PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Biased and inattentive responding contribute to apparent metacognitive biases in mental health

Noam Sarna[a,1], Reuven Dar[a], and Matan Mazor[b,c]

Affiliations are included on p. 11.

Edited by Timothy D. Wilson, University of Virginia, Charlottesville, VA; received August 11, 2025; accepted March 18, 2026

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 may be largely due to surface-level dimensions of questionnaire-filling behavior: systematic rating biases and inattentive responding. Specifically, a bias to report positive or negative values in self-report scales may generalize to confidence ratings, producing spurious correlations between the two. Additionally, systematic overconfidence 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 behavior. In a preregistered 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 substantial influence of inattentive and biased responding over both self-report psychiatric measures and confidence ratings. When not accounted for, these factors can produce a mirage of apparent metacognitive alterations in mental health. We discuss concrete precautionary measures that are needed to control for these biases.

computational psychiatry | self-report | metacognition | rating scales | mental health

Significance

Understanding the cognitive mechanisms underlying mental health is crucial for developing effective treatments that target causes rather than symptoms. To identify these mechanisms, researchers test large samples of online participants and seek associations between their self-reported mental health tendencies (e.g., compulsivity) and their task behavior (e.g., confidence ratings in perceptual decisions). Here we show that such associations can be driven by surface-level questionnaire-filling behaviors. In five datasets, we show that biased and inattentive responding are liable to systematically distort our understanding of the relationship between mental health and metacognition, potentially leading to false conclusions about the underlying mechanisms. We submit that a better understanding of these effects is essential for translating insights from the lab to the clinic.

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 (1–10). 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 (11). 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 (4).

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 et al. (12). In their analysis, three factors emerged from a pool of nine 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 (12).

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 underconfident in one's decisions) (6–8, 13–15). 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 (6). This finding has since been replicated in other, independent samples (13–17) and extended to a variety of cognitive tasks [e.g., a predictive inference task (8), a gamified version of the perceptual-decision-making task (3), an external reminder-usage task (2)].

Author contributions: N.S. and M.M. designed research; N.S. performed research; N.S. and M.M. analyzed data; R.D. and M.M. secured funding; and N.S., R.D., and M.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: noamsarna@mail.tau.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2520832123/-/DCSupplemental.

Published May 4, 2026.

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 (7, 9; for review and discussion see ref. 18). Here, we suggest that the well-documented associations between metacognition and transdiagnostic dimensions can arise from the interaction of psychometric properties of self-report questionnaires with response biases, to the extent that these confounding biases pose a rival explanation for any true association with mental health. 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. Therefore, true relationships between psychiatric dimensions and confidence become indistinguishable from confounds driven by response biases.

We consider two properties of questionnaire-filling behavior 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. 1 *A*). In this model, a questionnaire item (I) 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 their content (19) (Fig. 1 *A*). In this paper, we use acquiescence to refer to 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 refs. 20 and 21). Critically, acquiescence is likely to affect both questionnaire responses and subjective confidence ratings, potentially producing an appearance of a link between decision confidence and symptom severity (Fig. 1 *B*, *Left*).

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 (22). Inattentive respondents are thought to sample their responses semirandomly from a nearly uniform distribution (23, 24) (Fig. 1 *A*, *Rightmost*). 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 (Fig. 1 *B*, *Middle*). 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 symptoms and decision confidence (Fig. 1 *B*, *Middle*).

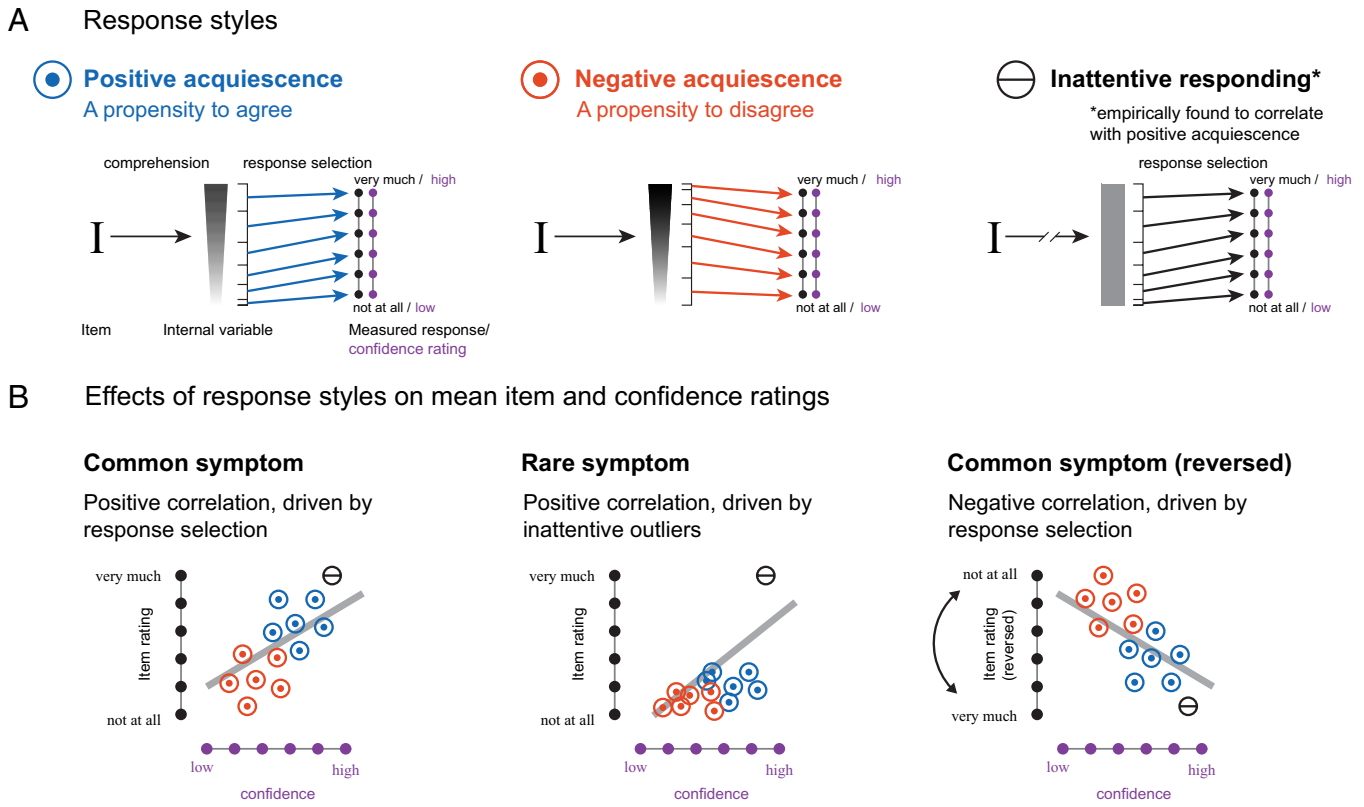

Fig. 1. A schematic illustration of the effects of acquiescence and inattentive responding on item-confidence correlations. (*A*) We describe the production of a self-report as a two-step process. First, a questionnaire item is read, generating an internal variable that represents a subjective level 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 a general tendency to agree or disagree with an item regardless of its 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 symptom (reversed)' refers to symptoms with high prevalence in the population which are articulated in a reversed tense (e.g., "I am happy"; STAI item 10).

We elaborate on these two effects in the *Methods* section and demonstrate their respective contributions to the reported associations between mental health and metacognition in three large datasets: Hoven, Luigjes et al. (17), Rouault et al. (6), and Seow and Gillan (7) in the Results section. We then show that transdiagnostic factors of mental health are robustly aligned with surface-level questionnaire-filling behavior. Finally, analysis of a new dataset with direct measures of inattentive responding and acquiescence reveals that these surface-level properties of questionnaire-filling behavior are reliably correlated with psychiatric questionnaire scores and with confidence ratings in a perceptual task, consequently contribute to the positive correlation between the two.

Results

We start by reporting the reanalysis of three 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 (25)], apathy [Apathy Evaluation Scale, AES (26)], depression [Self-Rating Depression Scale, SDS (27)], eating attitudes [Eating Attitudes Test, EAT-26 (28)], impulsivity [Barratt Impulsivity Scale, BIS-11 (29)], obsessive-compulsive tendencies [Obsessive-Compulsive Inventory – Revised, OCI-R (30)], schizotypy [Short scales for measuring schizotypy (31)], social anxiety [Liebowitz Social Anxiety Scale, LSAS (32)], and anxiety. For anxiety, Rouault et al. (6) and Seow and Gillan (8) used the State-Trait Anxiety Inventory [STAI (33)], whereas Hoven et al. (17) used the Generalized Anxiety Disorder scale [GAD-7 (34)]. The same participants also rated their confidence in perceptual decisions. In Rouault et al. (6) (Exp. 2) and Hoven et al. (17) (*SI Appendix*, Fig. SA1, *Left*), participants decided which of two briefly presented boxes had more dots in it and rated their subjective confidence on a 6-point or 50-point scale, respectively. In Seow and Gillan (8) (*SI Appendix*, Fig. SA1, *Right*), participants positioned a bucket to catch a flying particle and rated their subjective confidence on a 100-point scale. For more details about the datasets reanalyzed here, see *SI Appendix*, *Supplementary Methods*.

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 (Fig. 1 *B*, *Leftmost*). 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 an acquiescence proxy based on the average rating across all questionnaire items. As some of the self-report items are reversed-phrased, i.e., phrased such that agreement indicates lower endorsement of the construct, they must be reverse-scored to align with the scale's direction (we elaborate on this in the next section). To account for this, we calculated this proxy as the rating delta between standard and reversed items across all psychiatric inventories (for more details, see *SI Appendix*, *Supplementary Methods*).

In Seow and Gillan (8), there was a positive correlation between mean item rating and mean confidence rating ($r = 0.31$, 95% CI $[0.23, 0.40]$, $t(435) = 6.92$, $P < 0.001$; Fig. 2 *A*, *Middle*), such that higher mean ratings across items were associated with higher mean confidence ratings. A positive correlation was also found in Rouault et al. (6) ($r = 0.23$, 95% CI [0.14, 0.31], $t(495) = 5.15$, $P < 0.001$) and in Hoven et al. (17) ($r = 0.19$, 95% CI [0.10, 0.27], $t(485) = 4.20$, $P < 0.001$). 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.

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 [STAI (33)], impulsivity [BIS-11 (29)], depression [SDS (27)], and apathy [AES (26)] 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 STAI 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 (21).

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 (8) 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 $[0.14, 0.20]$, $t(85.12) = 11.28$, $P < 0.001$; Fig. 2 *B*, *Middle*), with a large effect size [Cohen's d = 2.41, 95% CI (1.85, 2.96)]. A similar pattern was also observed in the Rouault et al. (6) 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 $[0.09, 0.14]$, $t(78.29) = 8.98$, $P < 0.001$; Fig. 2 *B*, *Left*), with a large effect size (Cohen's d = 1.93, 95% CI [1.41, 2.43]). Finally, in the

Fig. 2. Associations between acquiescence-related measures and confidence across three datasets. *Left* column: Rouault et al. (6). *Middle* column: Seow and Gillan (8). *Right* column: Hoven et al. (17). (*A*) Correlation between rating delta and mean confidence rating. Each point represents a single participant's rating delta (mean rating for standard items minus mean rating for reversed items, after reversal) across all inventories and mean confidence rating across all trials. The red line represents a linear regression fit, and the shaded gray area represents the SE of the fit. (*B*) Relationship between reversed-coded items and item confidence correlation. Each point represents a questionnaire item with the shapes indicating different questionnaires. Standard and reversed items are color-coded. Black markers denote the mean item–confidence correlation for standard items (*Left*) and reversed items (*Right*), and the vertical bars represent the SEM. The reference line at y = 0 indicates zero item confidence correlation. Black dashed horizontal lines denote the correlation significance thresholds at alpha = 0.05, two-sided: For positive correlations, values above the line are significant, and for negative correlations, values below the line are significant. STAI items do not appear in the Hoven et al. (17) panel because that study used a different anxiety inventory that did not include reversed items. Asterisks denote significance ($P < 0.001$) of a *t* test for independent samples.

Hoven et al. (17) dataset, the same pattern was also found: standard items (mean *r* across the 32 items = 0.02) were on average more positively correlated with confidence than reversed items (mean *r* across the 36 items = −0.08, $\Delta M = 0.10$, 95% CI [0.07, 0.13], $t(57.95) = 7.23$, $P < 0.001$; Fig. 2 *B*, *Right*), with a large effect size (Cohen's $d = 1.78$, 95% CI [1.21, 2.34]). This effect remained significant in all three datasets when accounting for the main (intercept) effect of questionnaire in a mixed-effect model [Seow and Gillan (8): $t(84.24) = -13.45$, $P < 0.001$; Rouault et al. (6): $t(85.97) = -8.92$, $P < 0.001$; Hoven et al. (17): $t(40.79) = -5.82$, $P < 0.001$; see *SI Appendix*].

Analysis 2: Testing the Effect of Inattentive Responding with Item-Level Skewness. The semirandom 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 (23). 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 nonzero 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 nonzero ratings will be more likely to be inattentive responders than participants who provide zero ratings (Fig. 1 *B*, *Middle*).

Given that inattentive participants were previously found to be biased toward using the positive ("agree") half of a survey rating scale (24), we reasoned that inattentive responders may rate their confidence as higher on average, which can produce spurious positive correlation between the endorsement of rare (right skewed) psychiatric symptoms and decision confidence. Supporting this, we found a positive correlation between item skewness and its correlation with confidence in all three datasets [Seow and Gillan (8): $r_s = 0.71$, $P < 0.001$; Rouault et al. (6): $r_s = 0.55$, $P < 0.001$; Hoven et al. (17): $r_s = 0.52$, $P < 0.001$; Fig. 3*A*], 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 found in all three datasets using models that allowed the strength of the relationship to vary across questionnaires (random slope): Seow and Gillan (8) ($\hat{\beta} = 0.12$, 95% CI [0.08, 0.17], $t(6.02) = 6.00$, $P < 0.001$), Rouault et al. (6) ($\hat{\beta} = 0.03$, 95% CI [0.01, 0.05], $t(5.66) = 2.65$, $P = 0.040$), and Hoven et al. (17) ($\hat{\beta} = 0.03$, 95% CI [0.01, 0.05], $t(4.87) = 3.47$, $P = 0.019$), indicating that the association is not attributable solely to between-questionnaire differences and is, on average, observed within questionnaires (*SI Appendix*). This finding also elucidates a curious pattern which can be observed in Fig. 2*A*: in all three datasets, the positive correlation between mean item rating and mean confidence was mostly driven by a subset of participants with high mean item ratings, who also gave high confidence ratings. As a result, the correlation was much stronger in the dataset that featured more such subjects [Seow and Gillan (8)]. We suggest that this qualitatively distinct group of subjects may be inattentive responders and that they may contribute to the observed effect of acquiescence on item-confidence correlations.

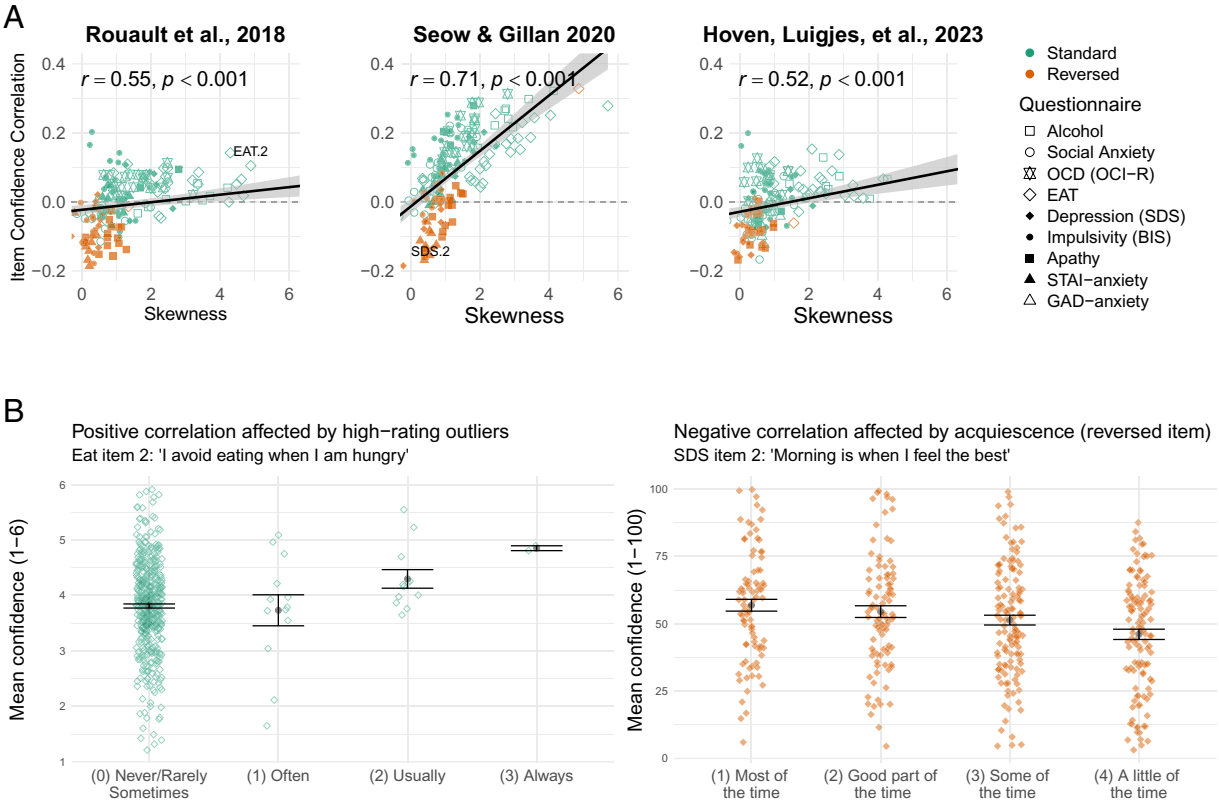

Fig. 3. Correlation between item-level skewness and item-confidence correlation. (*A*): Each point represents an item from the self-report questionnaires, with the shapes indicating different questionnaires. Standard and reversed items are color-coded. The x-axis represents the skewness score of each item. 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 SE of the fit. The dashed horizontal line at y = 0 marks zero correlation. (*B*): mean confidence ratings by questionnaire item responses. *Left*: EAT item 2, 'I avoid eating when I am hungry,' from Rouault et al. (6), showing a positive correlation affected by high-rating outliers in the "Usually" and 'Always' response category. *Right*: SDS item 2, 'Morning is when I feel the best," from Seow and Gillan (8), demonstrating a negative correlation influenced by acquiescence to reversed items. Error bars represent SE of the mean. For visualization purposes only, the x-axis of panel *A* was set to the range 0 to 6, thereby not showing two extreme outliers. See *SI Appendix* for a figure including these two items.

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;" Fig. 3, *B Left*). Endorsement of this item in Rouault et al. (6) is significantly correlated with mean confidence ($t(495) = 2.35,\ P = 0.019$), 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.

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 higher mean confidence, whereas the AD (anxious-depression) dimension has been linked to lower mean confidence (6, 8, 17). One possibility is that these associations reflect true relationships between metacognition and psychopathology. 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 possibility, we tested the contribution of acquiescence and inattentive responding to the transdiagnostic factor structure itself, irrespective of confidence ratings. To that end, we obtained the factor weights of individual items as originally reported by Gillan et al. (12), as both Rouault et al. (6) and Seow and Gillan (8) relied on these weights in their analysis, and since these factors were computed based on a large sample (N = 1,413).* In Fig. 4*A*, we plotted 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 (Fig. 4 *A*, *Top* row) there is a positive correlation between item weight and skewness, such that more skewed items contribute more to the CIT factor. This association was large and significant in both datasets: in Seow and Gillan (8) (),$r_s = 0.68,\ P < 0.001$ and in Rouault et al. (6) ($r_s = 0.67,\ P < 0.001$). This finding is consistent with the conjecture that the positive correlation between the CIT dimension and confidence is 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 attentive, symptomatic participants (Fig. 1*B*, Rare Symptom; see also ref. 23). In the *SI Appendix*, we show that these results hold when controlling for the random effect of the questionnaire. We also present the results of a simulation which demonstrates that a positive correlation between item skewness and factor weights is unexpected under reasonable assumptions about the link between skewness and diagnostic content.

A second clear trend in Fig. 4 is that in the CIT factor, standard items (green markers) and reversed items (orange markers) have

*As Hoven et al. (17) used a questionnaire set that did not match the original set of Gillan et al. (12) (included the GAD-7 instead of the STAI to measure anxiety) we report its analysis separately in *SI Appendix*.

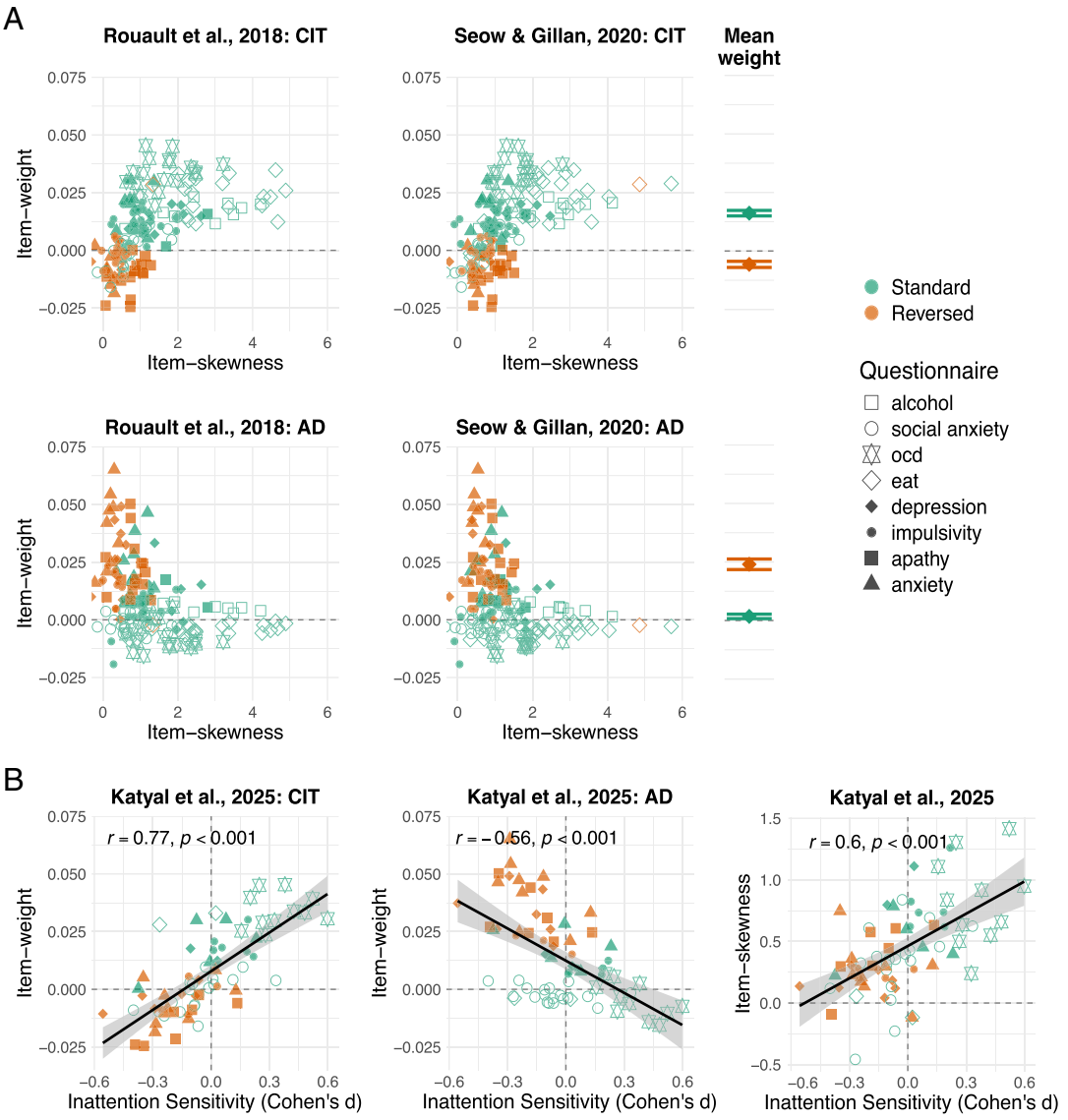

Fig. 4. The effect of reversed items and skewness on the CIT and AD dimensions. Each point represents a questionnaire item, with point shapes indicating the questionnaire. Standard and reversed items are color-coded. The dashed horizontal line at y = 0 serves as a reference for zero item weight. Item weights were taken from Gillan et al. (12). The higher the item's weight, the larger its contribution to the factor. Note that item weights are assigned to reversed items after reversal. (*A*) *Top* row: CIT—compulsive behavior and intrusive thought. *Bottom* row: AD—anxious-depression. *Left* column: Rouault et al. (6). *Right* column: Seow and Gillan (8). Diamond symbols in the rightmost column display the mean ± SEM for item weights, with separate values shown for standard (green) and reversed (orange) items. For visualization purposes only, the x-axis was set to the range 0 to 6, thereby not showing two extreme outliers. See *SI Appendix* for a figure including these two items. (*B*) Item sensitivity to inattention is correlated with both CIT/AD item weights and item-level skewness in Katyal et al. (14). In all three panels, the x-axis shows each item's sensitivity to inattentive responding (Cohen's d, comparing inattentive vs. attentive participants). *Left* panel: the y-axis shows the CIT item-weight. *Middle* panel: the y-axis shows the AD item-weight. *Right* panel: the y-axis shows item-level skewness. The solid black line represents the linear regression fit, and the shaded region reflects the SE of the fit. The reported correlation is the Spearman correlation coefficient.

average weights with opposite signs, whereas in the AD factor, standard item weights are near zero on average, while the average weight for reversed items is positive. When calculating total scores in psychiatric questionnaires, reversed items (such as "I feel that I am useful and needed" in a depression scale) are coded such that low endorsement results in a higher symptom score. Crucially, the CIT and AD dimensions effectively *undo this semantic reversal*. As a result, their factor scores likely reflect acquiescence bias: a general tendency to agree or disagree. Specifically, on average, the CIT dimension is correlated with a tendency to agree, and the AD dimension with a tendency to disagree. This mirrored pattern is not surprising given the high negative correlation between the weights of these two factors across items ($r = -0.58$, 95% CI $[-0.67, -0.47]$, $t(164) = -9.09$, $P < 0.001$).

Analysis 4: Examining the Relation between Skewness and Inattention Sensitivity. In Analyses 2 and 3, we used item-level skewness to estimate the potential impact of inattentive responding. This choice was grounded in the empirical observation that highly skewed items are more affected by inattention (23, 24). However, skewness also reflects the underlying prevalence of symptoms, such that the distribution of rarer symptoms (e.g., compulsivity) tends to be more skewed than the distribution of common symptoms (e.g., anxiety and depression). It is therefore important to determine the extent to which skewness tracks inattention rather than symptom prevalence. To address this, we analyzed a fourth dataset, Katyal et al. (14), which includes a reliable inattention index derived from infrequency items [a gold-standard approach for detecting inattentive responding (35)]. This enabled us to compute an

inattention-sensitivity index for each psychiatric questionnaire item by comparing endorsement rates between inattentive and attentive respondents (expressed as Cohen's d). If skewness is a sensitive index of inattention, items with higher skewness should show larger inattentive–attentive differences. This is exactly what we found: inattention sensitivity was strongly associated with item-level skewness $r_s = 0.60$, $P < 0.001$ (Fig. 4 *B*, *Right*). We also tested the correlation between CIT and AD factor weights and inattention sensitivity and found the same patterns obtained for skewness: inattention sensitivity positively correlated with CIT weights ($r_s = 0.77$, $P < 0.001$; Fig. 4 *B*, *Left*), and a negatively correlated with AD weights ($r_s = -0.56$, $P < 0.001$; Fig. 4 *B*, *Middle*). In other words, items that received higher ratings from inattentive participants loaded high on CIT and low on AD, and the opposite was true for items that received lower ratings from inattentive participants. This validates our use of skewness as a sensitive proxy for inattention, and reaffirms that the factor structure strongly aligns with inattentive responding.

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. (6) and Hoven et al. (17), 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 (30)] and depression [SDS (27)]. 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 et al. (24).

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 of content (acquiescence). We used a subset of items from the validated Extreme Response Style measure (e.g., "I like to visit places that are totally different from my home") (36), 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."

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 ($M = 0.54$, $SD = 0.17$;;$t(73.51) = 3.56$, $P < 0.001$, Fig. 5, 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 ($r_s = 0.28$, $P < 0.001$). To maintain consistency with the other analyses, we also report Pearson's correlation, which yielded a similar effect ($r = 0.30$, 95% CI $[0.17, 0.43]$, $t(188) = 4.38$, $P < 0.001$; Fig. 5, panel *D*). The effects of inattentiveness and acquiescence remained significant when controlling for age and sex (*SI Appendix*).

Although our experiment was not designed to explain how inattention inflates confidence ratings, we examined one factor that may account for higher confidence among inattentive participants: variation in task difficulty between groups. We found that inattentive participants performed on average an easier task than attentive participants, with a large effect size ($t(188) = 5.20$, $P < 0.001$, Cohen's d = 0.88; *SI Appendix*, Fig. SA14, *Left*). 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 (6, 9, 37). Task difficulty was in turn negatively associated with mean confidence, such that as the task became easier, mean confidence increased ($r_s = -0.21$, $P = 0.004$; *SI Appendix*, Fig. SA14, *Right*). 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. While outside the focus of the present paper, this effect of staircasing on overall confidence has important implications for the study of individual differences in metacognition in general, not only in the context of mental health.

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 = 0.28, 95% CI [0.14, 0.41], t(188) = 3.98, $P < 0.001$; Fig. 5, panel *G*). This finding aligns with previous reports of overconfidence in participants with high OCI-R scores (7) and those with high CIT factor scores (6, 15–17, 38). Yet, this positive correlation contrasts with the clinical presentation of doubt, indecisiveness, and heightened uncertainty among individuals with OCD (39–41) and with experimental findings of underconfidence in OCD from lab-based experiments [(39, 42–46), for review see ref. 47]. We were therefore especially interested to see how this correlation would relate to the two surface-level properties of questionnaire-filling behavior, 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 behaviors 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 uncontrolled acquiescence effects.

Indeed, 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, $P < 0.001$; Fig. 5, panel *B*), 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 ($r = 0.27$, 95% CI [0.14, 0.40], t(188) = 3.91, $P < 0.001$; Fig. 5, panel *E*). When we excluded inattentive participants, the correlation

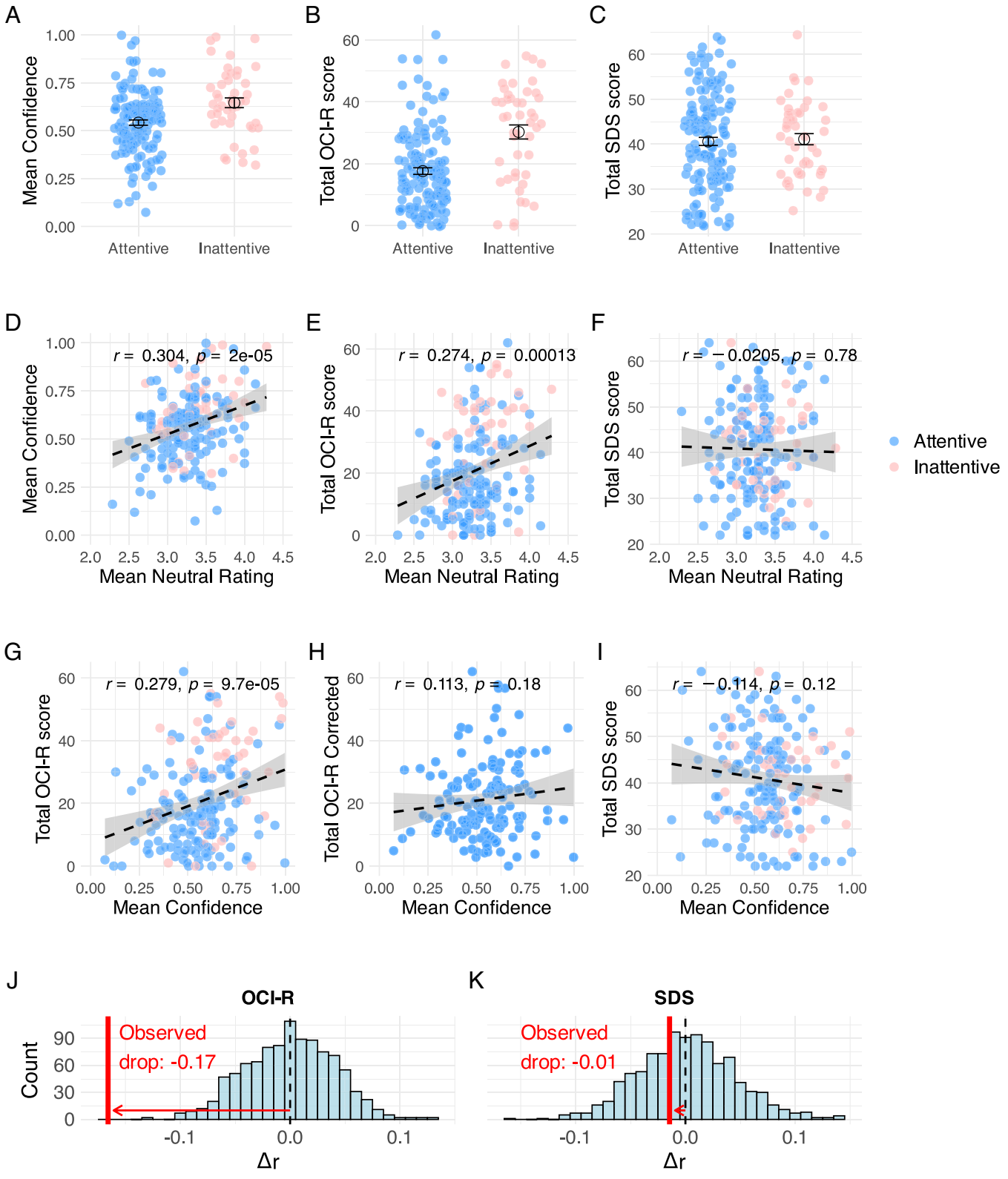

Fig. 5. Effects of inattentiveness and acquiescence on confidence ratings, OCI-R and SDS scores. Panels (*A*–*C*) show comparisons between attentive and inattentive participants for: (*A*) mean confidence, (*B*) total OCI-R scores, and (*C*) total SDS scores. Each point represents scores for one participant. The hollow black circles represent the group mean and error bars show 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% CI. 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. (*J*) the relative drop in correlation strength between OCI-R and confidence when controlling for acquiescence and removing inattentive participants ($\Delta r = r_{corrected} - r_{original}$). More negative values indicate a larger drop after applying the correction. The red line shows the observed drop in correlation after controlling for inattentive responding and acquiescence, and the blue histogram is a nonparametric null distribution of values (see *SI Appendix* for details). (*K*) similar to *J*, for SDS.

between OCI-R and confidence weakened ($r = 0.16$, 95% CI [0.00, 0.32], $t(142) = 1.97$, $P = 0.051$) and then weakened further when further controlling for acquiescence by regressing out the mean response to content-neutral items ($r = 0.11$, 95% CI [−0.05, 0.27], $t(142) = 1.36$, $P = 0.176$ Fig. 5, panel *H*). Readers should not overinterpret the nonsignificance of the final correlation, but instead focus on the magnitude of the drop in the correlation coefficient (from 0.27 to 0.11; Fig. 5, panel *J*), which was highly significant ($P < 0.001$), and 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; for full details see *SI Appendix*, *Supplementary Analysis*). While it is tempting to assert that the controlled OCI-R confidence correlation (r = 0.11, $P = 0.176$; Fig. 5, panel *H*) might become significant with a larger sample, it is also crucial to bear in mind that in cases of strong confounding variables with imperfect control, the significance of the effects of interest becomes difficult to interpret; in such cases, as sample size increases, the error rate also increases, raising the likelihood of a confusion between true effects and confounds (48) (see also *SI Appendix*, Reply to Gillan et al., 2025, for more details).

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, $r = -0.11$, CI [−0.25, 0.03], $t(188) = -1.58$, $P = 0.117$; Fig. 5, panel *I*). Low confidence among depressed individuals is in line with the clinical picture of negative appraisal, low self-esteem, and low self-efficacy that characterize depression (49–52). 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; $t(96.57) = -0.32$, $P = 0.747$; Cohen's d = 0.05, 95% CI [−0.28, 0.38]; Fig. 5, panel *C*). Furthermore, with half of the items being reverse-coded, the SDS is robust to effects of acquiescence, and indeed, SDS scores were uncorrelated with acquiescence (r = −0.02, CI [−0.16, 0.12], t(188) = −0.28, $P = 0.779$; Fig. 5, panel *F*). This null effect was due to opposing effects of acquiescence on SDS standard items (r = 0.16, CI [0.02, 0.30], t(188) = 2.28, $P = 0.024$) and reversed items (r = −0.16, 95% CI [−0.29, −0.02], t(188) = −2.20, $P = 0.029$), which canceled each other out. Consequently, unlike the OCI-R, the correlation between SDS and confidence was unaffected both by the removal of inattentive responders (r = −0.13, 95% CI [−0.29, 0.04], t(142) = −1.55, $P = 0.124$) and by controlling for acquiescence ($r = -0.13$, 95% CI [−0.29, 0.04], $t(142) = -1.55$, $P = 0.124$), also reflected in the nonsignificant correlation drop after applying these corrections (Fig. 5, panel *K*).

Discussion

Decades of psychological research have identified limitations in the use of self-reports to measure psychological traits and mental health (19, 21–23, 36, 53–60), devising partial solutions and best-practice recommendations (21). 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 (23, 59), with evidence that inattentive responding leads to spurious negative correlations between the endorsement of rare items and task performance (24). Our findings expand on and amplify these concerns. Specifically, 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. We further show that, even among attentive participants, biases common in responses to self-report inventories may generalize to confidence ratings—a form of self-reports in themselves—thus contributing to spurious correlations between psychiatric dimensions and metacognitive biases that are indistinguishable from the ones driven by true metacognitive effects.

We would like to emphasize that while our findings reveal how population variability in questionnaire-filling behavior can yield spurious correlations between confidence and mental health, it is *not* our claim that true, meaningful correlations between the two do not exist. Specifically, the negative correlation between the AD (anxiety depression) factor and confidence aligns with the clinical presentation of anxiety and depression, and most likely reflects a true negative correlation over and beyond any contribution from acquiescence. With respect to the positive correlation of confidence with the CIT (compulsivity and intrusive thought) factor, more work is needed to determine the extent to which it reflects a true association, over and above the joint effect of acquiescence and inattentive responding. While CIT is repeatedly found to positively correlate with confidence in online samples (2, 3, 6, 8, 13–17), confidence is reduced in both clinical OCD samples [(42, 43, 45, 61–63), see refs. 18 and 47 for a review], including studies that control for anxiety [(64), repeated recall paradigm: (65)] and depression (see analysis in ref. 47) and in samples of individuals with high OCD tendencies (66, 67). This tension between the diverging metacognitive patterns of clinical OCD patients and individuals with high compulsivity was discussed in Hoven et al. (68), who cautioned against generalizing metacognitive findings from the general population to clinical samples. We acknowledge that this discrepancy between patient and online studies is not fully resolved here, as the correlation between OCI-R scores and confidence did not become negative when controlling for acquiescence and inattentive responding. More work is needed to fully understand these diverging patterns.

Importantly, this discrepancy may be explained in part by the CIT factor being more than a pure index of obsessive-compulsive tendencies, with significant contributions from the EAT questionnaire of eating disorders among others. And yet, this factor is also particularly susceptible to influences from surface-level questionnaire filling-behavior, negatively weighting reversed items (thereby undoing the semantic coding reversal) and positively weighting highly skewed items from all questionnaires. Indeed, high CIT scores are associated with poor task performance when no staircasing is employed (15, 16), as is expected from inattentive participants, but not necessarily from compulsive individuals. Furthermore, while CIT correlated with explicit confidence ratings (which are subject to acquiescence biases), it is not normally found to correlate with implicit measures of confidence such as information seeking (15) and postdecision wagering (69). An important exception is a recent finding of a negative association between CIT and reminder-setting (2), which may reflect an implicit signature of CIT overconfidence, reduced engagement with the task, or a combination of both. Notably, the OCI-R questionnaire suffers from similar vulnerabilities, with no control for acquiescence (no reversed items),

and multiple items that are rarely endorsed by attentive participants. In our experiment, OCI-R scores were significantly higher among inattentive participants, and correlated with acquiescence, measured as response to content-neutral items. Given these issues, we note that researchers should be mindful of overinterpreting links between compulsivity or OC tendencies and self-report or task performance.

We suggest one more factor that may contribute to higher confidence among inattentive responders—gender. Roughly two thirds of inattentive responders in our study were self-declared males, compared to roughly half of all attentive responders ($\chi^2(1, n = 188) = 3.21$, $P = 0.073$; *SI Appendix*, Fig. SA12). Given that male participants were, on average, more confident than female participants ($t(184.87) = 3.65$, $P < 0.001$; Cohen's d = 0.53, 95% CI [0.24, 0.82]; *SI Appendix*, Fig. SA13), it is possible that part of the association between inattentiveness and high confidence is related to these gender differences.

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 all studies reanalyzed here (6, 8, 17), and more recent studies incorporated infrequency items as well (3, 14). 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 (24). 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 (*SI Appendix*, Fig. SA15). This means that even with four infrequency items, 15.5% of all inattentive participants in our sample were not classified as such. Given that detection rates increase substantially with each additional infrequency item, studies relying on only one or two such items cannot rule out residual confounding effects of inattention (we elaborate on this further in our Reply to Gillan et al., 2025; section 2). 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 *SI 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; (70)].

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 postdecisional wagering ("am I confident enough to bet on this decision?") (71–73) and information seeking ("do I require more evidence before committing a decision?") (74–76). Similarly, it has been suggested that decisions about absence, experimentally measured as decisions about missing targets and nonlearned words, open a window into metacognitive knowledge about perception ("I would have seen the target if it was present") (77–79) and memory ("I would have remembered this face if I had seen it before") (80, 81).

Finally, whenever theory-based predictions about interactions between metacognition and test conditions are implied by theoretical models, 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') (82), a difficulty to generate a feeling of knowing ('yedasentience') (83), or attenuated access to one's internal states, including memory (84). 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 behaviors. Causal interventions can provide an additional support for a true link between metacognition and mental health. For example, Fox et al. (38) 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 behavior 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 behavior 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.

Methods

Analysis of Existing Datasets. We report a reanalysis of data from four published articles that include both raw scores of psychiatric inventories and data from a cognitive task with confidence rating and that publicly shared their raw data: Rouault et al. (6); Seow and Gillan (8); Hoven et al. (17); Katyal et al. (14). More details on these datasets is available in *SI Appendix, Supplementary Methods*.

Assessing Acquiescence.

Rating delta between standard and reversed items. To measure acquiescence, we calculated for each participant their rating delta, as the mean rating for standard items minus mean rating for reversed items (after reversal) across all inventories. We used this rating delta as a proxy for acquiescence, as these studies did not include items with neutral content that could be used to independently assess acquiescence (21). This method allowed us to identify consistent agreement/disagreement patterns across diverse content. Positive acquiescence is expected to result in large positive rating deltas, as individuals would rate both standard and reversed items highly, creating a substantial difference after reverse-scoring the reversed items (e.g., endorsing both SDS 1 'I feel down-hearted and blue' and SDS 17 'I feel that I am useful and needed' would yield a large positive delta after reverse-scoring). Negative acquiescence is expected to produce large negative rating deltas, as individuals would rate both item types low, again creating a large difference in the opposite direction after reverse-scoring (e.g., low endorsement of both items, would yield a large negative delta).

Reversed items inconsistency. Another marker of acquiescence is an inconsistency between responses to reversed and regular items (21). 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 ref. 22 for a review).

Here, we were particularly interested in a specific phenomenon discussed by refs. 23, 24, and 85 whereby inattentive participants appear symptomatic when symptom frequencies are rare (Fig. 1*B* 'rare symptom'). Specifically, Zorowitz et al. (24) found that when a self-report inventory probes for symptoms 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. (24); Fig. 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 ref. 85). 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 (86).

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. The online experiment was approved by the Research Ethics Committee of Tel-Aviv University (Study ID No. 0009312-1). All participants provided informed consent before beginning the experiment. After providing 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 ms 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 task started with 25 practice trials. 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 s to emphasize the error, whereas feedback for correct selections was shown for 1.5 s. 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. After finishing the task, participants answered two comprehension questions (details in *SI Appendix, Supplementary Methods*). Then, participants were redirected to Qualtrics to complete the self-report section, which comprised the OCI-R (30) SDS (27), and the Content-neutral items inventory. Each psychiatric questionnaire included two "infrequency" items to assess inattentive responding (details in *SI Appendix, Supplementary Methods*).

The complete set of survey items, including participants' instructions, is provided in *SI Appendix* section *Full Survey Questions*. The presentation order of the OCI-R, SDS, and mean neutral items inventory was randomized between participants. Within each questionnaire, item order was also randomized, with the constraint that infrequency items were never presented as the first item in either the OCI-R or SDS.

Sampling design, modeling, and weighting assumptions. Participants were recruited via Prolific online research platform.

No modeling or weighting assumptions were applied. Respondent recruitment and question-related panel conditioning factors

Participants were English-speaking located in the United Kingdom or United States. Participants who took part in previous pilot studies were excluded from participating again.

Attrition rates and potential implications of attrition. Five participants were excluded for an average accuracy below 60%, in accordance with our preregistered exclusion criterion.

Data, Materials, and Software Availability. Our analysis code for both the reanalyses of existing datasets and analysis for our experiment are available on GitHub at https://github.com/self-model/BIRDAM (87). Anonymized data from our online experiment are openly available on OSF at https://osf.io/6npd9/ (88). This study also reanalyzed previously published datasets from Rouault et al. (6), Seow and Gillan (8), Hoven et al. (17) Gillan et al. (12), and Katyal et al. (14), which are available from their original publications and repositories. The datasets as used in the present analyses are available on OSF at https://osf.io/6npd9/ (88).

ACKNOWLEDGMENTS. We wholeheartedly wish to thank Tricia Seow, Marion Rouault, Sucharit Katyal, Ruth Van Holst, 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, Ido Ben Artzi, Chris Benwell, and Roni Maimon-Mor for their valuable feedback. Finally, we thank Julia Rohrer for useful advice regarding collider bias in multiple regression.

Author affiliations: [a]School of Psychological Sciences, Tel Aviv University, Tel Aviv 6997801, Israel; [b]All Souls College, University of Oxford, OX1 4AL Oxford, United Kingdom; and [c]Department of Experimental Psychology, University of Oxford, OX1 3EL Oxford, United Kingdom

1. C. M. Gillan, R. Whelan, What big data can do for treatment in psychiatry. *Curr. Opin. Behav. Sci.* **18**, 34–42 (2017).
2. A. Boldt, C. A. Fox, C. M. Gillan, S. Gilbert, Transdiagnostic compulsivity is associated with reduced reminder setting, only partially attributable to overconfidence. *Elife* **13**, RP98114 (2025).
3. C. A. Fox *et al.*, Reliable, rapid, and remote measurement of metacognitive bias. *Sci. Rep.* **14**, 14941 (2024).
4. C. M. Gillan, N. D. Daw, Taking psychiatry research online. *Neuron* **91**, 19–23 (2016).
5. Q. J. M. Huys, T. V. Maia, M. J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications. *Biol. Psychiatry Nat. Neurosci.* **19**, 404–413 (2016).
6. M. Rouault, T. Seow, C. M. Gillan, S. M. Fleming, Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* **84**, 443–451 (2018).
7. T. X. Seow, M. Rouault, C. M. Gillan, S. M. Fleming, How local and global metacognition shape mental health. *Biol. Psychiatry* **90**, 436–446 (2021).
8. T. X. Seow, C. M. Gillan, Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Sci. Rep.* **10**, 1–11 (2020).
9. T. Wise, O. J. Robinson, C. M. Gillan, Identifying transdiagnostic mechanisms in mental health using computational factor modeling. *Biol. Psychiatry* **93**, 690–703 (2023).

10. T. Wise, R. J. Dolan, Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nat. Commun.* **11**, 4179 (2020).
11. T. Insel *et al.*, Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* **167**, 748–751 (2010).
12. C. M. Gillan, M. Kosinski, R. Whelan, E. A. Phelps, N. D. Daw, Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife* **5**, e11305 (2016).
13. T. X. F. Seow, S. M. Fleming, T. U. Hauser, Metacognitive biases in anxiety-depression and compulsivity extend across perception and memory. *PLoS Ment. Health* **2**, e0000259 (2025).
14. S. Katyal, Q. J. Huys, R. J. Dolan, S. M. Fleming, Distorted learning from local metacognition supports transdiagnostic underconfidence. *Nat. Commun.* **16**, 1854 (2025).
15. G. Mohr, R. A. A. Ince, C. S. Y. Benwell, Information search under uncertainty across transdiagnostic psychopathology and healthy ageing. *Transl. Psychiatry* **14**, 353 (2024).
16. C. S. Y. Benwell, G. Mohr, J. Wallberg, A. Kouadio, Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *NPJ Ment. Health Res.* **1**, 10 (2022).
17. M. Hoven, J. Luigjes, D. Denys, M. Rouault, R. J. van Holst, How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nat. Ment. Health* **1**, 337–345 (2023).
18. M. Hoven *et al.*, Abnormalities of confidence in psychiatry: An overview and future perspectives. *Transl. Psychiatry* **9**, 268 (2019).
19. P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, N. P. Podsakoff, Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**, 879–903 (2003).
20. R. E. McGrath, M. Mitchell, B. H. Kim, L. Hough, Evidence for response bias as a source of error variance in applied assessment. *Psychol. Bull.* **136**, 450–470 (2010).
21. B. Weijters, H. Baumgartner, N. Schillewaert, Reversed item bias: An integrative model. *Psychol. Methods* **18**, 320–334 (2013).
22. A. W. Meade, S. B. Craig, Identifying careless responses in survey data. *Psychol. Methods* **17**, 437–455 (2012).
23. J. Chandler, I. Sisso, D. Shapiro, Participant carelessness and fraud: Consequences for clinical research and potential solutions. *J. Abnorm. Psychol.* **129**, 49–55 (2020).
24. S. Zorowitz, J. Solis, Y. Niv, D. Bennett, Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nat. Hum. Behav.* **7**, 1667–1681 (2023).
25. J. B. Saunders, O. G. Aasland, T. F. Babor, J. R. De La Fuente, M. Grant, Development of the alcohol use disorders identification test (audit): Who collaborative project on early detection of persons with harmful alcohol consumption-ii. *Addiction* **88**, 791–804 (1993).
26. R. S. Marin, R. C. Biedrzycki, S. Firinciogullari, Reliability and validity of the apathy evaluation scale. *Psychiatry Res.* **38**, 143–162 (1991).
27. W. W. Zung, A self-rating depression scale. *Arch. Gen. Psychiatry* **12**, 63–70 (1965).
28. D. M. Garner, M. P. Olmsted, Y. Bohr, P. E. Garfinkel, The eating attitudes test: Psychometric features and clinical correlates. *Psychol. Med.* **12**, 871–878 (1982).
29. J. H. Patton, M. S. Stanford, E. S. Barratt, Factor structure of the Barratt impulsiveness scale. *J. Clin. Psychol.* **51**, 768–774 (1995).
30. E. B. Foa *et al.*, The obsessive-compulsive inventory: Development and validation of a short version. *Psychol. Assess.* **14**, 485 (2002).
31. O. Mason, Y. Linney, G. Claridge, Short scales for measuring schizotypy. *Schizophr. Res.* **78**, 293–296 (2005).
32. M. R. Liebowitz, Social phobia. Moder prob. *Pharmacopsychiatry* **22**, 141–173 (1987).
33. C. D. Spielberger, R. Gorsuch, R. Lushene, P. R. Vagg, G. A. Jacobs, *Manual for the State-Trait Anxiety Inventory* (Consulting Psychologists Press, Palo Alto, CA, 1983).
34. R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **166**, 1092–1097 (2006).
35. J. L. Huang, N. A. Bowling, M. Liu, Y. Li, Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *J. Bus. Psychol.* **30**, 299–311 (2015).
36. E. A. Greenleaf, Measuring extreme response style. *Public Opin. Q.* **56**, 328–351 (1992).
37. T. U. Hauser *et al.*, Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Sci. Rep.* **7**, 6614 (2017).
38. C. A. Fox *et al.*, An observational treatment study of metacognition in anxious-depression. *Elife* **12**, RP87193 (2023).
39. R. Dar, Elucidating the mechanism of uncertainty and doubt in obsessive-compulsive checkers. *J. Behav. Ther. Exp. Psychiatry* **35**, 153–163 (2004).
40. R. Dar, A. Lazarov, N. Liberman, Seeking proxies for internal states (SPIS): Towards a novel model of obsessive-compulsive disorder. *Behav. Res. Ther.* **147**, 103987 (2021).
41. S. A. Rasmussen, J. L. Eisen, Clinical features and phenomenology of obsessive-compulsive disorder. *Psychiatr. Ann.* **19**, 67–73 (1989).
42. J. R. Cougle, P. M. Salkovskis, K. Wahl, Perception of memory ability and confidence in recollections in obsessive-compulsive checking. *J. Anxiety Disord.* **21**, 118–130 (2007).
43. F. Karadag, N. Oguzhanoglu, O. Ozdel, F. C. Atesci, T. Amuk, Memory function in patients with obsessive compulsive disorder and the problem of confidence in their memories: A clinical study. *Croat. Med. J.* **46**, 282–287 (2005).
44. T. Marton *et al.*, Validating a dimension of doubt in decision-making: A proposed endophenotype for obsessive-compulsive disorder. *PLoS One* **14**, e0218182 (2019).
45. R. J. McNally, P. A. Kohlbeck, Reality monitoring in obsessive-compulsive disorder. *Behav. Res. Ther.* **31**, 249–253 (1993).
46. W. Zitterl *et al.*, Memory deficits in patients with DSM-IV obsessive-compulsive disorder. *Psychopathology* **34**, 113–117 (2001).
47. R. Dar, N. Sarna, G. Yardeni, A. Lazarov, Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychol. Med.* **52**, 2404–2412 (2022).
48. J. Westfall, T. Yarkoni, Statistically controlling for confounding constructs is harder than you think. *PLoS One* **11**, e0152719 (2016).
49. A. T. Beck, K. Bredemeier, A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clin. Psychol. Sci.* **4**, 596–619 (2016).
50. J. Hancock, "Depressive realism" assessed via confidence in decision making. *Cogn. Neuropsychiatry* **1**, 213–220 (1996).
51. D. Richards, Prevalence and clinical course of depression: A review. *Clin. Psychol. Rev.* **31**, 1117–1125 (2011).
52. T. Szu-Ting Fu, W. Koutstaal, L. Poon, A. J. Cleare, Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *J. Behav. Ther. Exp. Psychiatry* **43**, 699–704 (2012).
53. R. P. Bagozzi, Y. Yi, Assessing method variance in multitrait-multimethod matrices: The case of self-reported affect and perceptions at work. *J. Appl. Psychol.* **75**, 547–560 (1990).
54. H. Baumgartner, J. B. E. Steenkamp, Response styles in marketing research: A cross-national investigation. *J. Mark. Res.* **38**, 143–156 (2001).
55. D. T. Campbell, D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81 (1959).
56. P. G. Curran, Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* **66**, 4–19 (2016).
57. J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, R. P. DeShon, Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* **27**, 99–114 (2012).
58. D. S. Nichols, R. L. Greene, P. Schmolck, Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *J. Clin. Psychol.* **45**, 239–250 (1989).
59. Y. Ophir, I. Sisso, C. S. C. Asterhan, R. Tikochinski, R. Reichart, The turker blues: Hidden factors behind increased depression rates among Amazon's mechanical turkers. *Clin. Psychol. Sci.* **8**, 65–83 (2020).
60. P. E. Spector, Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *J. Appl. Psychol.* **72**, 438–443 (1987).
61. E. B. Foa, N. Amir, B. Gershuny, C. Molnar, M. J. Kozak, Implicit and explicit memory in obsessive-compulsive disorder. *J. Anxiety Disord.* **11**, 119–129 (1997).
62. D. Hermans *et al.*, Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behav. Res. Ther.* **46**, 98–113 (2008).
63. S. Moritz *et al.*, Enhanced perceived responsibility decreases metamemory but not memory accuracy in obsessive–compulsive disorder (OCD). *Behav. Res. Ther.* **45**, 2044–2052 (2007).
64. R. Dar, S. Rish, H. Hermesh, M. Taub, M. Fux, Realism of confidence in obsessive-compulsive checkers. *J. Abnorm. Psychol.* **109**, 673–678 (2000).
65. D. F. Tolin *et al.*, Memory and memory confidence in obsessive–compulsive disorder. *Behav. Res. Ther.* **39**, 913–927 (2001).
66. A. Lazarov, R. Dar, N. Liberman, Y. Oded, Obsessive–compulsive tendencies may be associated with attenuated access to internal states: Evidence from a biofeedback-aided muscle tensing task. *Conscious. Cogn.* **21**, 1401–1409 (2012).
67. Z. Zhang *et al.*, Individuals with high obsessive-compulsive tendencies or undermined confidence rely more on external proxies to access their internal states. *J. Behav. Ther. Exp. Psychiatry* **54**, 263–269 (2017).
68. M. Hoven, M. Rouault, R. Van Holst, J. Luigjes, Differences in metacognitive functioning between obsessive–compulsive disorder patients and highly compulsive individuals from the general population. *Psychol. Med.* **53**, 7933–7942 (2023).
69. S. Sookud, I. Martin, C. M. Gillan, T. Wise, Impaired goal-directed planning in transdiagnostic compulsivity is explained by uncertainty about learned task structure. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **11**, 357–365 (2026).
70. B. Maniscalco, H. Lau, A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21**, 422–430 (2012).
71. A. Ben Shachar, A. Lazarov, M. Goldsmith, R. Moran, R. Dar, Exploring metacognitive components of confidence and control in individuals with obsessive-compulsive tendencies. *J. Behav. Ther. Exp. Psychiatry* **44**, 255–261 (2013).
72. E. Hembacher, S. Ghetti, Subjective experience guides betting decisions beyond accuracy: Evidence from a metamemory illusion. *Memory* **25**, 575–585 (2017).
73. N. Persaud, P. McLeod, A. Cowey, Post-decision wagering objectively measures awareness. *Nat. Neurosci.* **10**, 257–261 (2007).
74. M. H. Siegel, R. W. Magid, M. Pelz, J. B. Tenenbaum, L. E. Schulz, Children's exploratory play tracks the discriminability of hypotheses. *Nat. Commun.* **12**, 3598 (2021).
75. L. Schulz, S. M. Fleming, P. Dayan, Metacognitive computations for information search: Confidence in control. *Psychol. Rev.* **130**, 604–639 (2023).
76. D. Selmeczy, I. G. Dobbins, Metacognitive awareness and adaptive recognition biases. *J. Exp. Psychol. Learn. Mem. Cogn.* **39**, 678–690 (2013).
77. M. Mazor, S. M. Fleming, Efficient search termination without task experience. *J. Exp. Psychol. Gen.* **151**, 2494–2510 (2022).
78. M. Mazor, R. Moran, C. Press, Beliefs about perception shape perceptual inference: An ideal observer model of detection. *Psychol. Rev.* **133**, 271–295 (2026).
79. N. Sarna, M. Mazor, R. Dar, Obsessive-compulsive visual search: A reexamination of presence–absence asymmetries. *Clin. Psychol. Sci.* **13**, 425–433 (2024).
80. S. Ghetti, Memory for nonoccurrences: The role of metacognition. *J. Mem. Lang.* **48**, 722–739 (2003).
81. M. Mazor, Inference about absence as a window into the mental selfmodel. *Open Mind* **9**, 635–651 (2025).
82. S. Rachman, R. Shafran, Cognitive distortions: Thought–action fusion. *Clin. Psychol. Psychother.* **6**, 80–85 (1999).
83. H. Szechtman, E. Woody, Obsessive-compulsive disorder as a disturbance of security motivation. *Psychol. Rev.* **111**, 111–127 (2004).
84. N. Liberman, A. Lazarov, R. Dar, Obsessive-compulsive disorder: The underlying role of diminished access to internal states. *Curr. Dir. Psychol. Sci.* **32**, 118–124 (2023).
85. K. M. King, D. S. Kim, C. J. McCabe, Random responses inflate statistical estimates in heavily skewed addictions data. *Drug Alcohol Depend.* **183**, 102–110 (2018).
86. L. Komsta, F. Novomestky, moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. R package version 0.14.1. https://CRAN.R-project.org/package=moments. Accessed 5 March 2025.
87. N. Sarna, R. Dar, M. Mazor, Analysis code for "Biased and inattentive responding contribute to apparent metacognitive biases in mental health." GitHub. https://github.com/self-model/BIRDAM. Deposited 5 March 2025.
88. N. Sarna, R. Dar, M. Mazor, Data from "Biased and inattentive responding contribute to apparent metacognitive biases in mental health." Open Science Framework. https://osf.io/6npd9/. Deposited 5 March 2025.