asking about hypomanic behaviors), inattentive responders would appear more symptomatic than attentive ones. The reason is that attentive responders will mostly give zero ratings to a rare symptom, while inattentive responders will use the entire rating scale equally (for an illustration see Zorowitz et al., 2023, Figure 2). Statistically, the rarer the symptom, the more skewed its distribution; hence, as the distribution becomes more skewed, the effect of inattentive responding becomes more pronounced (a phenomenon that has been documented by King et al., 2018). We harnessed this phenomenon as a proxy for evaluating the effects of inattentive responding.

For every item in each questionnaire (148 items in total), we computed its skewness score and its Pearson correlation coefficient with the mean confidence ratings. We computed skewness using the “moments” package (Komsta & Novomestky, 2022).

Correlation tests

We report Pearson correlations when the population distribution is assumed to be normal. When normality is not assumed, we report Spearman correlations.

Experiment

After giving their informed consent, participants were instructed on the structure of the experiment, which included two parts: a perceptual task and a set of questions. They then received specific instructions regarding the perceptual decision task. In this task, participants viewed two black squares filled with black dots for 300 milliseconds and decided which square contained more black dots, the left or the right (with no time restriction). They were instructed to press ‘S’ for the left square and ‘F’ for the right. After making their perceptual decision, participants reported their confidence using a slider, ranging from ‘Guessing’ on the left to ‘Certainly correct’ on the right, with no numeric values displayed.

The number of dots in each square varied across trials, with the difference between the two squares adjusted by a staircase procedure (described below). One square contained a fixed number of 313 dots, while the other square had either more or fewer dots, depending on the trial’s difficulty level. With a difference of fewer dots creating a more difficult task. Each square contained 625 possible positions for black dots (arranged in a 25x25 grid). The specific positions of the dots within each square were randomly selected from these 625 possible locations on every trial, and the square with the greater number of dots (the target) was randomly assigned to appear on the left or right side of the screen on each trial.

Next, 25 practice trials were administered. In the first 6 trials, feedback on the perceptual decision was provided (after the confidence rating). The feedback stated either ‘Your box selection was correct’ or ‘Your box selection was incorrect.’ Feedback for incorrect decisions was shown for 3 seconds to emphasize the error, whereas feedback for correct selections was shown for 1.5 seconds. Participants then completed 19 additional trials without feedback. The purpose of these practice trials was to familiarize participants with the structure of the task. Upon completing the practice phase, participants received instructions for the main task, which included 300 trials divided into 4 blocks.

Staircase Procedure

A staircase procedure was used to adjust the task difficulty based on participants’ performance. The difference in the number of dots between the two squares (task difficulty) was initially set to 40 and then adjusted according to participants’ accuracy: following a 2-down 1-up procedure with a step size of 2 and a minimum difference of 0. At the limit, this procedure converges to a proportion of 72% correct responses.

The experiment was programmed in jsPsych and the experiment code and a demo of the task is available at github/noamsarna/BIRDAM. The order of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized to ensure registration time-locking (Mazor et al., 2019).

Comprehension questions.

Lastly, participants answered the following two comprehension questions:

3. “If you are certain you made the correct judgment, where on the scale would you place your confidence from 50% ‘Guessing’ to 100% ‘Certainly correct’?”
4. “If you are completely unsure whether you made a correct judgment, where on the scale would you place your confidence from 50% ‘Guessing’ to 100% ‘Certainly correct’?”

Psychiatric questionnaires

Upon completing the comprehension questions, participants were redirected to Qualtrics to complete the self-report section, which comprised the OCI-R (Foa et al., 2002) and the SDS (Zung, 1965).

Infrequency items

Each questionnaire included two “infrequency” items to assess inattentive responding. Specifically, we used the following four items, (the first written by us and the last two adapted from Zorowitz et al., 2023):

1. I was worried about the leprechauns who guard the hidden treasure (expected answer: ‘0- not at all’).
2. I often rearrange the furniture in my home to prepare for the arrival of magical beans (expected answer: ‘0- not at all’).
3. I find that relying on food and water is essential to my survival (expected answer: ‘4- Most of the time’; ‘3- Good part of the time’).
4. I am worried about the canine World Cup (expected answer: ‘1- A little of the time’).

Content-neutral items

Lastly, we used 14 neutral items (a mixture of items adapted from Greenleaf, 1992 and ones created by us) to assess participants' global tendencies in using the rating scale (ranging from: 1- "Strongly agree" to 5- "Strongly disagree"). These items were intended to be heterogeneous in content, therefore, not expect to share common content, and neutral in the sense that, as a group, they are not minimally related to psychopathology.

1. I think quantitative information is difficult to understand.
2. When I go shopping, I find myself spending very little time checking out new products and brands.
3. Everyone should use a mouthwash to help control bad breath.
4. A college education is very important for success in today's world.
5. I like to visit places that are totally different from my home.
6. I work very hard most of the time.
7. I will probably have more money to spend next year than I have now.
8. I think fashion is irrelevant.
9. I believe there are relatively few different breeds of cats.
10. I think the moon is very far from earth.
11. As I see it, Madrid is a small place.
12. Book covers are important in my opinion.
13. These days, matchboxes are no longer useful.
14. I find the taste of apples different to that of pears.

Data availability

Anonymized data from our online experiment is openly available on GitHub at https://github.com/noamsarna/BIRDAM

Code availability

Our analysis code for both the re-analyses of existing datasets and analysis for our experiment are available on GitHub at https://github.com/noamsarna/BIRDAM

References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology*, *75*(5), 547.

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., & Jagan, M. (2024). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Ben Shachar, A., Lazarov, A., Goldsmith, M., Moran, R., & Dar, R. (2013). Exploring metacognitive components of confidence and control in individuals with obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 255–261. https://doi.org/10.1016/j.jbtep.2012.11.007

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Benwell, C. S. Y., Mohr, G., Wallberg, J., Kouadio, A., & Ince, R. A. A. (2022). Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *Npj Mental Health Research*, *1*(1), 10. https://doi.org/10.1038/s44184-022-00009-4

Berrios, G. E. (1989). Obsessive-compulsive disorder: Its conceptual history in france during the 19th century. *Comprehensive Psychiatry*, *30*(4), 283295. Retrieved from https://www.sciencedirect.com/science/article/pii/0010440X89900527

Boldt, A., Fox, C. A., Gillan, C. M., & Gilbert, S. (2024). *Transdiagnostic compulsivity is associated with reduced reminder setting, only partially attributable to overconfidence*. https://doi.org/10.7554/eLife.98114.1

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56(2), 81.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., … Moffitt, T. E. (2014). The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49–55. https://doi.org/10.1037/abn0000479

Cougle, J. R., Salkovskis, P. M., & Wahl, K. (2007). Perception of memory ability and confidence in recollections in obsessive-compulsive checking. *Journal of Anxiety Disorders*, *21*(1), 118130.

Craddock, M. (2021). *metaSDT: Calculate type 1 and type 2 signal detection measures*. Retrieved from https://github.com/craddm/metaSDT

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Dar, R. (2004). Elucidating the mechanism of uncertainty and doubt in obsessive-compulsive checkers. *Journal of Behavior Therapy and Experimental Psychiatry*, *35*(2), 153–163. https://doi.org/10.1016/j.jbtep.2004.04.006

Dar, R., Lazarov, A., & Liberman, N. (2021). Seeking proxies for internal states (SPIS): Towards a novel model of obsessive-compulsive disorder. *Behaviour Research and Therapy*, *147*, 103987. https://doi.org/10.1016/j.brat.2021.103987

Dar, R., Rish, S., Hermesh, H., Taub, M., & Fux, M. (2000). Realism of confidence in obsessive-compulsive checkers. *Journal of Abnormal Psychology*, *109*(4), 673.

Dar, R., Sarna, N., Yardeni, G., & Lazarov, A. (2022). Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychological Medicine*, *52*(13), 2404–2412. https://doi.org/10.1017/S0033291722001908

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 112.

Foa, E. B., Amir, N., Gershuny, B., Molnar, C., & Kozak, M. J. (1997). Implicit and explicit memory in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *11*(2), 119129.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485.

Fox, Celine A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., … Gillan, C. M. (n.d.). *Metacognition in anxious-depression is state-dependent: an observational treatment study*. https://doi.org/10.7554/eLife.87193.2

Fox, Celine A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R. J., Hanlon, A. K., … Gillan, C. M. (2024). Reliable, rapid, and remote measurement of metacognitive bias. *Scientific Reports*, *14*(1), 14941. https://doi.org/10.1038/s41598-024-64900-0

Fu, T., Koutstaal, W., Fu, C. H. Y., Poon, L., & Cleare, A. J. (2005). Depression, Confidence, and Decision: Evidence Against Depressive Realism. *Journal of Psychopathology and Behavioral Assessment*, *27*(4), 243–252. https://doi.org/10.1007/s10862-005-2404-x

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: Psychometric features and clinical correlates. *Psychological Medicine*, *12*(4), 871878.

Ghetti, S. (2003). Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language*, *48*(4), 722739.

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42. https://doi.org/10.1016/j.cobeha.2017.07.003

Gillan, C. M. & Daw, N. D. (2016). Taking psychiatry research online. *Neuron*, *91*(1), 19–23. https://doi.org/10.1016/j.neuron.2016.06.002

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*, e11305. https://doi.org/10.7554/eLife.11305

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*(3), 328. https://doi.org/10.1086/269326

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25.

Hancock, J. A. (1996). “Depressive Realism” assessed via Confidence in Decision-making. *Cognitive Neuropsychiatry*, *1*(3), 213–220. https://doi.org/10.1080/135468096396514

Hauser, T. U., Allen, M., NSPN Consortium, Bullmore, E. T., Goodyer, I., Fonagy, P., … Dolan, R. J. (2017). Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Scientific Reports*, *7*(1), 6614. https://doi.org/10.1038/s41598-017-06116-z

Hembacher, E., & Ghetti, S. (2017). Subjective experience guides betting decisions beyond accuracy: evidence from a metamemory illusion. *Memory*, *25*(5), 575–585. https://doi.org/10.1080/09658211.2016.1197946

Hermans, D., Engelen, U., Grouwels, L., Joos, E., Lemmens, J., & Pieters, G. (2008). Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behaviour Research and Therapy*, *46*(1), 98–113. https://doi.org/10.1016/j.brat.2007.11.001

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & Van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational Psychiatry*, *9*(1), 268. https://doi.org/10.1038/s41398-019-0602-7

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & Holst, R. J. van. (2023). How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nature Mental Health*, 19. Retrieved from https://www.nature.com/articles/s44220-023-00062-8

Hoven, M., Rouault, M., Holst, R. van, & Luigjes, J. (2023). Differences in metacognitive functioning between obsessivecompulsive disorder patients and highly compulsive individuals from the general population. *Psychological Medicine*, *53*(16), 7933–7942. https://doi.org/10.1017/S003329172300209X

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., … Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Janet, P., & Raymond, F. (1903). *Les obsessions et la psychasthénie* (Vol. 2). F. Alcan.

Karadag, F., Oguzhanoglu, N., Ozdel, O., Atesci, F. C., & Amuk, T. (2005). Memory function in patients with obsessive compulsive disorder and the problem of confidence in their memories: A clinical study. *Anxiety*, *6*, 8.

Katyal, S., Huys, Q. J., Dolan, R. J., & Fleming, S. M. (2025). Distorted learning from local metacognition supports transdiagnostic underconfidence. Nature Communications, 16(1), 1854.

Kassambara, A. (2023). *Ggpubr: 'ggplot2' based publication ready plots*. Retrieved from https://CRAN.R-project.org/package=ggpubr

King, K. M., Kim, D. S., & McCabe, C. J. (2018). Random responses inflate statistical estimates in heavily skewed addictions data. *Drug and Alcohol Dependence*, *183*, 102–110. https://doi.org/10.1016/j.drugalcdep.2017.10.033

Komsta, L., & Novomestky, F. (2022a). *Moments: Moments, cumulants, skewness, kurtosis and related tests*. Retrieved from https://CRAN.R-project.org/package=moments

Komsta, L., & Novomestky, F. (2022b). *Moments: Moments, cumulants, skewness, kurtosis and related tests*. Retrieved from https://cran.r-project.org/web/packages/moments/index.html

Lazarov, A., Dar, R., Liberman, N., & Oded, Y. (2012). Obsessive compulsive tendencies may be associated with attenuated access to internal states: Evidence from a biofeedback-aided muscle tensing task. *Consciousness and Cognition*, *21*(3), 14011409.

Liberman, N., Lazarov, A., & Dar, R. (2023). Obsessive-Compulsive Disorder: The Underlying Role of Diminished Access to Internal States. *Current Directions in Psychological Science*, *32*(2), 118–124. https://doi.org/10.1177/09637214221128560

Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, *22*(141), e173. Retrieved from https://karger.com/book/chapter-pdf/2055981/000414022.pdf

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatry Research*, *38*(2), 143162.

Marton, T., Samuels, J., Nestadt, P., Krasnow, J., Wang, Y., Shuler, M., … Nestadt, G. (2019). Validating a dimension of doubt in decision-making: A proposed endophenotype for obsessive-compulsive disorder. *PLOS ONE*, *14*(6), e0218182. https://doi.org/10.1371/journal.pone.0218182

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*(2-3), 293296.

Mazor, M. (2021). Inference about absence as a window into the mental self-model. *PsyArXiv.(10.31234/Osf. Io/Zgf6s)*.

Mazor, M., & Fleming, S. M. (2022). Efficient search termination without task experience. *Journal of Experimental Psychology: General*.

Mazor, M., Mazor, N., & Mukamel, R. (2019). A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, *49*(9), 11491156.

Mazor, M., Moran, R., & Press, C. (2024). The role of counterfactual visibility in inference about absence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. https://doi.org/10.1037/a0019216

McNally, R. J., & Kohlbeck, P. A. (1993). Reality monitoring in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *31*(3), 249253.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Moritz, S., Wahl, K., Zurowski, B., Jelinek, L., Hand, I., & Fricke, S. (2007). Enhanced perceived responsibility decreases metamemory but not memory accuracy in obsessive compulsive disorder (OCD). *Behaviour Research and Therapy*, *45*(9), 20442052.

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239–250. https://doi.org/10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The Turker Blues: Hidden Factors Behind Increased Depression Rates Among Amazon's Mechanical Turkers. *Clinical Psychological Science*, *8*(1), 65–83. https://doi.org/10.1177/2167702619865973

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–261. https://doi.org/10.1038/nn1840

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rachman, S., & Shafran, R. (1999). Cognitive distortions: thought-action fusion. *Clinical Psychology & Psychotherapy*, *6*(2), 80–85. https://doi.org/10.1002/(SICI)1099-0879(199905)6:2<80::AID-CPP188>3.0.CO;2-C

Rasmussen, S. A., & Eisen, J. L. (1989). *Clinical features and phenomenology of obsessive compulsive disorder*.

Richards, D. (2011). Prevalence and clinical course of depression: A review. *Clinical Psychology Review*, *31*(7), 1117–1125. https://doi.org/10.1016/j.cpr.2011.07.004

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Sarig, S., Dar, R., & Liberman, N. (2012). Obsessive-compulsive tendencies are related to indecisiveness and reliance on feedback in a neutral color judgment task. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(1), 692697.

Sarna, N., Mazor, M., & Dar, R. (2024). Obsessive-Compulsive Visual Search: A Reexamination of Presence Absence Asymmetries. *Clinical Psychological Science*, 21677026241258380. https://doi.org/10.1177/21677026241258380

Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction*, *88*(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, *130*(3), 604.

Selmeczy, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 678.

Seow, T. X., Fleming, S. M., & Hauser, T. U. (2025). Metacognitive biases in anxiety-depression and compulsivity extend across perception and memory. PLOS Mental Health, 2(3), e0000259.

Seow, T. X., & Gillan, C. M. (2020). Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Scientific Reports*, *10*(1), 111.

Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological Psychiatry*.

Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B., & Schulz, L. E. (2021). Children's exploratory play tracks the discriminability of hypotheses. *Nature Communications*, *12*(1), 3598. https://doi.org/10.1038/s41467-021-23431-2

Simonsohn, U., & Gruson, H. (2023). *Groundhog: Version-control for CRAN, GitHub, and GitLab packages*. Retrieved from https://CRAN.R-project.org/package=groundhog

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, *72*(3), 438.

Spielberger, C. D. (1970). Manual for the state-trait anxiety inventory (self-evaluation questionnaire). *(No Title)*.

Szechtman, H., & Woody, E. (2004). Obsessive-Compulsive Disorder as a Disturbance of Security Motivation. *Psychological Review*, *111*(1), 111–127. https://doi.org/10.1037/0033-295X.111.1.111

Szu-Ting Fu, T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. Negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 699–704. https://doi.org/10.1016/j.jbtep.2011.09.014

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessivecompulsive disorder. *Behaviour Research and Therapy*, *39*(8), 913927.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Wilke, C. O. (2024). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, *11*(1), 4179. https://doi.org/10.1038/s41467-020-17977-w

Wise, T., Robinson, O. J., & Gillan, C. M. (2023). Identifying Transdiagnostic Mechanisms in Mental Health Using Computational Factor Modeling. *Biological Psychiatry*, *93*(8), 690–703. https://doi.org/10.1016/j.biopsych.2022.09.034

Zhang, Z., Wang, M., Miao, X., Li, Y., Hitchman, G., & Yuan, Z. (2017). Individuals with high obsessive-compulsive tendencies or undermined confidence rely more on external proxies to access their internal states. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 263269. Retrieved from https://www.sciencedirect.com/science/article/pii/S0005791616302294

Zitterl, W., Urban, C., Linzmayer, L., Aigner, M., Demal, U., Semler, B., & Zitterl-Eglseer, K. (2001). Memory deficits in patients with DSM-IV obsessive-compulsive disorder. *Psychopathology*, *34*(3), 113117.

Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, *7*(10), 1667–1681. https://doi.org/10.1038/s41562-023-01640-7

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry*, *12*(1), 6370.

Acknowledgements

We wholeheartedly wish to thank Tricia Seow, Marion Rouault, Steve Fleming, and Claire Gillan for their collaborative scientific spirit, critical and insightful comments, and constructive feedback. Without their generous and open sharing of data and code, this work would not have been possible. We also wish to thank Nitzan Shahar and Roni Maimon-Mor for their valuable feedback.

Appendix

Supplementary analysis

Mixed linear model analysis of item reversal effects on item-confidence correlation

In order to control for the possible effect of questionnaire content on the correlation between standard and reversed items with confidence, we performed a linear mixed model predicting item-correlation with confidence from item coding (standard/reversed) with random intercept for questionnaire (correlation ~ reversed + (1|questionnaire)). Results revealed that in both datasets reversed items showed lower correlations with confidence compared to standard items, even when accounting for questionnaire-level variation in the model: Seow and Gillan, 2020 ($\hat{\beta} = -0.17$, 95% CI $[-0.20, -0.15]$, $t(84.24) = -13.45$, $p < .001$), Rouault et al, 2018 ($\hat{\beta} = -0.11$, 95% CI $[-0.14, -0.09]$, $t(85.97) = -8.92$, $p < .001$). We also tested a model allowing the effect of reversal to vary across questionnaires (random slopes), but model diagnostics indicated this model was too complex for the available data.

Within-questionnaire analysis of skewness-confidence correlations

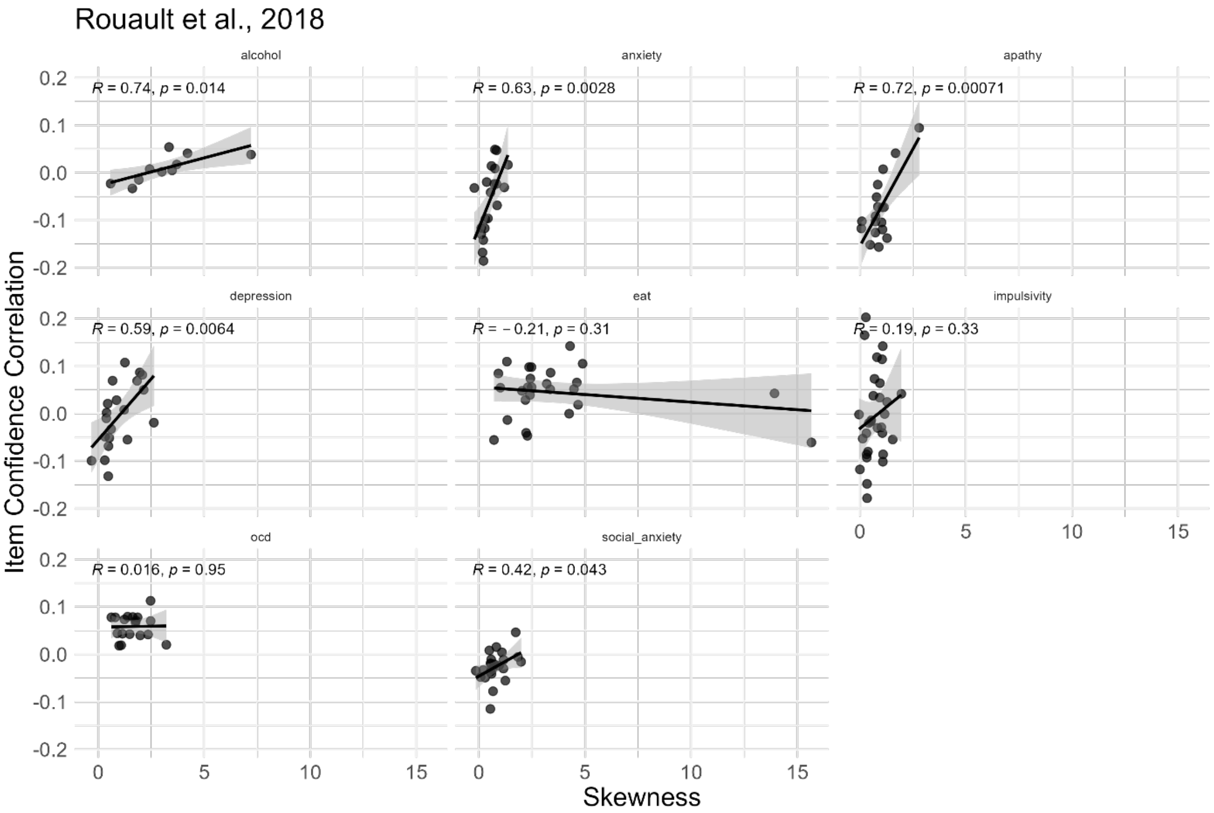

Figure A1 *Within-Questionnaire Analysis of Item Skewness and Item-Confidence Correlations in Rouault et al. (2018).* Each panel represents a different questionnaire, and each point represents an individual item. The x-axis shows the skewness of item responses. Y-axis shows the correlation between item ratings and confidence across participants. The black line represents the linear fit with 95% confidence intervals (gray shading). Pearson correlation coefficients (R) and corresponding p-values are shown for each questionnaire.

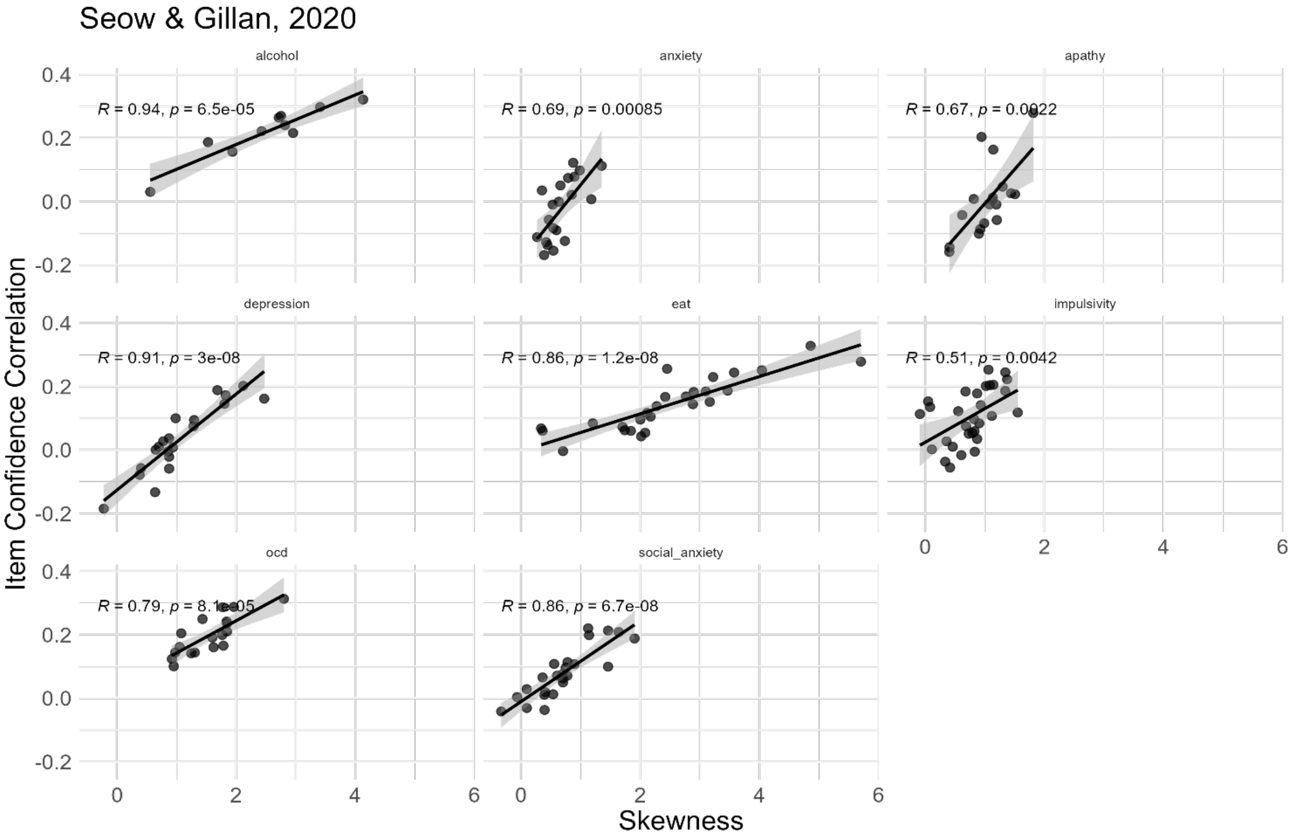

Figure A2 *Within-Questionnaire Analysis of Item Skewness and Item-Confidence Correlations in Seow and Gillan (2020). The same conventions as in Figure A1 are used.*

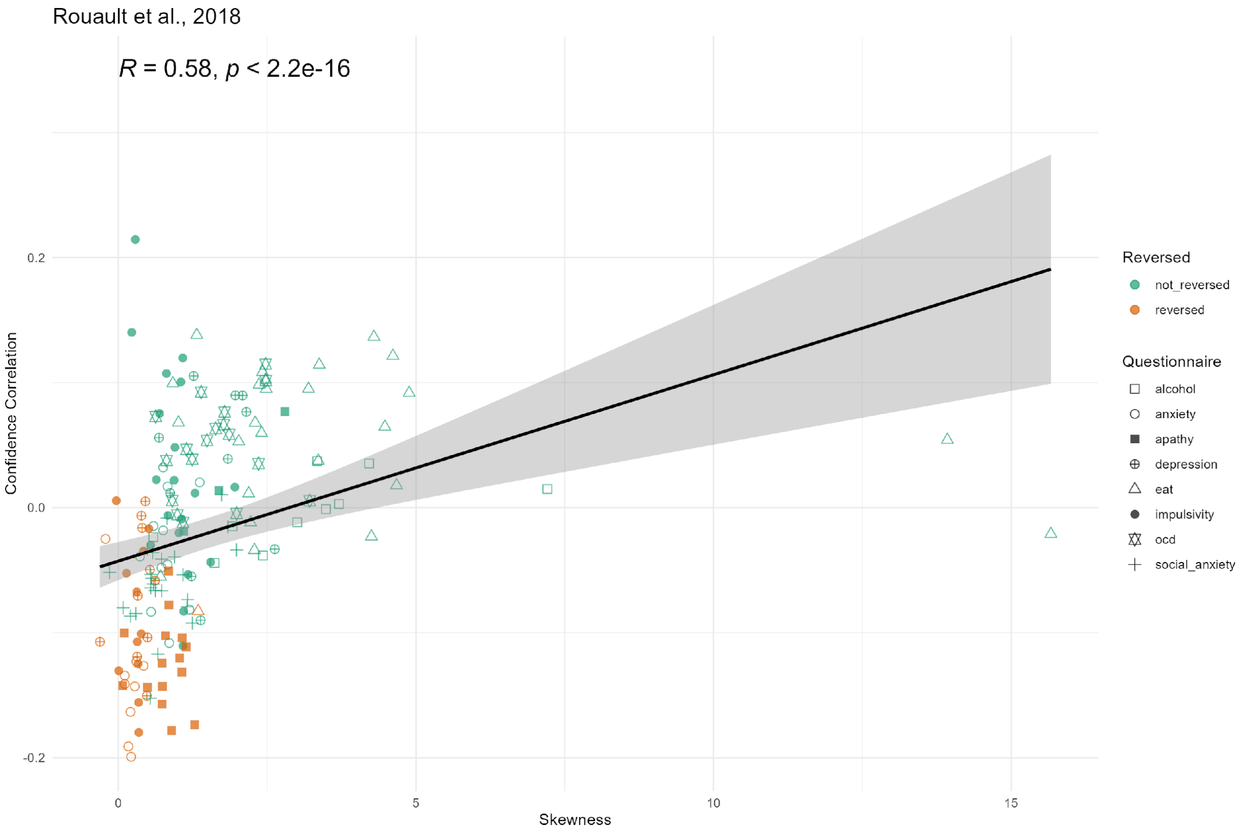

Figure A3 *Correlation between Item -Level Skewness and Confidence with Extreme Skewed Items Included.* Relationship between item-level skewness and item-confidence correlation in Rouault et al., (2018) including full skewness scale. Each point represents an item from the self-report questionnaires, with the shapes indicating different questionnaires and color indicating whether the item was reversed or not. The x-axis represents the skewness score of each item; the y-axis represents the Pearson correlation coefficient between the item's skewness and mean confidence ratings.

Effects of inattentiveness and acquiescence on confidence controlling for sex and age

We conducted multiple regression analyses to examine the effects of inattentiveness and acquiescence on mean confidence while controlling for sex and age. In the first model with inattentiveness, sex, and age as independent variables, we found that inattentiveness significantly predicted mean confidence even when controlling for age and sex (b = 0.09, 95% CI [0.03, 0.14], t(184) = 3.03, p = .003). In the second model with acquiescence, sex, and age as independent variables, we found that acquiescence also significantly predicted mean confidence when controlling for sex and age (b = 0.13, 95% CI [0.06, 0.20], t(184) = 3.75, p < .001).

Sex differences in inattentive responding

In our sample, the distribution of Sex (Male/Female self-reported on Prolific; two participants who did not report their sex were excluded from this analysis) differed between attentive and inattentive participants. Among attentive responders, there was no significant sex difference (Male: n = 69; Female: n = 73), whereas among inattentive participants, males were more frequent (Male: n = 30; Female: n = 16). This effect was only marginally significant in a chi-square test ($\chi^2(1, n = 188) = 3.21, p = .073$)

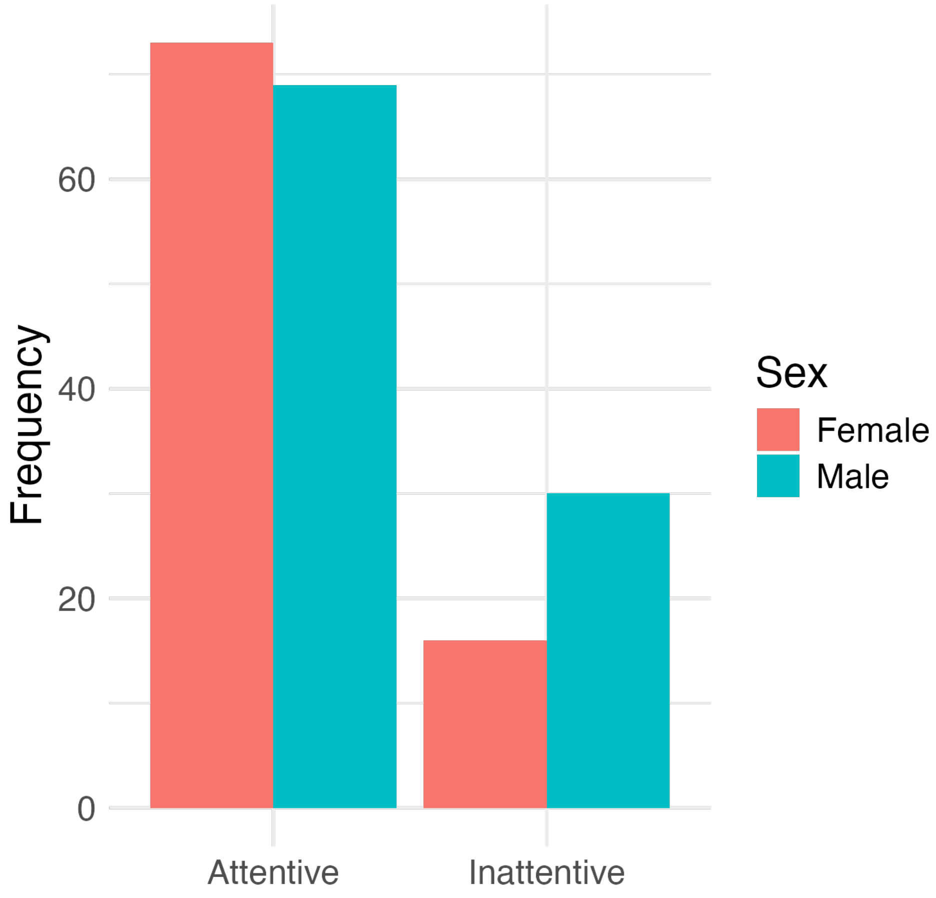

Figure A4 *Sex differences in inattentiveness prevalence.*

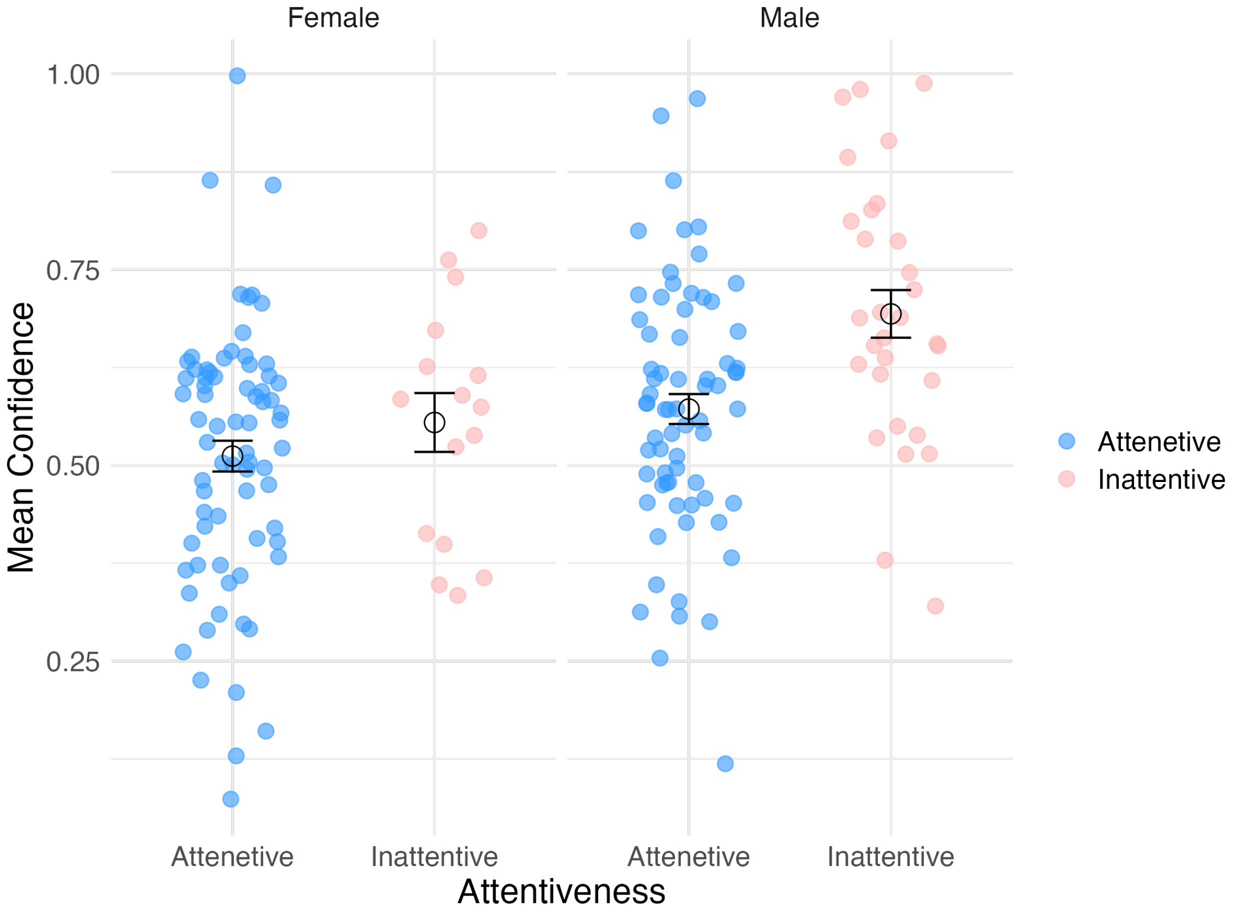

Figure A5 *Mean Confidence by Attentiveness and Sex.* Scatter plot showing mean confidence ratings across attentive and inattentive responders, separated by sex. Blue and pink points represent attentive and inattentive participants, respectively. Error bars indicate the standard error for each subgroup.

Inattentive responders encounter an easier task due to staircasing

As inattentive responders generally perform worse than attentive responders, they encounter an overall easier task difficulty when using a staircase procedure (see Methods for details on this procedure). This occurs because more mistakes lead to a reduction in task difficulty. In our experiment, task difficulty was defined as the difference in dots (increment) between two stimuli. The greater the difference, the easier the task, as the distinction between the two squares becomes more apparent. Therefore, we computed difficulty as the negative of the increment:

$$difficulty = -increment$$

Higher increments result in easier perceptual decisions. The mean difficulty was then calculated for each participant to assess overall task difficulty. We found that inattentive participants encountered significantly easier task on average compared to attentive participants, $t(188) = 5.20, p < .001$ (figure A6 left panel). We also found a negative correlation between task difficulty and mean confidence ratings, indicating that as the task became easier, mean confidence increased, $p = .004$ (figure A6, right panel).

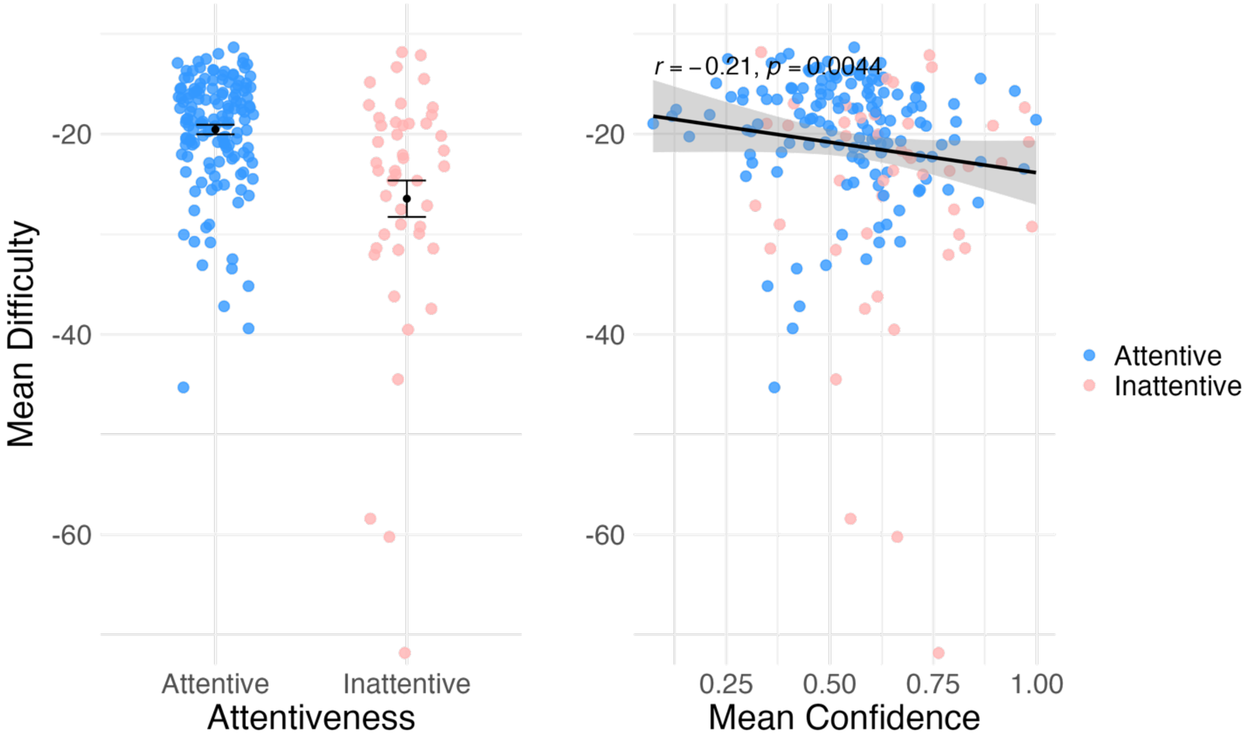

Figure A6 *Task difficulty and inattentiveness.* Left panel: mean difficulty ratings as a function of attentiveness, with data points for attentive and inattentive responses shown in red and blue, respectively. Black markers and error bars represent the mean and standard error, respectively. Right panel: mean confidence as a function of mean difficulty, with attentive and inattentive responses again differentiated by color. A negative correlation is observed, as shown by the black regression line (shaded area represents the confidence interval).

Analysis excluding participants that failed the confidence comprehension check

At the end of the perceptual task, participants' comprehension was assessed with two comprehension items (see Methods). Our pre-registered plan was not to exclude participants according to these items. However, as an exploratory analysis, we report our main analyses after excluding participants that failed the comprehension items. Initially, we only included participants who gave confidence ratings above 85 in response to the item: "If you are certain you made the correct judgment, where on the scale would you place your confidence from 50% ‘Guessing’ to 100% ‘Certainly correct’?” and below 65 in response to the item: “If you are completely unsure whether you made a correct judgment, where on the scale would you place your confidence from 50% ‘Guessing’ to 100% ‘Certainly correct’?”. Only 52 participant passed these stringent criteria. Instead, we adopted a more lenient criterion, and included all participants who gave higher rating to the first item than to the second item. This resulted in a total of 149 participants. We then tested our two main hypotheses which remained significant (hypothesis 1: $t(46.07) = 4.53, p < .001$, hypothesis 2: $r_s = .31, p < .001$).

A bootstrap permutation test for response style effects on OCI-R-confidence correlations

To validate that the reduction in correlation between total OCI scores and mean confidence ratings was stronger than expected by chance, we implemented a bootstrap permutation test. The procedure consisted of the following steps:

For each of 1,000 iterations, we:

1. Removed a random subset of 46 participants (equivalent to the number of inattentive participants in our original analysis) from the total sample.
2. Randomly assigned mean rating values (from content neutral items) to each participant, sampling without replacement from the original distribution of mean ratings.
3. For each OCI-R item we fitted a linear regression model predicting item response from the randomly assigned mean rating. Then, we extracted the residuals to use as an acquiescence-controlled item score.

4. Calculated new total OCI-R scores for each participant by summing the residualised item scores and scaling the corrected totals to match the original scale range (0-62). (Scaling was done for visualization purposes, and, being a linear transformation, does not affect the correlation with confidence).
5. Computed the Pearson correlation between corrected OCI-R total scores and mean confidence ratings. This procedure maintained the same sample size reduction and mathematical correction process as our main analysis but broke the relationship between participants and their mean ratings scores.

6. Computed the correlation drop between the original OCI-R total score confidence correlation coefficient and the correlation coefficient produced in each iteration, giving us an estimate of the expected drop in correlation.

The resulting distribution of correlation coefficients served as a null distribution, representing the expected drop in correlation coefficient if inattentive responding and acquiescence were not truly associated with OCI-R scores. We compared our observed drop in the correlation coefficient (r=0.19), calculated as the difference between the original correlation coefficient (r=0.28) and correlation coefficient observed after controlling for acquiescence and inattentiveness using actual mean ratings and removing truly inattentive participants (r=0.08), to this null distribution. We found that all 1,000 iterations produced smaller Pearson correlation coefficients drops than our observed value (p < .001), which suggests that the reduction in correlation in our main analysis represents a genuine effect of controlling for inattentiveness and acquiescence (Figure A7). We ran the same analysis for the SDS questionnaire. Out of 1,000 iterations in our null distribution, 626 produced correlation coefficients below our observed value, indicating that the observed reduction in correlation was not significantly different from what would be expected by chance (p = .374, Figure A8).

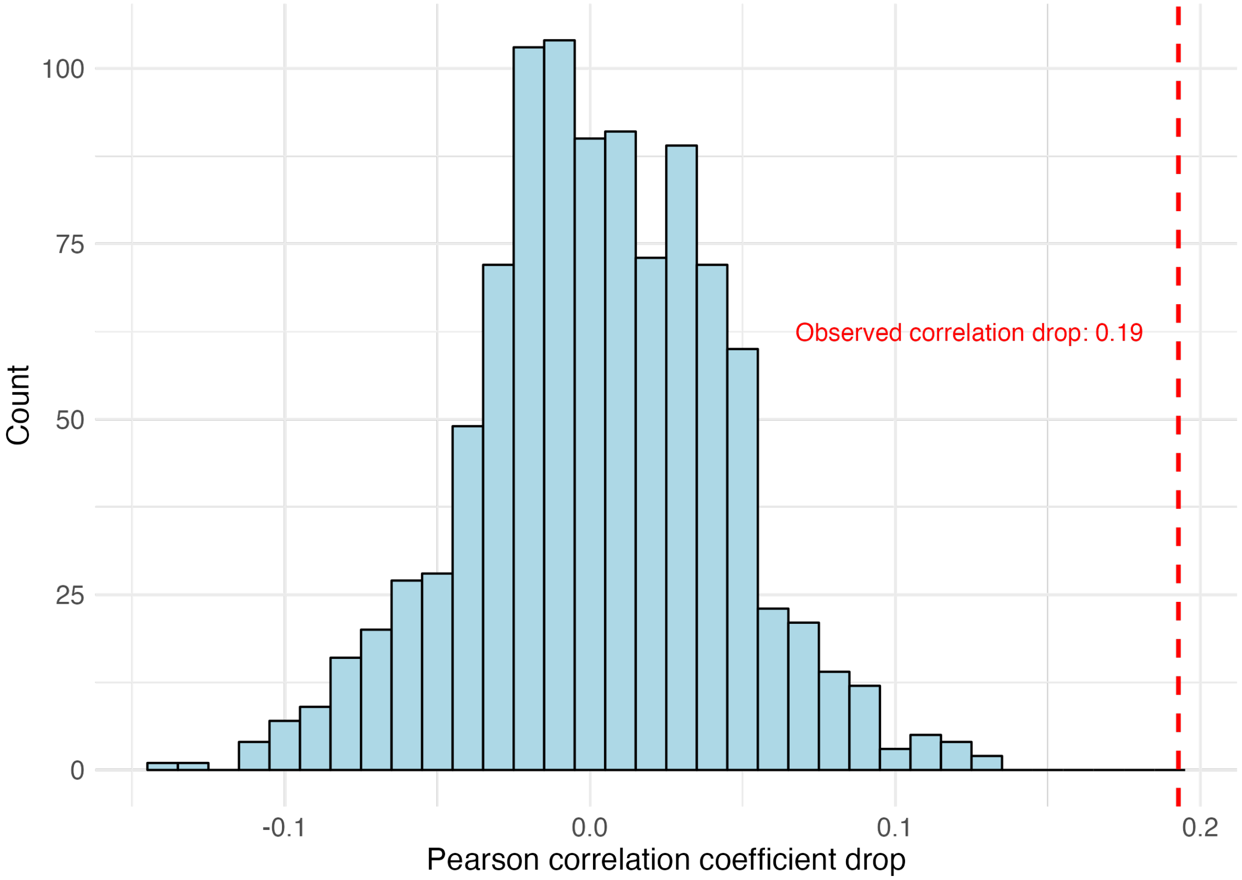

Figure A7 *Null distribution of OCI-R–confidence correlations drop from a permutation test.* Distribution of drop in Pearson correlation coefficients from 1,000 permutation iterations with randomised subset of participants (N=46) and mean rating values (blue). The red dashed line shows the observed drop in correlation after controlling for inattentive responding and acquiescence.

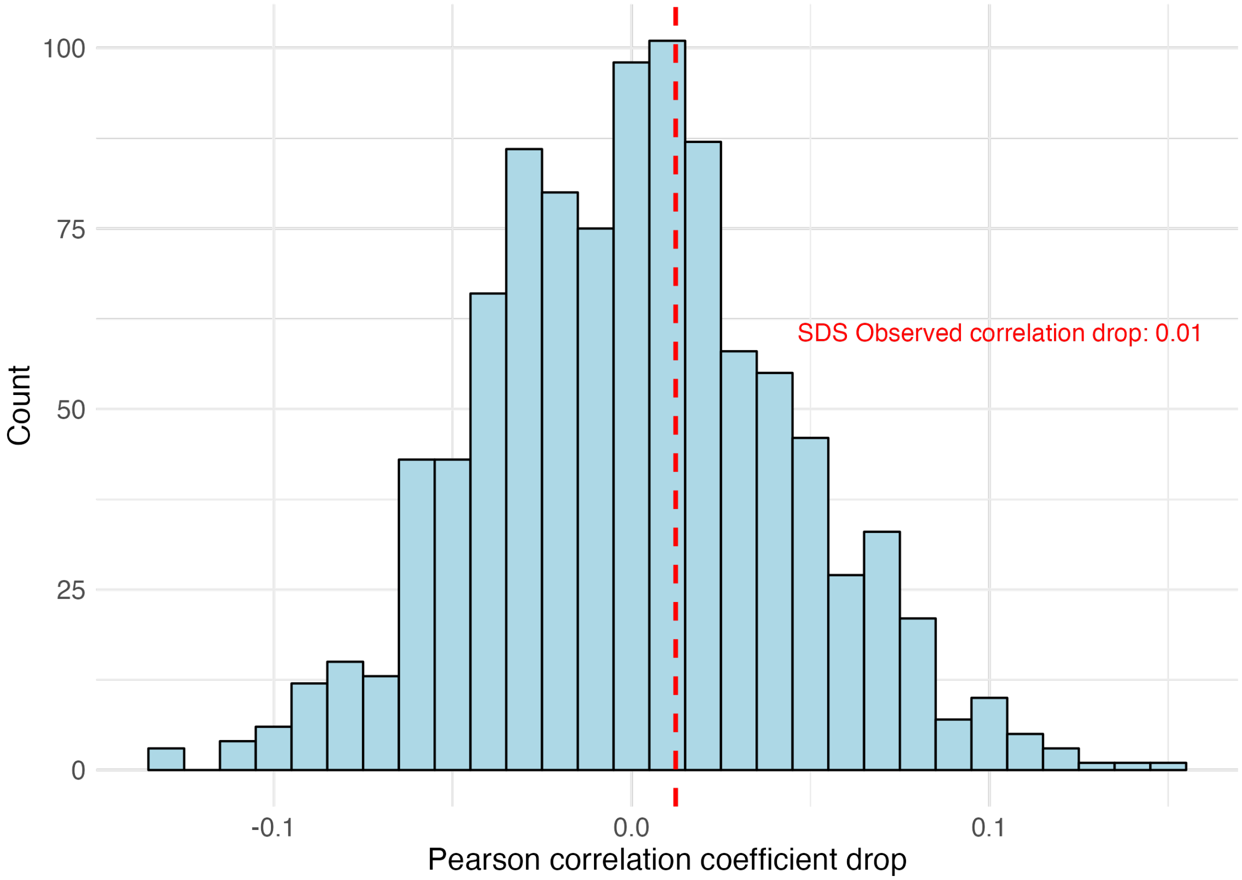

Figure A8 *Null distribution of SDS–confidence correlations drop from a permutation test.* The same conventions as in Figure A7 are used.

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Model estimation of inattentive responding

We estimated the prevalence of inattentive responding through a detection rate analysis and model fitting procedure. First, a detection rate analysis was conducted to examine the effectiveness of using multiple infrequency items. For this analysis, we calculated the percentage of participants identified as inattentive based on combinations of one to four infrequency items. For each possible combination of *k* items (where *k* = 1, 2, 3, or 4), we computed the percentage of participants who failed at least one item within that combination. We then averaged these percentages across all possible combinations of the same size to obtain the mean detection rate for each number of items. We found the following inattentive responders detection rates: 11.18% with one item, 17.54% with two items, 21.45% with three items, and 24.21% with four items.

To estimate the total proportion of inattentive respondents in our sample, including those not detected by our infrequency items, we fit a model assuming that: (1) a fixed proportion of participants (*x*) respond inattentively, and (2) each inattentive participant has a probability (*p*) of being captured by any single infrequency item. Under this model, the probability of detecting an inattentive participant with *n* items is given by:

$$\mathrm{Prob(detection)} = \mathrm{x} * (1 - (1 - \mathrm{p})\text{^}\mathrm{n})$$

Model parameters were estimated using a grid approximation approach. We created a 101 × 101 grid of possible parameter values, with both *x* (the proportion of inattentive participants) and *p* (the probability of an inattentive participants being captured by a single infrequency item) ranging from 0 to 1 in increments of 0.01. For each parameter combination, we calculated the negative log-likelihood (NLL) of observing our detection rates given those parameters:

$$\mathrm{NLL}(\mathrm{x},\mathrm{p}) \; = \; -\sum \log[\mathrm{Likelihood}(\mathrm{observed}_{\mathrm{n}}|\mathrm{N},\mathrm{x},\mathrm{p})]$$

where the likelihood was calculated using the binomial probability mass function, with *N* representing the total number of participants. The parameter combination that minimized the NLL was selected as the best-fitting model.

NLL reached a minimum for *x* = 0.28 and *p* = 0.39, suggesting that 28% of participants were responding inattentively, with each inattentive participant having a 39% probability of being caught by any single infrequency item. The model's predictions closely matched the observed detection rates (Figure A9).

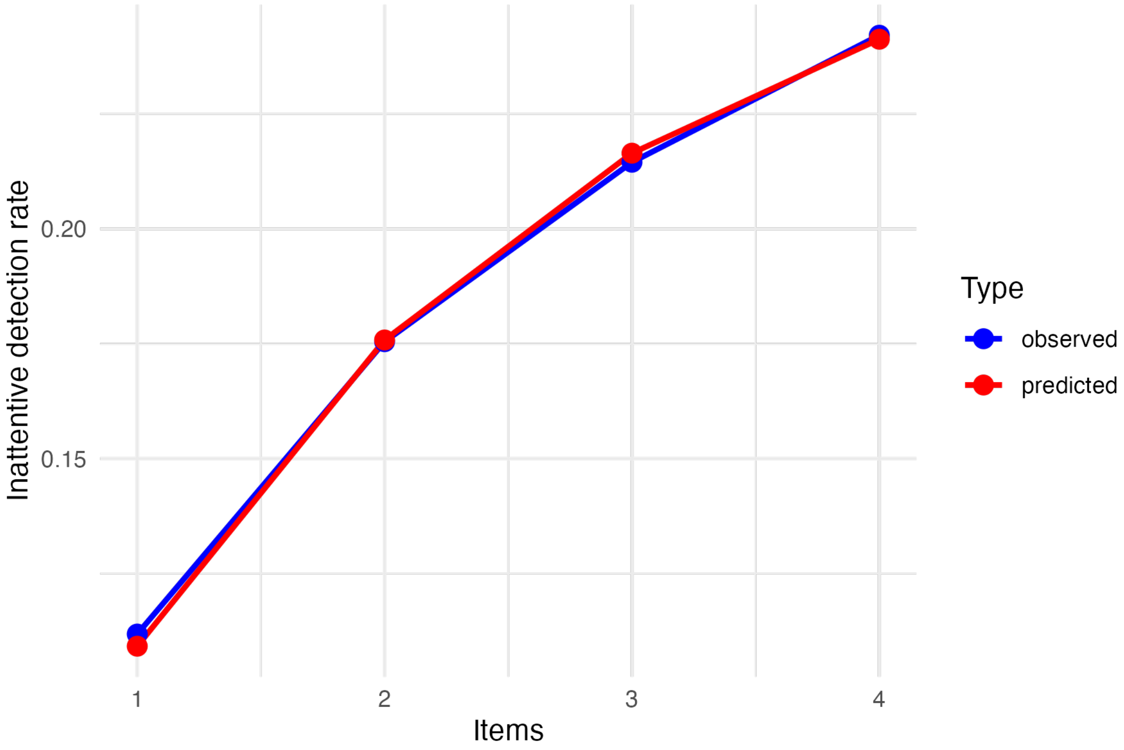

Figure A9 *Observed and predicted inattentive detection rates as a function of the number of infrequency items.* Observed detection rates (blue) represent the mean percentage of participants identified as inattentive across all possible combinations of k infrequency items. The red line represents the best-fitting model, which estimated that 28% of participants responded inattentively, with each inattentive participant having a 39% probability of being detected by a single infrequency item.

Biased and Inattentive Responding Drive Apparent Metacognitive Biases in Mental Health

Simulation of the relationship between item skewness and factor weights

To assess the relationship between item skewness and factor weights, we conducted a simulation using R. We simulated responses from 200 participants on a 20-item questionnaire measuring two distinct latent traits. The first 10 items measured the first latent trait, while the remaining 10 items measured the second latent trait. Data generation followed these steps: First, two underlying traits (a and b) were simulated for each participant from normal distributions (M = 0, SD = 0.5). Responses were then generated by adding random normal error (SD = 1) to the relevant trait value, using the mean of the latent variable. Then, the cumulative density function of the normal distribution *N(0,1)* was used to translate individual responses to quantiles, ranging from 0 to 1. Then, quantile scores were projected back to ratings, using distributions of different levels of skewness for different items. This was done by using the inverse cumulative density function of beta distributions with shape parameters *b=2* for all items, and *a* ranging from 2 to 11. Finally, scores were binned into seven bins of equal width.

Factor Analysis Procedure

For each simulated dataset, we followed the factors analysis procedure of Gillan et al. (2016). We calculated a heterogeneous correlation matrix using polychoric correlations (via polycor::hetcor() with Maximum likelihood estimation) to account for the ordinal nature of the data. We then conducted a factor analysis using the psych package (specifically psych::fa()) with maximum likelihood estimation, a fixed two-factor solution, and oblimin rotation ensuring the mean weight per factor is positive. Item skewness was measured using moment package, and

Spearman correlations between item skewness and the corresponding factor weights were calculated. We then took the averaged correlation across the factors.

We repeated this procedure 1000 times using different random seeds. On average, we observed a pattern of weak negative correlations between item skewness and item weight ($M = -0.04$, 95% CI $[-0.04, -0.03]$, $t(999) = -14.40$, $p < .001$), meaning that item weights were slightly closer to zero for more skewed items. This effect is much weaker, and in the opposite direction, to what we find for the CIT dimension in both datasets.