Metacognitive biases in anxious-depression and compulsivity extend across perception and memory

Tricia X. F. Seow[1,2]*, Stephen. M. Fleming[1,2,3], Tobias U. Hauser[1,2,4,5]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London

[2]Wellcome Centre for Human Neuroimaging, University College London

[3]Department of Experimental Psychology, University College London

[4]Department of Psychiatry and Psychotherapy, Faculty of Medicine, University of Tübingen, Tübingen, Germany

[5]German Center for Mental Health (DZPG), partner site Tübingen

*Corresponding author
t.seow@ucl.ac.uk

Abstract

Metacognitive biases are characteristic of common mental health disorders like depression and obsessive-compulsive disorder (OCD). However, recent transdiagnostic approaches consistently contradict traditional clinical studies, with overconfidence in perception among highly compulsive individuals versus underconfident memory in OCD patients. To reconcile these differences, we investigated whether these metacognitive divergences arise due to cognitive domain-specific effects, comorbid overshadowing effects, and/or different manifestations at disparate levels of a metacognitive hierarchy. Using a transdiagnostic individual differences approach (N=327), we quantified metacognitive patterns across memory and perception. Across cognitive domains, we found that underconfidence was linked to anxious-depression and overconfidence was linked to compulsivity. These associations varied across the confidence hierarchy, with anxious-depression being predominantly explained by global low self-esteem, whereas compulsivity exhibited more specific alterations at lower metacognitive levels. Our results support a domain-general alteration of metacognition in psychopathology, with differential contributions from distinct levels of a metacognitive hierarchy, akin to an overshadowing effect.

Introduction

Altered insight is a hallmark of many psychiatric disorders[1]. In experimental studies, quantifying metacognition provides one route towards studying the processes contributing to distorted insight[2], and consequently has been increasingly examined in clinical populations. Metacognition is often studied by asking people to evaluate their confidence in their performance in a particular domain, with "metacognitive bias" referring to cases in which these self-evaluations diverge from reality[3].

Several patient studies have found metacognitive biases across multiple psychiatric disorders[4]. For instance, patients with depression[5–7] and obsessive-compulsive disorder (OCD)[8] consistently express lowered confidence in their decisions/actions. The latter disorder is even cast as a 'disorder of doubt'[9,10] and lowered confidence was found across various memory paradigms[11–16]. However, recent work have observed more differentiated effects when adopting a transdiagnostic approach in metacognitive research[17]. These studies, including our own, report that individuals with high 'anxious-depression' scores express metacognitive underconfidence, whilst individuals with high 'compulsivity and intrusive thoughts' (hereafter: 'compulsivity') scores express overconfidence[18–23]. Taken together, these findings suggest a striking and hitherto unexplained dissociation in which transdiagnostic compulsivity is linked to overconfidence, whereas a diagnosis of OCD is linked to underconfidence.

The reasons for these opposing findings are unclear, with a range of explanations on offer. First, metacognition may only partially generalise across cognitive domains[24–29] meaning that poor metacognition in memory tasks does not necessarily predict poor metacognition in perceptual decision tasks (termed domain-generality). Metacognitive patterns may manifest differently between these domains, perhaps explaining findings of underconfidence in one domain, and overconfidence in another. Second, different findings may arise due a mixture of psychopathological symptoms across the different samples[30]. Due to high co-morbidity across disorders (e.g., depression in OCD patients), the effect of one symptom dimension could overshadow or mimic the effect of another i.e., OCD patients may exhibit lowered confidence due to their co-morbid depression, rather than their primary OCD symptoms[22,31]. Third, we and others have recently shown that the expression of confidence may reflect an internal hierarchy, ranging from local confidence in single decisions to high-level global assessments such as self-esteem[19,32,33]. Metacognitive biases across different psychopathologies could manifest at different levels of a confidence hierarchy, leading to apparent dissociations between different studies[17].

Here, we set out to disentangle these three possible explanations of conflicting findings relating metacognitive biases to compulsivity and OCD by testing metacognition across the cognitive domains of perception and memory using a transdiagnostic individual differences approach. We investigated the involvement of the different hierarchical levels of metacognition by probing their contributions to dimensions of anxious-depression and compulsivity. Our results (N=327) show a domain-general effect of low confidence linked to anxious-depression and high confidence linked to compulsivity, suggesting that the previously observed differences of memory underconfidence in OCD versus perceptual overconfidence in compulsivity are unlikely to arise due to domain-specific metacognitive differences, and are more likely due to overshadowing effects. Moreover, we show that mental health dimensions are linked to metacognitive impairments at distinct levels of the metacognitive hierarchy—with predominant contributions of low global self-esteem to anxious-depression versus bi-valenced multi-hierarchical contributions of confidence to compulsivity, including lower global task confidence and self-esteem together with higher self-efficacy and local task confidence.

Results

To evaluate metacognitive biases across hierarchies and domains, we conducted a large-scale (N=327) online study with a general population sample. Each participant completed, in a randomly assigned order, a memory and a perceptual task designed to measure performance and confidence in these cognitive domains (Fig. 1). On each trial, participants performed a two-alternative forced-choice (stimulus identification in previously shown picture array (memory); numerosity discrimination (perception)) before rating their confidence in their decision on a continuous confidence scale. Performances for both tasks were equalised using a staircase procedure (Supplementary Fig. 1 & 2). Thereafter, participants completed a battery of self-report questionnaires assessing a range of psychopathological symptoms, which were transformed into transdiagnostic dimension scores encompassing ‘anxious-depression’, ‘compulsive behaviour and intrusive thought’ (compulsivity) and ‘social withdrawal’[34].

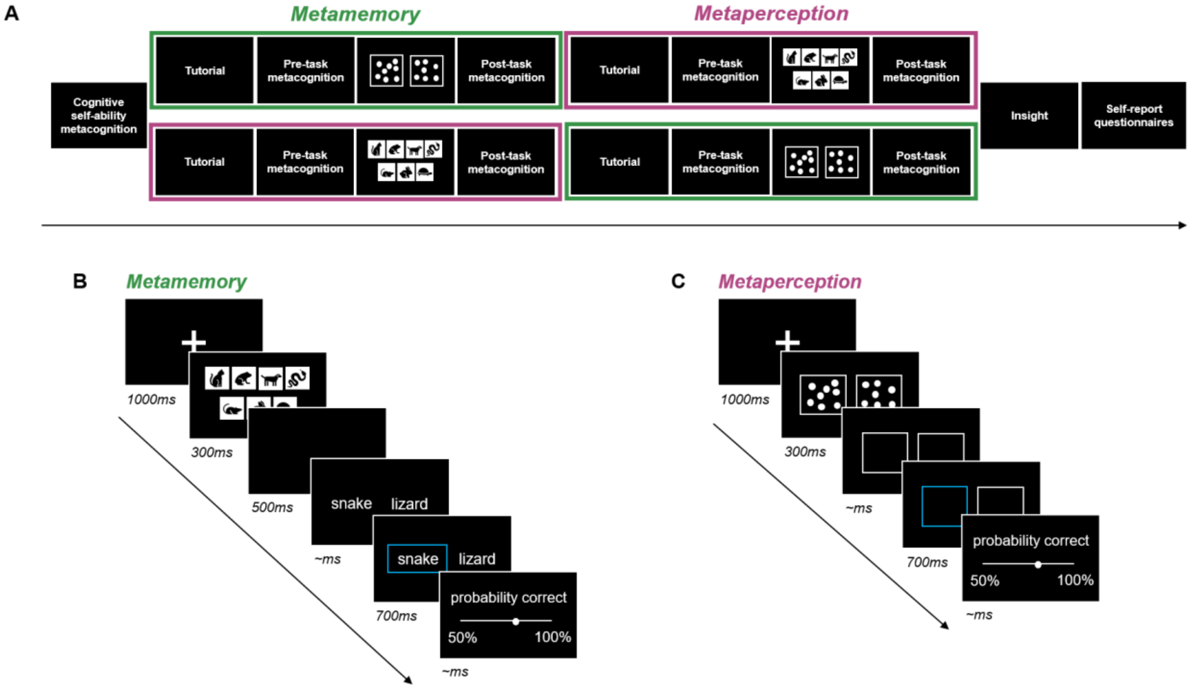


Fig. 1. Experimental task schedule and paradigms. (A) Task schedule. *Participants were randomly allocated to complete either the metamemory or metaperception task set first.* ***(B) Metamemory task.*** *Participants were to select the word of the stimuli that was previously displayed amongst an array.* ***(C) Metaperception task.*** *Participants were to select the box which displayed more dots. In both* ***(B)*** *&* ***(C)****, confidence ratings of the decisions were made on a continuous scale (50%-100% probability correct) for every trial.*

Local task confidence is closely linked across perception and memory

We first tested if there was a domain-general relationship between confidence across the two tasks. To ensure that confidence levels were not driven by differences in task performance, we examined task accuracies to ensure that the staircase procedures were successful in equating performance between individuals and tasks. Reassuringly, accuracy was well-calibrated for both memory (mean (M)=71.40% correct, standard deviation (SD)=2.22%) and perception (M=71.85% correct, SD=2.36%), with no significant difference between tasks (t(325)=0.57, p=0.57, 95% Confidence Interval (CI)=[-0.08 0.14]) (Fig. 2A). As accuracy was titrated, task difficulty (estimated as the titrated average dot numerosity difference (metaperception) or average stimuli number of picture array shown (metamemory)) reflected the participants' endogenous perceptual/working-memory ability. Higher difficulty was indexed by smaller (log) dots difference (M=3.20, SD=0.46) in the metaperception task and higher number of stimuli (M=7.94, SD=1.81) in the metamemory task (Fig. 2B). We found that average task difficulty was correlated across tasks (r=-0.28, p<0.001, 95% CI=[-0.38 -0.18]), indicating that participants who achieved a higher difficulty level at one task also did so in the other.

Given that task accuracies were similar between participants, we next examined confidence for the two tasks by estimating the mean trial-by-trial confidence level as a local metacognitive bias measure. We found that participants were overall slightly more confident (t(325)=2.79, p=0.006, 95% CI=[0.29 1.66]) in the metaperception task (M=76.70, SD=8.19) as compared to the metamemory task (M=75.73, SD=6.51) (Fig. 2C). Mean confidence ratings were highly positively correlated across tasks (r=0.65, p<0.001, 95% CI=[0.59 0.71]), and remain significantly correlated even after controlling for task difficulty (r=0.48, p<0.001, 95% CI=[0.39 0.56]), supporting a domain-general pattern of confidence.

We noted that correct trials had a higher mean confidence rating than incorrect trials for both memory (M=15.83, SD=5.39) and perception (M=5.69, SD=4.08) (Fig. 2D), indicating that participants effectively used their performance to guide their confidence ratings. We also saw a positive correlation between tasks of the difference between correct and incorrect trial confidence (r=0.17, p=0.002, 95% CI=[0.06 0.27]), suggesting a domain-general relationship between metacognitive sensitivity to performance as well. This was further supported by a positive cross-task correlation (mean group-level covariance ρ=0.11, 95% highest density interval (HDI)=[-0.10 0.31]) in metacognitive efficiency (meta-d'/d'; the ability to distinguish between correct and incorrect judgments with confidence ratings) estimated within a hierarchical Bayesian model (HMeta-d[35]; Supplementary Fig. 7).

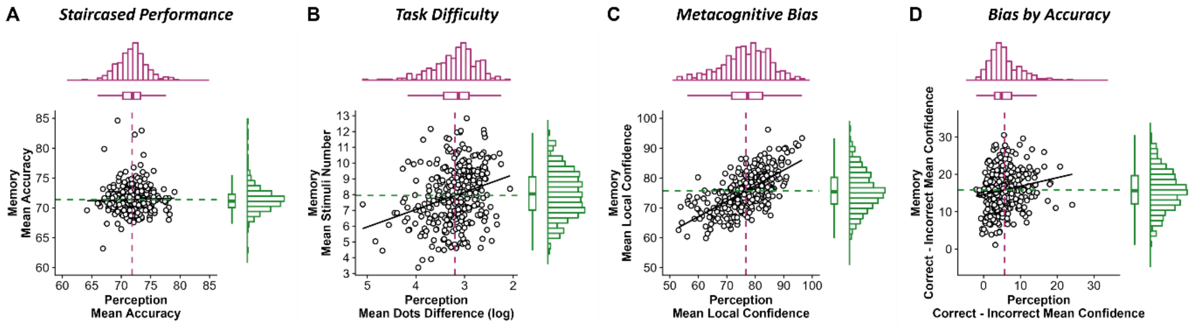


Fig. 2. Performance and confidence distributions from metaperceptual and metamemory tasks. (A) Task accuracy. *Both tasks reported similar levels of accuracy controlled by staircase procedures.* ***(B) Task difficulty.*** *As participants' accuracies were titrated, the difficulty at which the participants performed the task is another indicator of performance. Metaperception task difficulty level was indicated by the mean (log) dot difference between the stimuli (smaller difference = more difficult), while metamemory task difficulty level was indicated by the mean number of stimuli in the array during presentation time (more stimuli = more difficult), over the course of the experiment.* ***(C) Task confidence level.*** *Local metacognitive bias was estimated by the mean trial-by-trial confidence ratings of each task. Bias was positively correlated between tasks, suggesting a domain-general effect of confidence.* ***(D) Task confidence difference between correct and incorrect trials****. Correct trials garner higher confidence ratings than incorrect trials, indicating metacognitive awareness of performance. For hierarchical Bayesian model analyses of metacognitive efficiency, see Supplementary Fig. 7. For* ***(A)****, axes are in percentages, while for* ***(B)*** *and* ***(C)****, axes are in percentage of probability correct. Each circle represents an individual participant, dot lines represent means, diagonal lines indicate linear relationships. Histograms on the top and right insets of each figure represent the distribution of participants for each variable.*

Anxious-depression is associated with underconfidence and compulsivity is associated with overconfidence

We next asked whether previously observed confidence biases in psychopathology traits extended across domains. We first assessed the link between mean confidence level and each self-report questionnaire score we collected, controlled for the influence of task domain, task completion order, age, IQ and gender. Echoing prior work in perception[23], we observed that several questionnaires, particularly mood dysregulation disorders symptoms (e.g., apathy, depression, social anxiety, trait anxiety, schizotypy), were linked to lower confidence levels (β<-0.11, SE<0.04, p<0.005) across perception and memory (Fig. 3A). We then refactored our questionnaire scores into mental health dimension scores (see **Methods**), quantifying anxious-depression, compulsivity and social withdrawal levels for each participant. With the dimension scores, we observed specific bi-valenced relationships with confidence—individuals high in anxious-depression exhibited lower confidence (β=-0.12, SE=0.04, p=0.006) and individuals high in compulsivity exhibited higher confidence (β=0.09, SE=0.04, p=0.04; Fig 3A).

To test whether these confidence relationships differed between memory and perception, we investigated the interaction between task domain and mental health dimensions. We found no significant interaction between task domain and anxious-depression (β=-0.01, SE=0.04, p=0.81), suggesting that the negative relationship between anxious-depression and confidence is similar across perception and memory tasks (Fig. 3B). However, there was a trending interaction between task domain and compulsivity (β=0.08, SE=0.04, p=0.07), where individual differences in compulsivity showed a stronger positive association with higher confidence in memory than perception. The stronger association in memory than perception was also illustrated in task-specific regressions (Supplementary Fig. 8).

Surprisingly, we also observed an association between social withdrawal and lower confidence (β=-0.17, SE=0.05, p<0.001) (Fig 3A). No prior study has reported such an association. There was no significant interaction effect between task domain and social withdrawal (β=0.03, SE=0.04, p=0.56), meaning that lowered confidence was present across the tasks. Finally, to ensure that these effects were not driven by task performance, we established that task accuracy (ps>0.44) and difficulty (ps>0.16) did not show significant relationships with symptom dimensions. Overall, our findings thus suggest that relationships between mental health dimensions and local task confidence are relatively similar across task domains, supporting a domain-general relationship between metacognition and psychopathology.

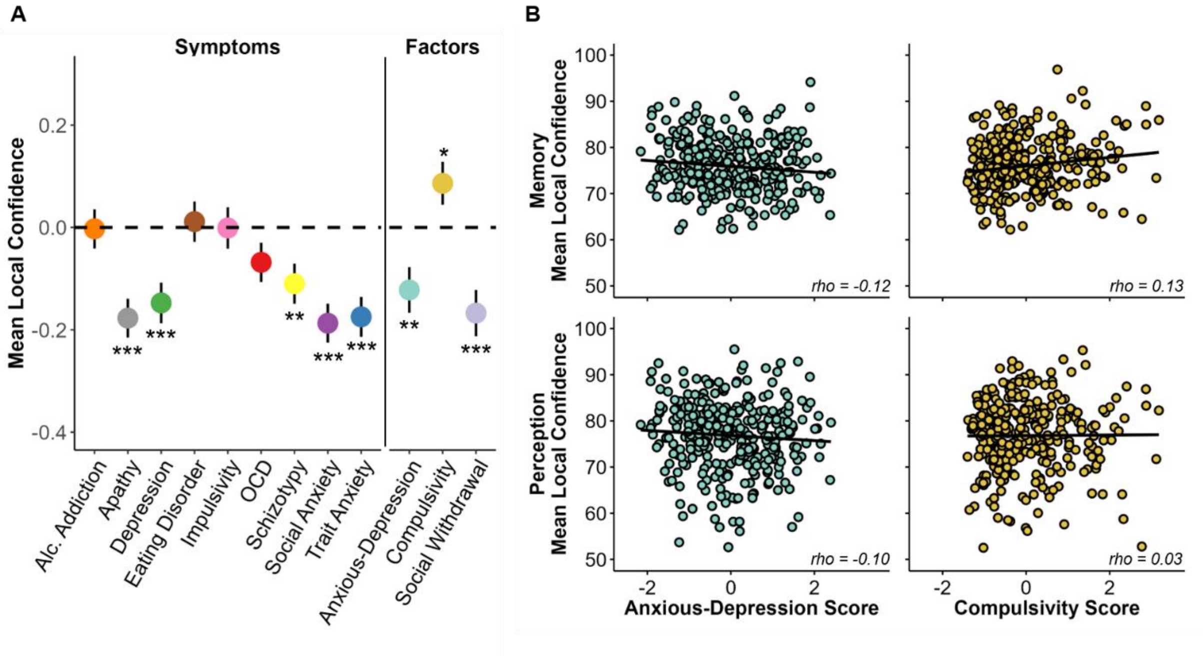


Fig. 3. Performance and confidence distributions from metaperceptual and metamemory task. (A) Regression predicting mean local confidence from questionnaire and dimension scores. *Higher apathy, depression, schizotypy, social anxiety and trait*

*anxiety scores were significantly linked to lower confidence levels. When we refactored the questionnaires scores into the transdiagnostic dimension scores, lower confidence was linked to anxious-depression and social withdrawal, while higher confidence was linked to compulsivity. Each questionnaire score was tested in a separate linear regression model, while dimension scores were included in the same model. The y-axis shows the z-scored change in mean confidence level as a function of 1 standard deviation increase of z-scored questionnaire/dimension scores. Error bars denote standard errors. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.* ***(B) Relationship between confidence and anxious-depression or compulsivity in metaperception or metamemory.*** *The relation of anxious-depression to lower confidence was similar between memory and perception, while for compulsivity, its relation to higher confidence was stronger in memory than perception. For illustrative purposes, we obtained the confidence residuals for each participant after controlling for anxious-depression/compulsivity (anxious-depression if compulsivity and confidence was to be examined, and vice versa), social withdrawal, task order, age, IQ and gender, separately for each task. We then correlated the confidence residuals and anxious-depression or compulsivity scores with Spearman's correlation. Each circle indicates the confidence (measured on a 50-100% probability correct scale) residuals and dimension score for each participant.*

<u>High compulsive individuals show blunted accuracy enhancement effect for harm versus non-harm stimuli</u>

Having established a link between memory confidence and compulsivity, we further investigated this association. Previous clinical metamemory studies have shown a stronger underconfidence effect for the recollection of "unsafe" objects (i.e., stimuli perceived to be harmful and trigger OCD behaviours) versus neutral (i.e., OCD irrelevant) ones[14]. We hypothesized that high compulsive individuals may similarly show a reduction in memory confidence for items identified as "harmful" (harm) versus "not harmful" (non-harm). In our metamemory task design, we thus purposefully chose a selection of high and low harm rated images from a prior rating study[36] as stimuli (Supplementary Fig. 9).

We first tested if trial accuracy would vary if the correct answer was a harm stimulus (versus non-harm), controlling for the contribution of task difficulty, age, IQ and gender. We found that trials where harm stimuli were the target stimulus were more likely to be answered correctly (β=0.11, SE=0.02, p<0.001) and with higher confidence (β=0.10, SE=0.009, p<0.001) (Fig. 4A), even when controlled for trial accuracy and difficulty. These findings replicate the boosting effect of "harmful" (affectively more salient) stimuli on memory, both in accuracy and confidence, even though the stimulus harmfulness was entirely task irrelevant.

We also hypothesized that high compulsive individuals would rate lower confidence on trials where harm items were the answer. We examined if confidence on each trial was linked to the stimulus category of the correct answer (controlling for trial accuracy, trial difficulty, age, IQ and gender). Contrary to expectations, we found no significant interaction effect of answer stimulus category with compulsivity on confidence (β=0.0008, SE=0.009, p=0.93) (Fig. 4B), suggesting that there was no impact of harm versus non-harm items on their confidence. Instead, we found that individuals high in compulsivity showed a trending decreased trial accuracy enhancement effect for harm stimuli (β=-0.05, SE=0.03, p=0.07) (Fig. 4B), where performance for harm and non-harm answer stimuli trials were more similar (Supplementary Fig. 9). This suggested that non-harmful stimuli were perceived as more similar to harmful stimuli in high compulsive participants.

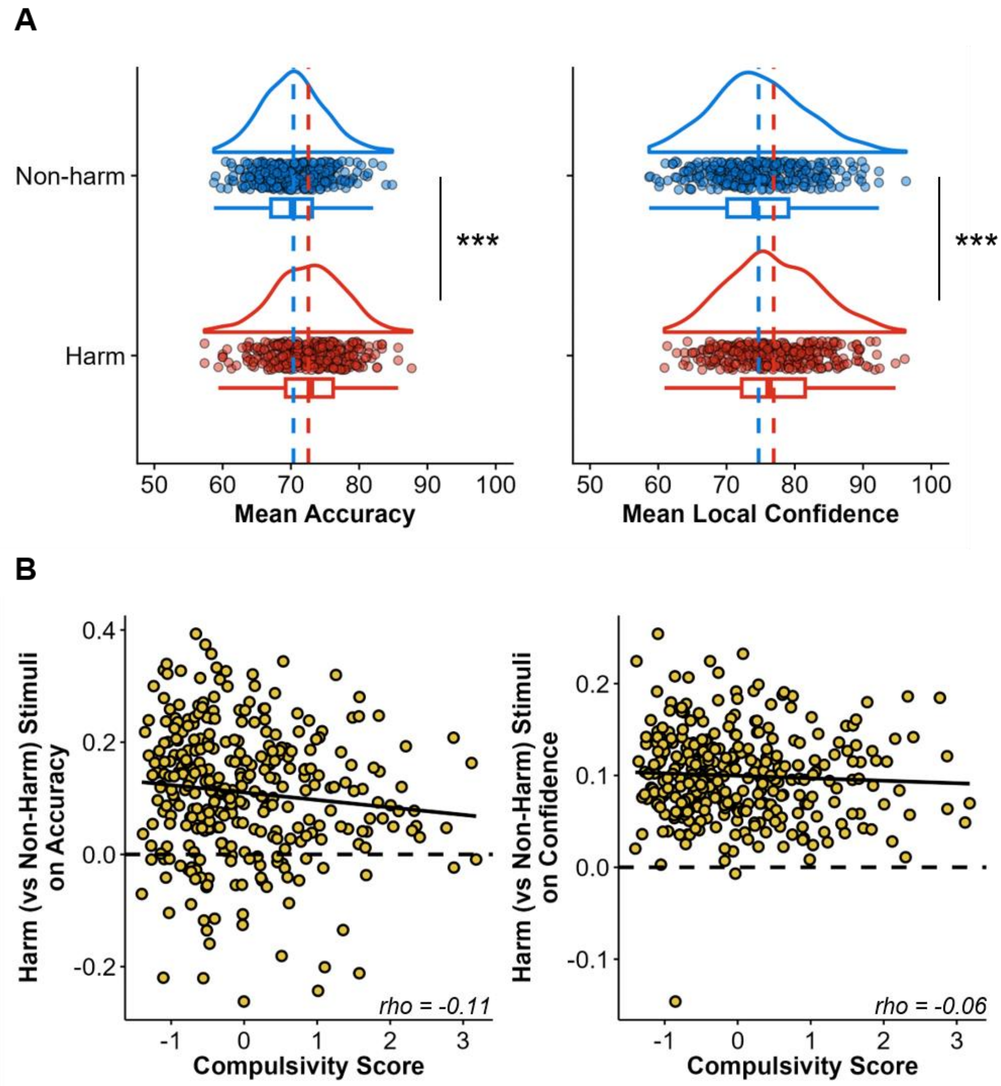


Fig. 4. Performance and confidence distributions, and their relation to compulsivity scores, from the metaperceptual and metamemory task. (A) Mean accuracy and confidence of trials split by whether the correct answer was a harm or non-harm stimulus*. Harm item trials were generally correct and had higher confidence ratings. Each circle represents each participant, the dotted lines represent the mean. ***p < 0.001.* ***(B) Correlation of compulsivity score with the stimuli category (harm versus non-harm) on accuracy or confidence****. For illustrative purposes, we obtained the accuracy or confidence*

*residual controlled for anxious-depression, social withdrawal, task difficulty, task accuracy (only for confidence), age, IQ and gender. The category*accuracy effect decreases as a function of increasing compulsivity scores, while category*confidence effect shows no significant relationship with compulsivity. Each circle represents each participant. Circles above the dotted lines indicate the boosting effect of harm (versus non-harm) stimuli on memory accuracy/confidence.*

Global metacognitive contributions to anxious-depression and compulsivity

Beyond trial-by-trial local confidence ratings, we and others have conceptualised metacognition operating across different hierarchical levels[17]. To assess links between these levels and psychopathology, we also obtained more global metrics of task-specific metacognition in the form of pre-task and post-task self-performance estimates (see **Methods**). We also asked participants to introspect on their global perceptual and memory abilities prior to completing the tasks ("self-ability estimates"; e.g., "how good do you think your memory/perceptual ability is?"), and complete self-report questionnaires to assess more general self-beliefs including self-esteem and domain-general self-efficacy[37,38]. We then tested if these global metacognitive estimates differed as a function of dimension scores, controlling for task domain and completion order (if relevant), age, IQ and gender.

We found no significant relationship between any of dimension scores with pre- ($\beta s>-0.07$, $SEs<0.06$, $ps>0.09$) and post-task ($\beta s>-0.15$, $SEs<0.05$, $ps>0.18$) global metacognition (Fig. 5A). However, we did find that individuals high in anxious-depression reported lower self-ability estimates ($\beta=-0.21$, $SE=0.05$, $p<0.001$), but no relationships with compulsivity ($\beta=0.07$, $SE=0.05$, $p=0.11$) or social withdrawal ($\beta=-0.05$, $SE=0.05$, $p=0.25$). There was no interaction effect between dimension scores and cognitive domain on self-ability estimates ($ps>0.78$), suggesting that anxious-depression was equally linked to lower self-evaluation of memory and perception abilities. For self-esteem, both anxious-depression ($\beta=-0.73$, $SE=0.03$, $p<0.001$) and social withdrawal ($\beta=-0.21$, $SE=0.03$, $p<0.001$) were linked to lower self-esteem, in the absence of relationships with compulsivity ($\beta=-0.03$, $SE=0.03$, $p=0.27$). For self-efficacy, anxious-depression ($\beta=-0.53$, $SE=0.05$, $p<0.001$) and social withdrawal ($\beta=-0.25$, $SE=0.05$, $p<0.001$) were similarly linked to lower self-efficacy scores while compulsivity was linked to higher scores ($\beta=0.12$, $SE=0.05$, $p=0.009$).

Differing local and global cumulative metacognitive contributions to anxious-depression and compulsivity

Lastly, a recent study has revealed differential associations between local (i.e., trial-by-trial (perceptual) task confidence ratings) and global (i.e., post-task confidence, self-esteem) measures of metacognition with dimension scores[19]. We extended this work by examining our six different local and global metacognitive metrics (across memory and perception) as we iteratively examined the importance of different metacognitive hierarchy levels in predicting psychiatric dimensions.

We started with a forward step-wise regression approach, utilising a base model of IQ, age and gender to predict dimension scores. We observed the impact of iterative addition of local confidence, pre-task metacognition, post-task metacognition, self-ability metacognition, self-esteem and self-efficacy on the model fit (Fig. 5B). For anxious-depression, low self-esteem (β=-0.76, SE=0.03, p<0.001) was the biggest contributor, followed by low self-efficacy (β=-0.11, SE=0.03, p<0.001), and low local task confidence (β=-0.06, SE=0.02, p=0.003). For compulsivity, the most impactful predictor was low self-esteem (β=-0.34, SE=0.04, p<0.001), then high self-efficacy (β=0.13, SE=0.04, p=0.002), low post-task metacognition (β=-0.07, SE=0.04, p=0.04) and high local task confidence (β=0.06, SE=0.04, p=0.12). Interestingly, we replicated a bi-valenced contribution of different metacognitive levels to compulsivity severity—high local confidence and low global self-esteem both emerged as key predictors[19]. To ensure the robustness of the step-wise regression models, we validated them by running a repeated 5-fold cross-validation procedure on 16 model variations for comparison (see Supplementary Fig. 11 for details). We found the same results as the step-wise regression, and saw out-of-sample prediction of the best metacognitive model was highly predictive for anxious-depression (r=0.86, p<0.001, 95% CI=[0.84 0.88]), and moderately predictive for compulsivity (r=0.44, p<0.001, 95% CI=[0.38 0.50]) (Fig. 5C).

Overall, our findings suggest that variance in anxious-depression is predominantly captured by the global factor of low self-esteem, while compulsivity seems to involve multiple and bi-valenced contributions of both local and global levels of metacognition.

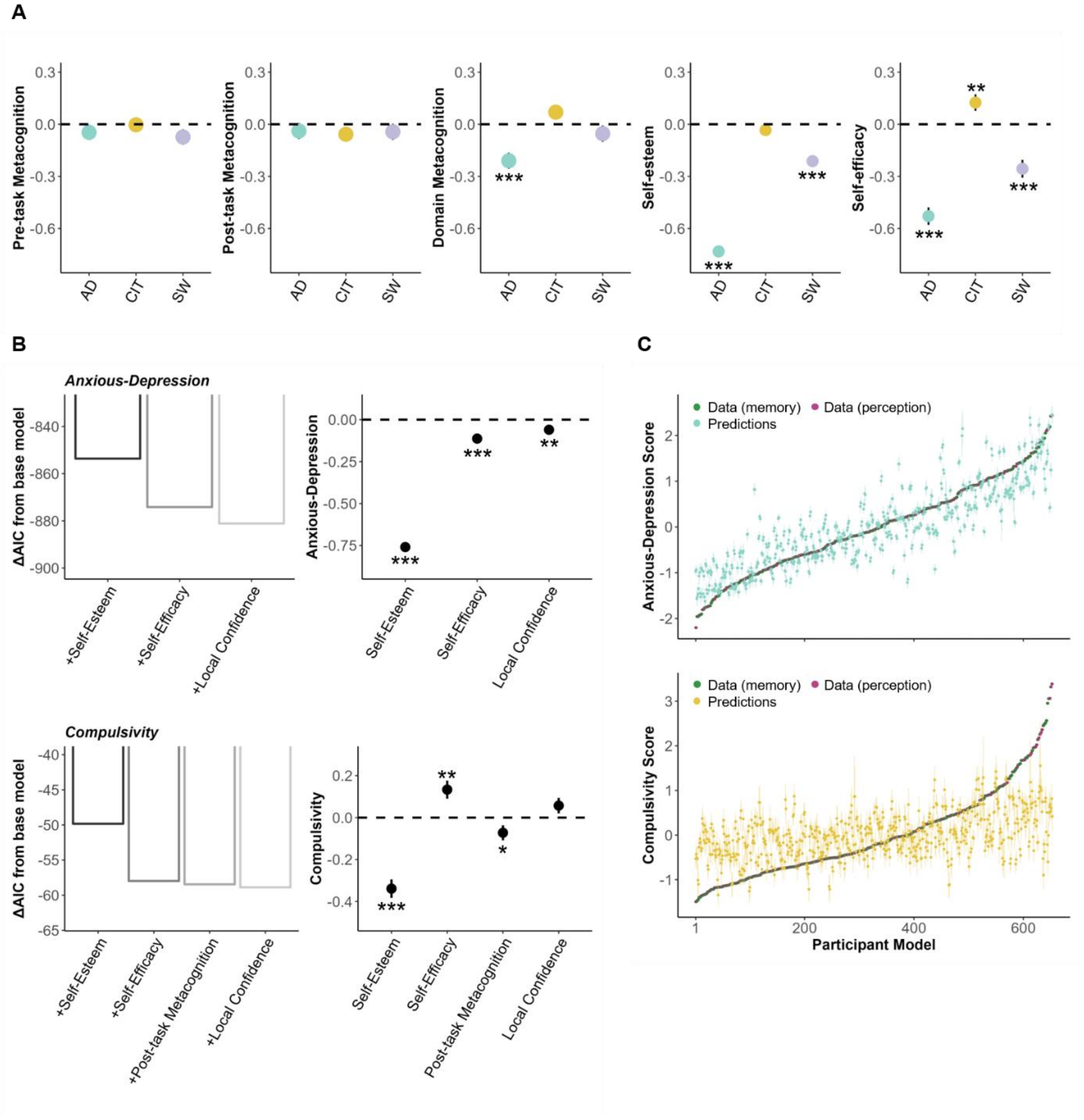


***Fig. 5. Global metacognition and its association to dimension severity. (A) Various measures of global metacognition and its association with dimension scores.** Transdiagnostic dimension scores were included in the same model, with a separate model for each metacognitive measure. The Y-axes shows the z-scored change in metacognition level as a function of 1 standard deviation increase of dimension scores. Error bars denote standard errors. *p < 0.05, **p < 0.01, ***p < 0.001*. **(B) Stepwise regression of metacognitive measures contributing to psychopathology.** Insets on the left depict the Akaike information criterion (AIC) scores from the models. Base model includes IQ, age and gender, and every additive iteration includes the regressor following rightward of the x-axes. The lower the AIC, the better the model fit. The right insets depict coefficients of the metacognitive regressors from the final model of the stepwise regression. The Y-axes shows the change in dimension score as a function of 1 standard deviation increase of z-scored metacognitive regressor. Error bars denote standard errors. *p < 0.05, **p < 0.01, ***p < 0.001. **(C) Out-of-sample model predictions from cross-validation of the winning model predicting dimension severity.** Models are the same as the ones revealed by step-wise*

regression analyses. X-axis represents participant models, where there are two data confidence models (purple for perception and green for memory) per participant, ranked by dimension score. Error bars indicate 95% confidence interval of predicted score. See Supplementary Fig. 11 for comparison of adjusted R^2 and root mean square error (RMSE) of all the tested models.

Discussion

In this study, we sought to reconcile diverging findings in how metacognition relates to psychiatric symptoms, particularly when comparing more traditional clinical approaches to metamemory[11–16] with recent transdiagnostic methods applied to metaperception[18–21,23]. We considered three possible reasons for this divergence: (i) a metacognitive domain-specificity leading to distinct results with memory and perception tasks, (ii) the co-occurrence of mental health symptoms overshadowing effects, and/or (iii) differing manifestations of psychopathology at different hierarchical levels of metacognition. We were able to refute the first explanandum by showing that lower confidence is linked to anxious-depression, and (partially) higher confidence is linked to compulsivity, across both perception and memory tasks. Secondly, using a dimensional approach, we replicated the opposing effects of anxious-depression and compulsivity, lending support for the overshadowing hypothesis. Lastly, we found that mental health dimensions were differentially linked to distinct levels of a metacognitive hierarchy, refining accounts of local versus global associations with mental health.

Whether metacognition is domain-general[28,29,33,39–44] or specific[24–29] is a subject of continued debate. We take the view that the extent of domain-generality is likely to differ between different metacognitive measures. For instance, we observed that metacognitive bias (average confidence level) was robustly correlated, whereas metacognitive efficiency (the mapping between performance and confidence) was more weakly associated (Supplementary Fig. 6 & 7), between perception and memory tasks. It is probable that metacognitive efficiency is influenced by other factors, such as interindividual differences in the utilisation of domain-specific cues (e.g., familiarity of images versus spatial information of dots) that determine the evaluative accuracy of confidence reports[45,46], and are thus more task-specific. Notably, our findings go beyond previous work on the domain-generality of metacognition by showing that mappings between confidence and symptomology are also task general: we saw that metacognitive bias was dissociatively linked to anxious-depression (underconfidence) and compulsivity (overconfidence) in both perception and memory. Even though this is indirect evidence, it does support the notion that the higher co-morbid anxiety/depression symptoms commonly found in OCD patients might explain the reduction in memory confidence seen in clinical OCD studies[47] but which is accounted for when studying compulsive individuals using a dimensional approach[22,31]. However, recent studies with OCD patients have found lowered confidence in comparison to healthy controls despite controlling for anxiety and depression scores[48,49]. It is thus also possible that there are specific features in patients that make them categorically distinct from “non-patients” or that other forms of co-morbid psychopathology could explain this lowered confidence. The true test will require a comparison with a transdiagnostically-characterised clinical cohort[50]. In line with this, we highlight that what we term 'compulsivity' does not only measure obsessive-compulsive symptoms, but also encompasses other compulsive-type symptoms such as addiction and eating disorders.

We note that the association between high confidence and compulsivity was not particularly strong in our data—the effect became non-significant when examined in the perceptual domain only (Supplementary Fig. 8). We considered several explanations for this through control analyses, and determined that it was not due to differences between the tasks in (i) tutorial accuracy titration, (ii) pre-task metacognitive evaluations, (iii) task order completion, (iv) task accuracy titration, or due to (v) using transformed scores versus de novo factor analysis scores (data not shown). Because a relationship between high perceptual confidence and compulsivity has been replicated numerous times[18,19,21,23], we suggest that our null finding might be driven by a slightly smaller effect size (β~=0.2) than we expected in our power calculation (which led to our somewhat smaller sample size). Future studies might note the importance of ensuring higher power for investigating relationships between metacognition and psychopathology.

Recent studies have started to unravel distinct contributions of different levels of a metacognitive hierarchy to mental health[19] and ageing[33]. Beyond local task mean confidence, we also comprehensively quantified five additional metacognitive metrics tracking more global levels of metacognition—pre-task metacognition, post-task metacognition, cognitive self-ability metacognition, self-esteem and self-efficacy. We first observed that all metacognitive metrics we assessed were positively associated with each other (Supplementary Fig. 10), but the associations between local and global levels were not particularly high (e.g., local confidence with any other metric, r=0.13-0.33). Thus, it is unsurprising that these different levels of a metacognitive hierarchy showed different relationships with anxious-depression and compulsivity. When we probed how much each metacognitive measure contributes to dimension severity, we found that anxious-depression and compulsivity were linked to different yet overlapping profiles of hierarchical metacognition across step-wise and cross-validation approaches.

For anxious-depression, a majority of the variance was explained by global metacognition in decreased self-esteem levels, followed by lower self-efficacy and lower local confidence. The predominant contribution of low global metacognition echoes clinical models of depression that emphasise negative schemas about general topics (oneself, the world, the future)[51,52]. For compulsivity, there was a contribution of lower self-esteem and lower post-task metacognition, but also higher self-efficacy and higher local confidence effects. This (replicated[19]) bi-valenced effect of different levels of a metacognitive hierarchy in compulsivity might encapsulate the seemingly contrasting hypotheses of OCD as a disorder of doubt[9,10,53] versus the rigidity of confidence and faulty world models underlying compulsive behaviour[22,54]. For the former, low self-esteem/post-task metacognition might reflect distrust of one's judgements/actions that promote repetitive behaviours like checking or assurance seeking[55]. For the latter, high local confidence/self-efficacy might reflect the greater endorsement of individuals' repetitive actions as being helpful despite evidence to the contrary[56,57]. These two phenomena are likely to act in tandem to aggravate persistent compulsive behaviours.

In summary, we found that metacognition exhibits similar patterns across perception and memory, but shows distinct, differing hierarchical contributions to psychopathology.

Methods

Participants

N=400 participants were recruited online via the worker platform Prolific (https://www.prolific.co/). All participants were 18-55 years old, currently residing in the United Kingdom, fluent in English, have normal or corrected-to-normal vision, and have at least 90% approval rate on Prolific. They provided informed consent online and received £8.26/hour plus a bonus of up to £4 based on their performance. We obtained and performed in accordance to the ethical approval for the study procedures by UCL's Research Ethics Committee (15301/001).

Power analysis. A power analysis was conducted using the effect size from a previous study examining associations between perceptual local confidence alterations and transdiagnostic dimensions of anxious-depression, compulsivity and intrusive thought ('compulsivity') and social withdrawal[23]. The prior study reported a positive association of anxious-depression ($\beta=-0.20$, $p<0.001$), and negative association of compulsivity ($\beta=0.23$, $p<0.001$), with mean confidence level. We used the smaller effect size (anxious-depression effect), to calculate sample size, which suggested that N=289 participants were required to achieve 90% power at 0.001 significance level with 210 task trials. Given that the current tasks consisted of 150 trials each, we required N=327 for the same power level and significance. We estimated N=400 as the target number of participants to reach the necessary sample size after accounting for expected data exclusions.

Exclusion criteria. To ensure data quality, multiple attention and comprehension checks were integrated into the study. For the experimental tasks, participants were required to get 100% correct on a short comprehension quiz on the task instructions before they were allowed to start either of the main tasks. Should they fail, they were directed back to the instructions before another quiz attempt. If they failed on their 3rd attempt, they were brought back to the very beginning to complete the practice trials again before another quiz attempt. For the questionnaires, the battery set included two attention check question items requiring participants to select a specific answer.

Following the exclusion criteria outlined in our preregistration, participants were excluded from the analysis if they:

- Had incomplete datasets, arising from mishaps in remote data collection (e.g., portions not saved) (N=18)
- Had >50 attempts to identify all the stimuli in the stimuli recognition training portion for the metamemory task (See Metamemory Task) (N=0)

- Had >5 attempts of the comprehension quiz (either task) (N=0)
- Failed at least one attention check question in the questionnaire battery (N=10)
- Had r>0.5 correlation of initial default confidence rating with final confidence rating across trials in tasks (N=1 (perception))
- Had task accuracy that diverged from expected of staircase (<60% or >85%) (N=6 (memory), N=1 (perception))
- Had same trial-by-trial confidence rating >80% of the trials (N=3 (memory), N=3 (perception))

We also excluded participants who:

- Chose the left/right option >80% the time (N=3 (perception))
- Recorded stimulus presentation time deviating >50ms from expected (N=33 (perception))
- Reported a task stimulus presentation error (N=1 (memory))
- Mean confidence reaction time >20s (N=1 (memory))

In total, N=73 (18.25%) were excluded, leaving N=327 participants (136 female (41.59%), 1 other gender (1.22%)) for analysis.

We additionally excluded trials with implausibly slow reaction times of >10s and/or ± 3 standard deviations from the per-subject mean for each participant. 1.59% (metamemory task) and 1.82% (metaperception task) of all trials were excluded.

Procedure

We employed a randomised, cross-over within-subject study design (Fig. 1A). We first assessed cognitive self-ability metacognition before each participant completed two experimental task sets assessing memory or perception. Each of these sets included a tutorial with practice trials, pre-task global metacognitive ratings, the main experimental task with trial-by-trial confidence ratings, followed by post-task global metacognitive ratings. After the task sets, we asked participants about their insight in how their confidence evaluations were related between the experimental tasks. Finally, self-report questionnaires measuring individual differences in demographics, psychopathology, self-beliefs and intellectual abilities (IQ) were administered.

The entire study was programmed in React v.18.2.0 (https://react.dev/) (JavaScript library), developed in an app bootstrapped by Create React App (https://github.com/facebook/create-react-app) and hosted on Scalingo (https://scalingo.com/).

Self-ability metacognition/estimates. Directly after the consent pages and prior to the experimental task sets, we asked participants to rate their abilities of memory and perception on a 1-10 scale, from 'worse than everyone' to 'better than everyone':

- How you do generally rate your memory ability? For example, in remembering events that you experienced a long time ago.
- How you do generally rate your perception ability? For example, how good you are at spotting hidden things, like birds in a forest.

Experimental tasks. Participants then began with one of the experimental task sets by random assignment in the order of which was performed first (N=159 (48.62%) memory first, N=168 (51.38%) perception first). We employed two-alternative forced choice memory and perceptual decision-making tasks with trial-by-trial confidence ratings (Fig. 1B & 1C). Both tasks were presented with a cover story. Participants were told that they boarded our spaceship to help out after an asteroid hit and damaged the spaceship. In the metamemory task, participants were to memorise sets of animal images and then were presented with two words (that each described an animal). They had to remember and select which animal was seen previously to help allocate them into their pods on the spaceship. In the metaperception task, participants were to distinguish high charged batteries (box with more dots) to help power up the spaceship. In both tasks, their confidence in their choice (higher charged battery/animal that was seen) was rated after every trial on a continuous 50-100% probability correct scale. Each task set consisted of a tutorial with 25 trials, plus the main task of 150 trials split over 3 blocks. Performance in both tasks were controlled by a two-down one-up staircase procedure (targeting ~71%) that was initiated during the 25 practice trials to minimize burn-in period (Supplementary Fig. 1 & 2).

Metamemory task. On each trial, participants were presented with a fixation cross (1000ms), followed by a presentation of a set of animal stimuli (1000ms) before the screen was cleared (500ms) (Fig. 1B). Participants were then shown two words, one which described one of the stimuli in the presented set and one that did not. They were to select the word representing the stimuli that was shown previously. The chosen word was then outlined in blue (700ms) before they rated their confidence in the decision. The staircase step-size was ±1 stimulus for the entire task.

Stimuli recognition task. As part of the tutorial, participants were shown each of the task stimuli one at a time, and asked to select the word that describes the stimulus versus another word that did not. The correct word for all stimuli had to be chosen to move on to the main task instructions and practice trials. This was to prevent misinterpretation of what the stimuli were depicting.

Harm versus non-harm stimuli. We selected an array of 15 animal pictures as the stimuli set (Supplementary Fig. 9). They were in greyscale and matched in luminesce. Items in 8 of these were perceived as highly harmful while the other 7 were rated to have low harm levels in an affective rating study [36]. We indexed these stimuli into two groups (Harm, Non-Harm) for subsequent analysis.

Metaperception task. On each trial, participants were presented with a fixation cross (1000ms) followed by a quick presentation of two boxes filled with differing number of dots (300ms) (Fig. 1C). One box would always have half the number of dots that could fill the box (313 dots out of 625 positions), whilst the other would have more than half (an increment of 1-312 dots). Empty boxes were then left on the screen, where participants were to select which box had the higher number of dots. The selected box was then outlined in blue (700ms) before confidence ratings were given. Staircase step-size was calculated in log-space as implemented in prior studies [23], with a starting point of 4.2 (+70 dots), changing by ±0.4 for the first 5 trials, ±0.2 for the next 5 trials and ±0.1 for the rest of the task. Dots were drawn on a trial-by-trial basis with react-konva v18.2.3 (https://konvajs.org/docs/react/index.html).

Global task metacognition. Pre- and post- task self-performance estimates were assessed before and after each of main tasks respectively, by asking how many trials participants think they would get/have gotten correct with a 0-150 correct trials scale.

Metamemory:

- Pre-task: Before we begin, out of 150 sets of animals, how many times you do think you will be able to select the correct animal seen in the set?
- Post-task: After going through all the 150 sets of animals, how many times do you think you selected the animal seen in the sets correctly?

Metaperception:

- Pre-task: Before we begin, out of 150 set pairs of battery cards, how many times do you think you will choose the higher charge battery card correctly?

- Post-task: After going through all the 150 set pairs of battery cards, how many times do you think you selected all the higher charge battery cards correctly?

Insight ratings. After completion of both task sets, we asked three insight questions:

- Did you prefer to complete the first task over the second task? (1-5 scale; 1: definitely not, 3: no difference, 5: definitely yes)
- How much did you feel that your confidence changed from completing the first task to finishing the second task? (1-7 scale; 1: decreased a lot, 4: no change, 7: increased a lot)
- How much did you feel that your confidence in the first task influenced your confidence on the second task? (1-5 scale; 1: not at all, 3: somewhat, 5: a lot)

Questionnaires. In the final section, participants completed a battery of 11 self-report questionnaires measuring psychopathology and self-beliefs, the order of which was fully randomised.

We assessed 9 mental health questionnaires to subsequently define transdiagnostic dimensions [34], which included:

- Alcohol addiction, using the Alcohol Use Disorder Identification Test (AUDIT)[58]
- Apathy, using the Apathy Evaluation Scale (AES)[59]
- Depression, using the Self-Rating Depression Scale (SDS)[60]
- Eating disorders, using the Eating Attitudes Test (EAT-26)[61]
- Impulsivity, using the Barratt Impulsivity Scale (BIS-11)[62]
- Obsessive-compulsive disorder, using the Obsessive-Compulsive Inventory - Revised (OCI-R)[63]
- Schizotypy, using the Short Scales for Measuring Schizotypy (SSMS)[64]
- Social anxiety, using the Liebowitz Social Anxiety Scale (LSAS)[65]
- Trait Anxiety, using the trait portion of the State-Trait Anxiety Inventory (STAI)[66]

We also collected 2 self-belief metrics as a form of global metacognition measures:

- Self-esteem, using the Rosenberg Self-esteem Scale (RSE)[37]
- Self-efficacy, using the General Self-efficacy Scale (GSE)[38]

Finally, we administered a short IQ evaluation using the International Cognitive Ability Resource (I-CAR) sample test[67].

Transdiagnostic dimensions. To quantify mental health dimensions in our sample, we applied a previously defined transdiagnostic definition based on a weighted combination of items drawn from the 9 mental health questionnaires[34]. We used weights derived from the previous study with N=1,413 to transform our scores as our sample size had a lower a subject-to-variable ratio (N=327) for de novo factor analysis. Consistent with prior literature, the resulting transformed dimension scores of anxious-depression, compulsivity and social withdrawal were moderately intercorrelated (r=0.30-0.53) (Supplementary Fig. 4). To note, in the prior study, item 13 on the SDS was mistakenly phrased "I am restless and can't sleep" rather than "I am restless and can't keep still", but subsequent work have demonstrated the stability of the factor structure in new data, with and without this error[22,23]. Likewise, applying de novo factor analysis to our current sample replicated the same factor structure (Supplementary Fig. 5).

See Supplementary Fig. 3 & 4 for the distribution of questionnaire and dimensional scores.

Analyses

All analyses were conducted in R version 3.6.0 via RStudio version 1.2.1335 (http://cran.us.r-project.org) and MATLAB 2019b (https://www.mathworks.com/). In R, we utilised the t.test() function (stats package) for two-sided paired t-tests, lm() function (lme4 package) for general linear modelling, lmer() function (lmerTest package) to estimate mixed-effects models, cor.test() function (stats package) for Pearson's and Spearman's correlation tests, and the step() function (stats package) for stepwise regressions. Repeated cross validation analyses on model regressions were performed in MATLAB using fitglm(), predict() and immse() functions. We outline further detail of the main analyses below:

Local metacognition estimates. We estimated local confidence (metacognitive bias) measure by the mean of the trial-by-trial confidence ratings over each task. Split-half reliability was high (r=0.99) for both tasks. We also ran both individual and hierarchical Bayesian meta-d' analyses (HMeta-d)[35,68] to obtain individual estimates of d' (sensitivity), meta-d' (metacognitive sensitivity) and m-ratio (meta-d'/d'; metacognitive efficiency), as well as group level estimates of m-ratio. Results of meta-d' model parameter analyses are reported in Supplementary Fig. 6 and 7.

General linear modelling. We asked if mean local confidence (Confidence) was associated with psychiatric questionnaire (Questionnaire) or dimension (anxious-depression (AD), compulsivity (CIT), social withdrawal (SW)) severity with a general linear model where the Task Order (memory:1, perception:-1), age, gender, IQ, and their interaction with Task Domain (memory:1, perception:-1) were taken as co-variates. Confidence, questionnaire/dimension scores, age and IQ were z-scored. The models were:

Confidence ~ (Questionnaire + Task Order + Age + Gender + IQ)*Task Domain
Confidence ~ (AD + CIT + SW + Task Order + Age + Gender + IQ)*Task Domain

The main effect of Questionnaire or AD/CIT/SW (Dimension) on Confidence was interpreted as a general confidence effect with psychopathology. The interaction effect of Dimension*Task Domain was interpreted as the change of dimension score on confidence effect compared between memory versus perception.

Mixed-effects modelling. To examine trial accuracy and confidence rating effects by stimuli category (Word Category; harm versus non-harm), we utilised a mixed-effects logistic regression model including number of stimuli presented (Task Difficulty), age, IQ and gender as main effect co-variates, with Word Category and Task Difficulty as random effect predictors. Task Difficulty, age and IQ were z-scored. The models were:

Trial Accuracy ~ Word Category + Task Difficulty + Age + IQ + Gender + (1 + Word Category + Task Difficulty | Subject)

Trial Confidence ~ Word Category + Trial Accuracy + Task Difficulty + Age + IQ + Gender + (1 + Word Category + Task Difficulty | Subject)

To examine if the effect of stimuli category on accuracy or confidence differed as a function of dimension scores, we interacted Word Category in the above model with the dimension scores. The models were:

Trial Accuracy ~ Word Category*(AD + CIT + SW) + Task Difficulty + Age + IQ + Gender + (1 + Word Category + Task Difficulty | Subject)

Trial Confidence ~ Word Category*(AD + CIT + SW) + Trial Accuracy + Task Difficulty + Age + IQ + Gender + (1 + Word Category + Task Difficulty | Subject)

Stepwise regression. We explored the importance of each metacognitive metric in contributing to dimension severity. With a base model of age, IQ and gender, we set the full model for forward iteration as:

Dimension ~ (Local Confidence + Pre-task Metacognition + Post-task Metacognition + Self-ability Metacognition)*Task Domain + Self-Esteem + Self-Efficacy + Age + IQ + Gender

All metacognitive regressors, age and IQ were z-scored.

Repeated cross validation. To validate the models revealed by stepwise regression, we conducted 5-fold cross validation on 16 variations of these regression models (Supplementary Fig. 11). We used cvpartition() with K-fold=5, running fitglm() on the training set and predicting dimension scores with predict() on the test set. immse() was used to obtain root mean squared error (RMSE) of the model fits.

Code and Data Availability

Stimuli for the memory task is publicly accessible (https://osf.io/mey48/)[36]. The data and analysis code to reproduce the main analyses and figures are available at https://osf.io/jm2at/.

Declarations of Interest

TUH consults for limbic ltd and holds shares in the company, which is entirely unrelated to the current project. The other authors declare no conflicts of interest.

Acknowledgements

We thank Claire Gillan for their discussion on the initial analyses of the study.

Authors' Contributions

TXFS: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Software, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing
SMF: Methodology, Writing – Review & Editing
TUH: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

Funding

TXFS is a Sir Henry Wellcome Postdoctoral Fellow (224051/Z/21/Z) based at the Max Planck UCL Centre for Computational Psychiatry and Ageing Research.

SMF is a CIFAR Fellow in the Brain, Mind & Consciousness Program, and is funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666).

TUH is supported by a Sir Henry Dale Fellowship (211155/Z/18/Z; 211155/Z/18/B; 224051/Z/21) from Wellcome/Royal Society, a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, a 2018 NARSAD Young Investigator grant (27023) from the Brain & Behavior Research Foundation and a Philip Leverhulme Prize from the Leverhulme Trust (PLP-2021-040). This research was funded in whole, or in part, by the Wellcome Trust (211155/Z/18/Z).

The Max Planck UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. American Psychiatric Association (APA). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. (American psychiatric association Washington, DC, 2013).

2. David, A. S., Bedford, N., Wiffen, B. & Gilleen, J. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 1379–1390 (2012).

3. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front Hum Neurosci* **8**, 443 (2014).

4. Hoven, M. *et al.* Abnormalities of confidence in psychiatry: an overview and future perspectives. *Transl Psychiatry* **9**, 268 (2019).

5. Hancock, J. A. "Depressive realism" assessed via confidence in decision-making. *Cogn Neuropsychiatry* **1**, 213–220 (1996).

6. Fu, T., Koutstaal, W., Fu, C. H. Y., Poon, L. & Cleare, A. J. Depression, confidence, and decision: Evidence against depressive realism. *J Psychopathol Behav Assess* **27**, 243–252 (2005).

7. Fu, T. S.-T., Koutstaal, W., Poon, L. & Cleare, A. J. Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *J Behav Ther Exp Psychiatry* **43**, 699–704 (2012).

8. Dar, R., Sarna, N., Yardeni, G. & Lazarov, A. Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychol Med* 1–9 (2022).

9. Dar, R. Elucidating the mechanism of uncertainty and doubt in obsessive-compulsive checkers. *J Behav Ther Exp Psychiatry* **35**, 153–163 (2004).

10. Dar, R., Rish, S., Hermesh, H., Taub, M. & Fux, M. Realism of confidence in obsessive-compulsive checkers. *J Abnorm Psychol* **109**, 673 (2000).

11. Macdonald, P. A., Antony, M. M., Macleod, C. M. & Richter, M. A. Memory and confidence in memory judgments among individuals with obsessive compulsive disorder and non-clinical controls. *Behaviour research and therapy* **35**, 497–505 (1997).

12. Tuna, Ş., Tekcan, A. I. & Topçuoğlu, V. Memory and metamemory in obsessive–compulsive disorder. *Behaviour Research and Therapy* **43**, 15–27 (2005).

13. Karadag, F., Oguzhanoglu, N., Ozdel, O., Atesci, F. C. & Amuk, T. 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).

14. Tolin, D. F. *et al.* Memory and memory confidence in obsessive–compulsive disorder. *Behaviour research and therapy* **39**, 913–927 (2001).

15. Moritz, S. & Jaeger, A. Decreased memory confidence in obsessive–compulsive disorder for scenarios high and low on responsibility: is low still too high? *Eur Arch Psychiatry Clin Neurosci* **268**, 291–299 (2018).

16. Zitterl, W. *et al.* Memory deficits in patients with DSM-IV obsessive-compulsive disorder. *Psychopathology* **34**, 113–117 (2001).

17. Seow, T. X. F., Rouault, M., Gillan, C. M. & Fleming, S. M. How local and global metacognition shape mental health. *Biol Psychiatry* **90**, 436–446 (2021).

18. Benwell, C. S. Y., Mohr, G., Wallberg, J., Kouadio, A. & Ince, R. A. A. Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *Npj Mental Health Research* **1**, 10 (2022).

19. Hoven, M., Luigjes, J., Denys, D., Rouault, M. & van Holst, R. J. How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health* 1–9 (2023).

20. Katyal, S., Huys, Q. & Fleming, S. How underconfidence is maintained in anxiety and depression. *Psyarxiv* (2023).

21. Fox, C. A. *et al.* Metacognition in anxious-depression is state-dependent: an observational treatment study. *Elife* **12**, (2023).

22. Seow, T. X. F. & Gillan, C. M. Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Sci Rep* **10**, 2883 (2020).

23. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol Psychiatry* **84**, 443–451 (2018).

24. Baird, B., Smallwood, J., Gorgolewski, K. J. & Margulies, D. S. Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *Journal of Neuroscience* **33**, 16657–16665 (2013).

25. Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T. & Schooler, J. W. Regional white matter variation associated with domain-specific metacognitive accuracy. *J Cogn Neurosci* **27**, 440–452 (2015).

26. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).

27. Morales, J., Lau, H. & Fleming, S. M. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience* **38**, 3534–3546 (2018).

28. Mazancieux, A., Fleming, S. M., Souchay, C. & Moulin, C. J. A. Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *J Exp Psychol Gen* **149**, 1788 (2020).

29. Rouault, M., McWilliams, A., Allen, M. G. & Fleming, S. M. Human metacognition across domains: insights from individual differences and neuroimaging. *Personal Neurosci* **1**, e17 (2018).

30. Hyman, S. E. Can neuroscience be integrated into the DSM-V? *Nat Rev Neurosci* **8**, 725–732 (2007).

31. Gillan, C. M., Fineberg, N. A. & Robbins, T. W. A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychol Med* **47**, 1528–1548 (2017).

32. Rouault, M., Dayan, P. & Fleming, S. M. Forming global estimates of self-performance from local confidence. *Nat Commun* **10**, 1141 (2019).

33. McWilliams, A., Bibby, H., Steinbeis, N., David, A. S. & Fleming, S. M. Age-related decreases in global metacognition are independent of local metacognition and task performance. *Cognition* **235**, 105389 (2023).

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

35. Fleming, S. M. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neurosci Conscious* **2017**, nix007 (2017).

36. Seow, T. X. & Hauser, T. U. What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli. *Wellcome Open Res* **9**, 83 (2024).

37. Rosenberg, M. Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package* **61**, 18 (1965).

38. Jerusalem, M. & Schwarzer, R. Self-efficacy as a resource factor in stress appraisal processes. in *Self-efficacy* 195–214 (Taylor & Francis, 2014).

39. Palmer, E. C., David, A. S. & Fleming, S. M. Effects of age on metacognitive efficiency. *Conscious Cogn* **28**, 151–160 (2014).

40. McCurdy, L. Y. *et al.* Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience* **33**, 1897–1906 (2013).

41. Rouault, M., Lebreton, M. & Pessiglione, M. A shared brain system forming confidence judgment across cognitive domains. *Cerebral Cortex* **33**, 1426–1439 (2023).

42. Bellon, E., Fias, W. & De Smedt, B. Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLoS One* **15**, e0229932 (2020).

43. Lund, A. E., Correa, C., Fardo, F., Fleming, S. & Allen, M. Domain Generality in Metacognitive Ability: A Confirmatory Study Across Visual Perception, Memory, and General Knowledge. *OSF* (2023).

44. Hoogervorst, K., Banellis, L. & Allen, M. Domain-specific updating of metacognitive self-beliefs. *OSF* (2023).

45. Hu, X., Yang, C. & Luo, L. Are the contributions of processing experience and prior beliefs to confidence ratings domain-general or domain-specific? *J Exp Psychol Gen* **152**, 28 (2023).

46. Hu, X. *et al.* A Bayesian inference model for metamemory. *Psychol Rev* **128**, 824 (2021).

47. Moritz, S., Jacobsen, D., Willenborg, B., Jelinek, L. & Fricke, S. A check on the memory deficit hypothesis of obsessive–compulsive checking. *Eur Arch Psychiatry Clin Neurosci* **256**, 82–86 (2006).

48. Hoven, M., Rouault, M., van Holst, R. & Luigjes, J. Differences in metacognitive functioning between obsessive–compulsive disorder patients and highly compulsive individuals from the general population. *Psychol Med* 1–10 (2022).

49. Hoven, M., Mulder, T., Denys, D., van Holst, R. & Luigjes, J. OCD patients show lower confidence and higher error sensitivity while learning under volatility compared to healthy and highly compulsive samples from the general population. *Psyarxiv* (2023).

50. Gillan, C. M. *et al.* Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry* **77**, 77–85 (2019).

51. Beck, A. T. Cognitive models of depression. *Clinical advances in cognitive psychotherapy: Theory and application* **14**, 29–61 (2002).

52. Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R. & Dolan, R. J. Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol Med* **44**, 579–592 (2014).

53. Radomsky, A. S. & Alcolado, G. M. Don’t even think about checking: Mental checking causes memory distrust. *J Behav Ther Exp Psychiatry* **41**, 345–351 (2010).

54. Seow, T. X. F. *et al.* Model-based planning deficits in compulsivity are linked to faulty neural representations of task structure. *Journal of Neuroscience* **41**, 6539–6550 (2021).

55. Parrish, C. L. & Radomsky, A. S. Why do people seek reassurance and check repeatedly? An investigation of factors involved in compulsive behavior in OCD and depression. *J Anxiety Disord* **24**, 211–222 (2010).

56. Doron, G., Kyrios, M. & Moulding, R. Sensitive domains of self-concept in obsessive–compulsive disorder (OCD): Further evidence for a multidimensional model of OCD. *J Anxiety Disord* **21**, 433–444 (2007).

57. Doron, G., Szepsenwol, O., Elad-Strenger, J., Hargil, E. & Bogoslavsky, B. Entity perceptions of morality and character are associated with obsessive compulsive phenomena. *J Soc Clin Psychol* **32**, 733–752 (2013).

58. Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R. & Grant, M. 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).

59. Marin, R. S., Biedrzycki, R. C. & Firinciogullari, S. Reliability and validity of the apathy evaluation scale. *Psychiatry Res* **38**, 143–162 (1991).

60. Zung, W. W. A self rating depression scale. *Arch Gen Psychiatry* **12**, 63–70 (1965).

61. Garner, D. M., Bohr, Y. & Garfinkel, P. E. The eating attitudes test: Psychometric features and clinical correlates. *Psychol Med* **12**, 871–878 (1982).

62. Patton, J. H., Stanford, M. S. & Barratt, E. S. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol* **51**, 768–774 (1995).

63. Foa, E. B. *et al.* The obsessive-compulsive inventory: Development and validation of a short version. *Psychol Assess* **14**, 485–496 (2002).

64. Mason, O., Linney, Y. & Claridge, G. Short scales for measuring schizotypy. *Schizophr Res* **78**, 293–296 (2005).

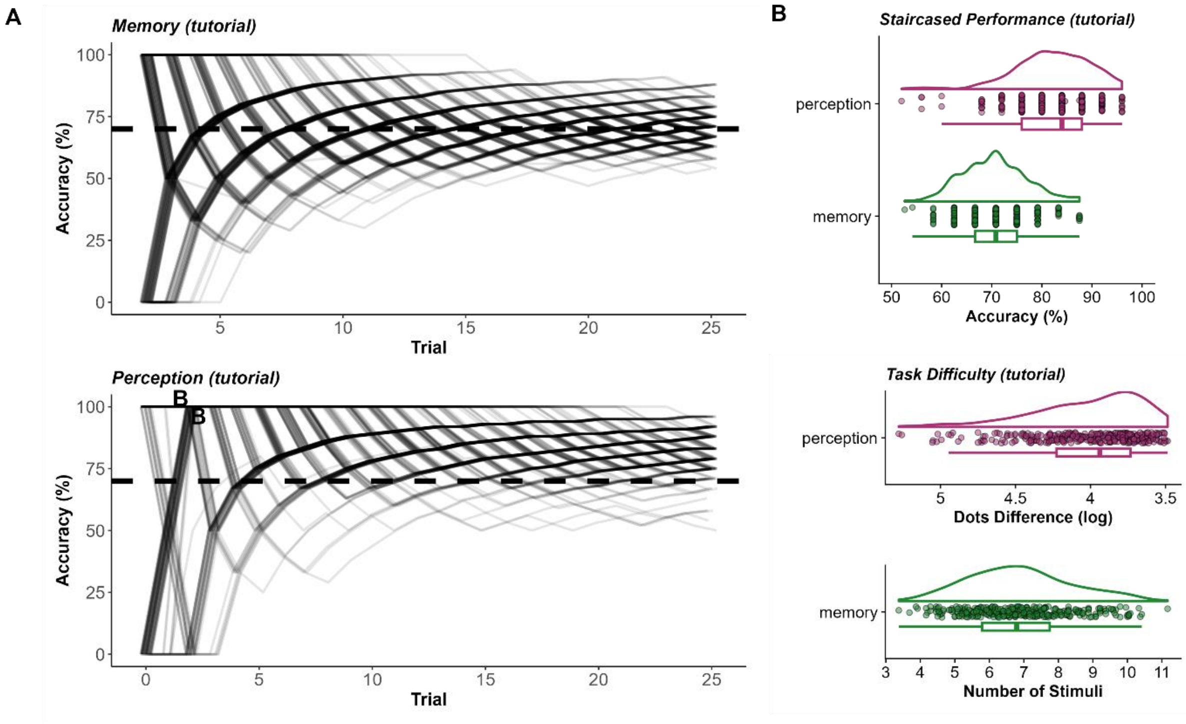
65. Liebowitz, M. R. Social phobia. *Modern Problems of Pharmapsychiatry* **22**, 141–173 (1987).

66. Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R. & Jacobs, G. A. *Manual for the State-Trait Anxiety Inventory.* (Consulting Psychologists Press, Palo Alto, CA, 1983).

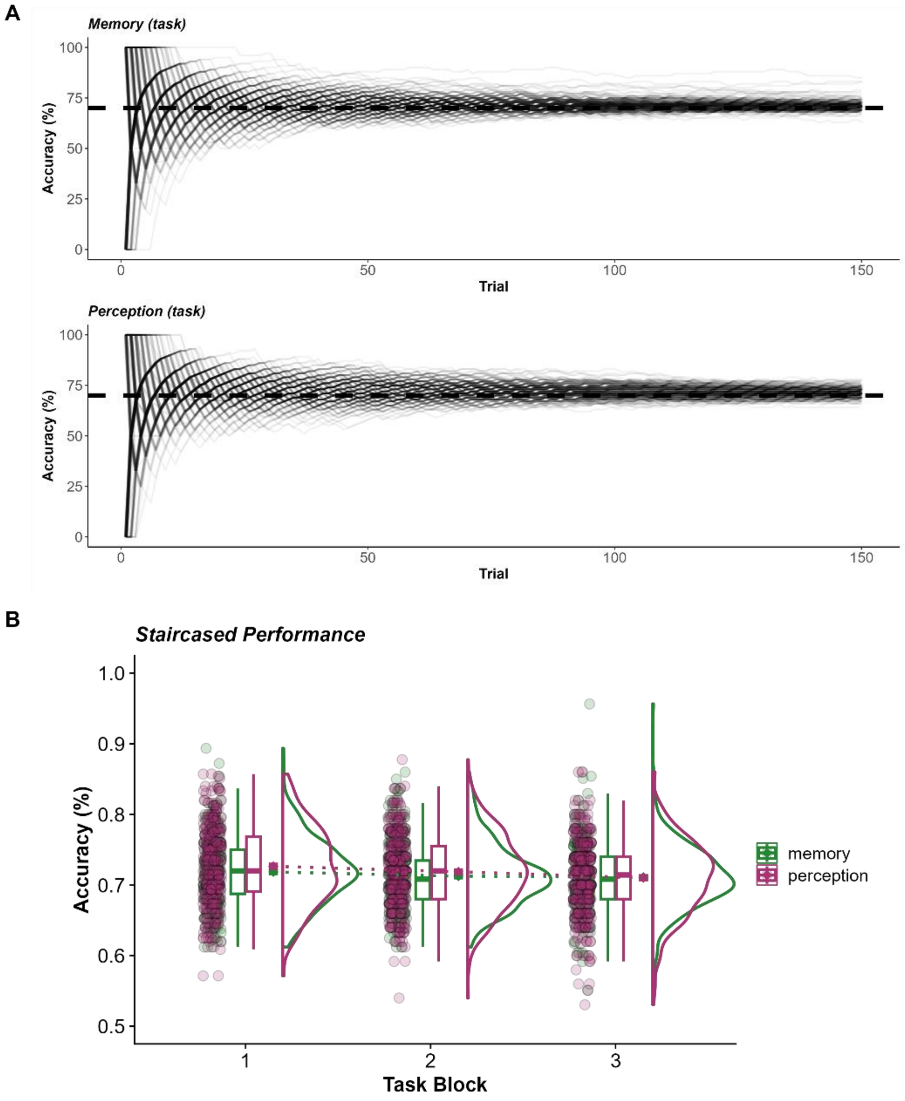
67. Condon, D. M. & Revelle, W. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* **43**, 52–64 (2014).

68. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn* (2012) doi:10.1016/j.concog.2011.09.021.

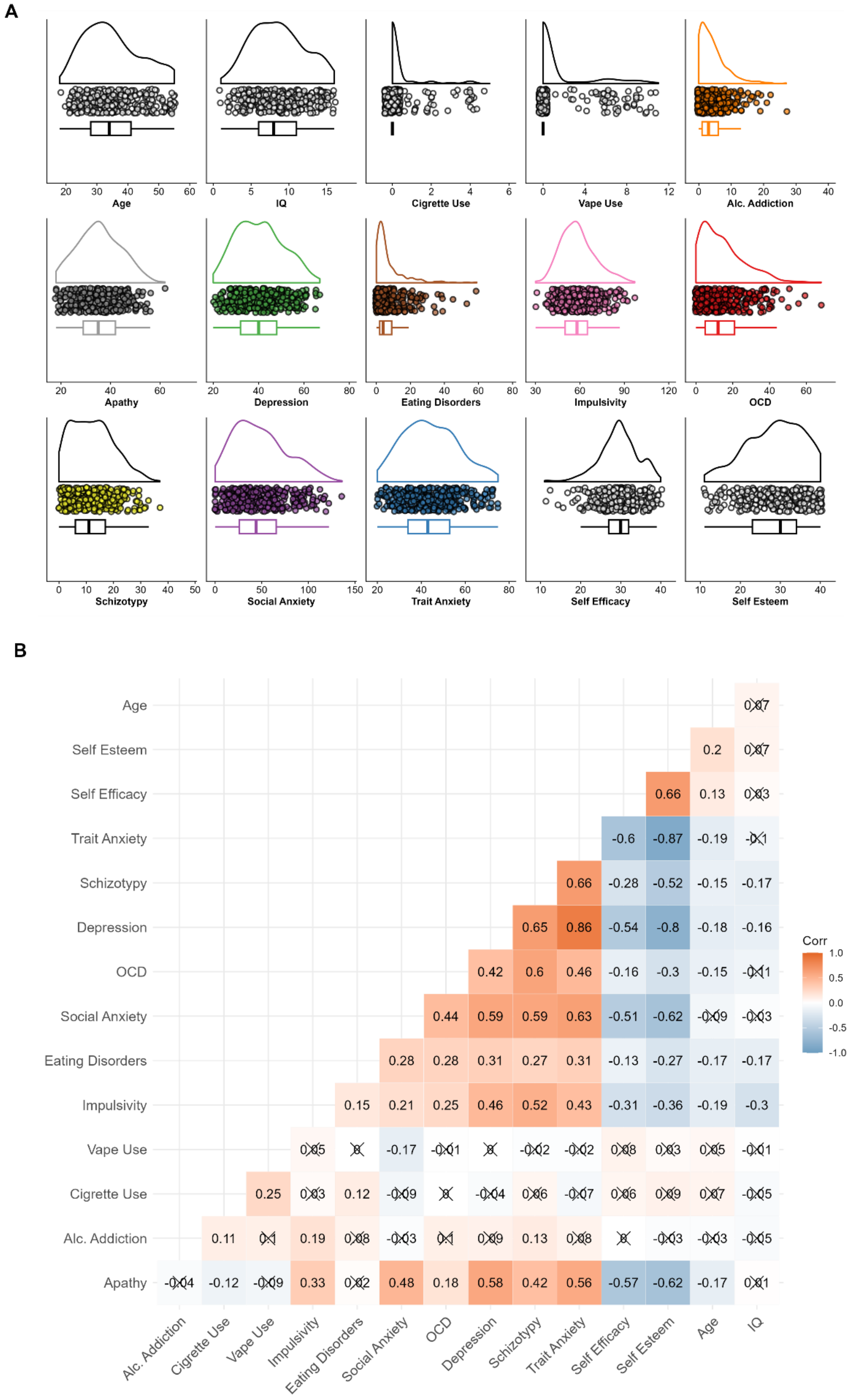
Supplemental Information



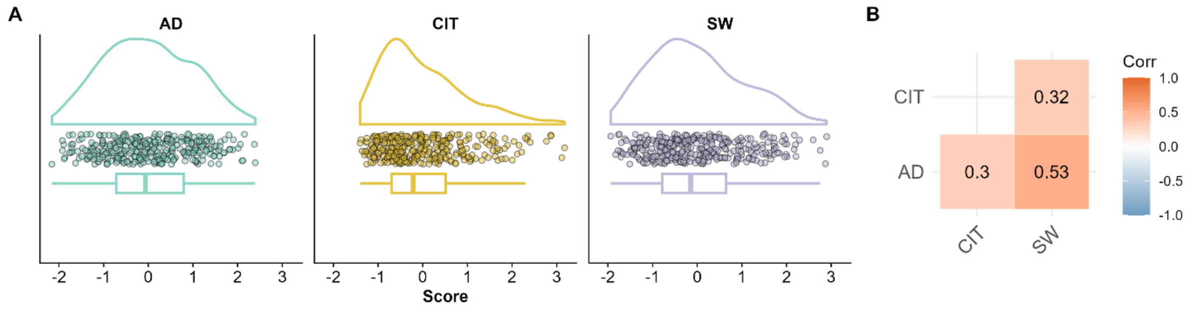
Supplementary Fig. 1. Tutorial (practice) performance titration via staircase procedure. We initiated the staircase procedure from the beginning of the practice trials. **(A) Mean accuracy over practice trials.** Accuracy was calculated as a sliding window including each trial over the accuracy of prior past trials. Each line represents a participant. At the end of the practice trials, the average performance was titrated close to the expected level in the memory task (M=70.59, SD=6.46), but not as much for the perception task (M=82.28, SD=7.50). Dotted lines represent 71% accuracy, where the staircase aimed to titrated performance to. **(B) Distributions of mean accuracy and task difficulty of the practice.** Participants on average finished the perceptual practice trials at a higher accuracy rate than the memory trials (t(325)=23.23, p<0.001, 95% CI=[10.70 12.68]). Task difficulty between tasks were correlated (r=-0.31, p<0.001, 95% CI=[-0.40 -0.21]).



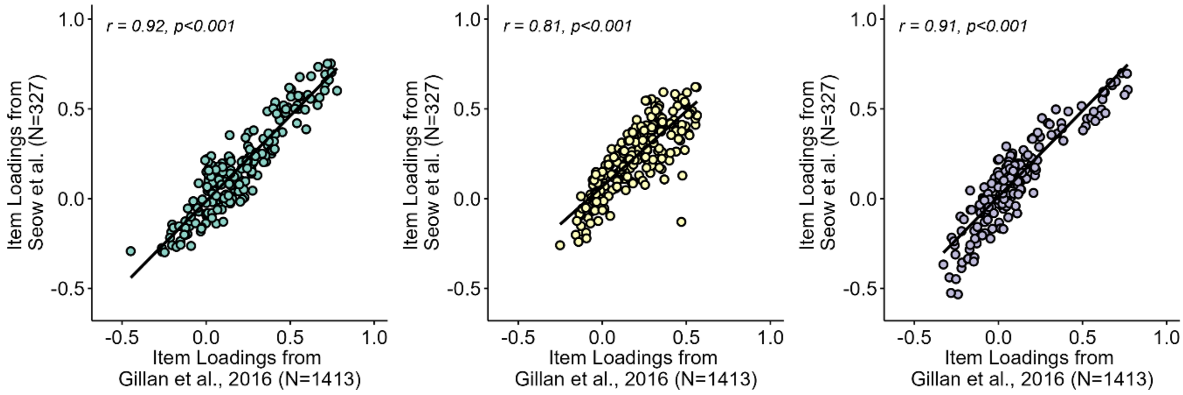
Supplementary Fig. 2. Task performance titration via staircase procedure. (A) Mean accuracy over task trials. Accuracy was calculated as a sliding window including each trial over the accuracy of prior past trials. Each line represents a participant. Performance in both tasks titrated to the expected level. Dotted lines represent 71% accuracy, where the staircase aimed to titrated performance to. **(B) Mean accuracy by task block.** We also examined accuracy within each of the three task blocks (50 trials each) for both tasks. Each circle indicates the accuracy of each participant per block.



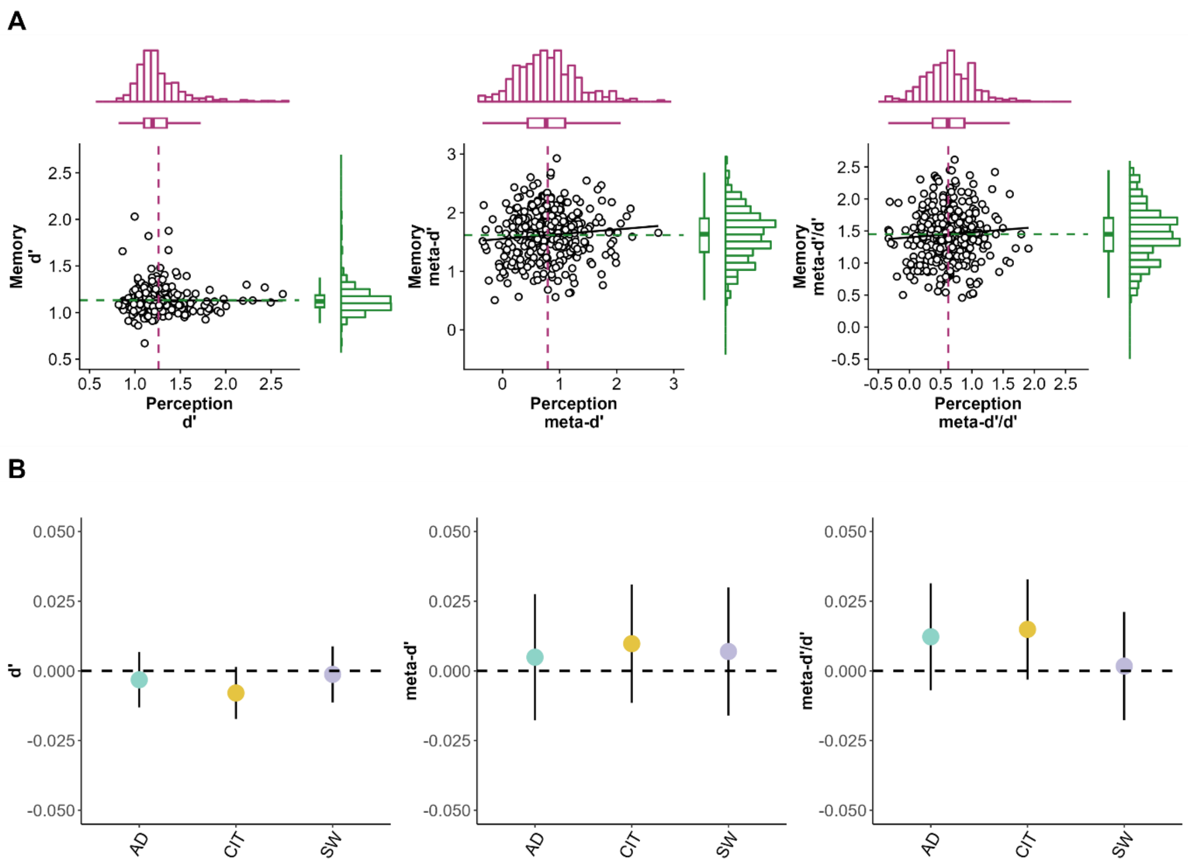
Supplementary Fig. 3. Age and questionnaire battery distributions and correlations. (A) Distributions of age and questionnaire scores. Each circle indicates an individual participant. **(B) Correlations between questionnaire scores.** Boxes that are crossed out indicate non-significant (p>0.05) correlations.



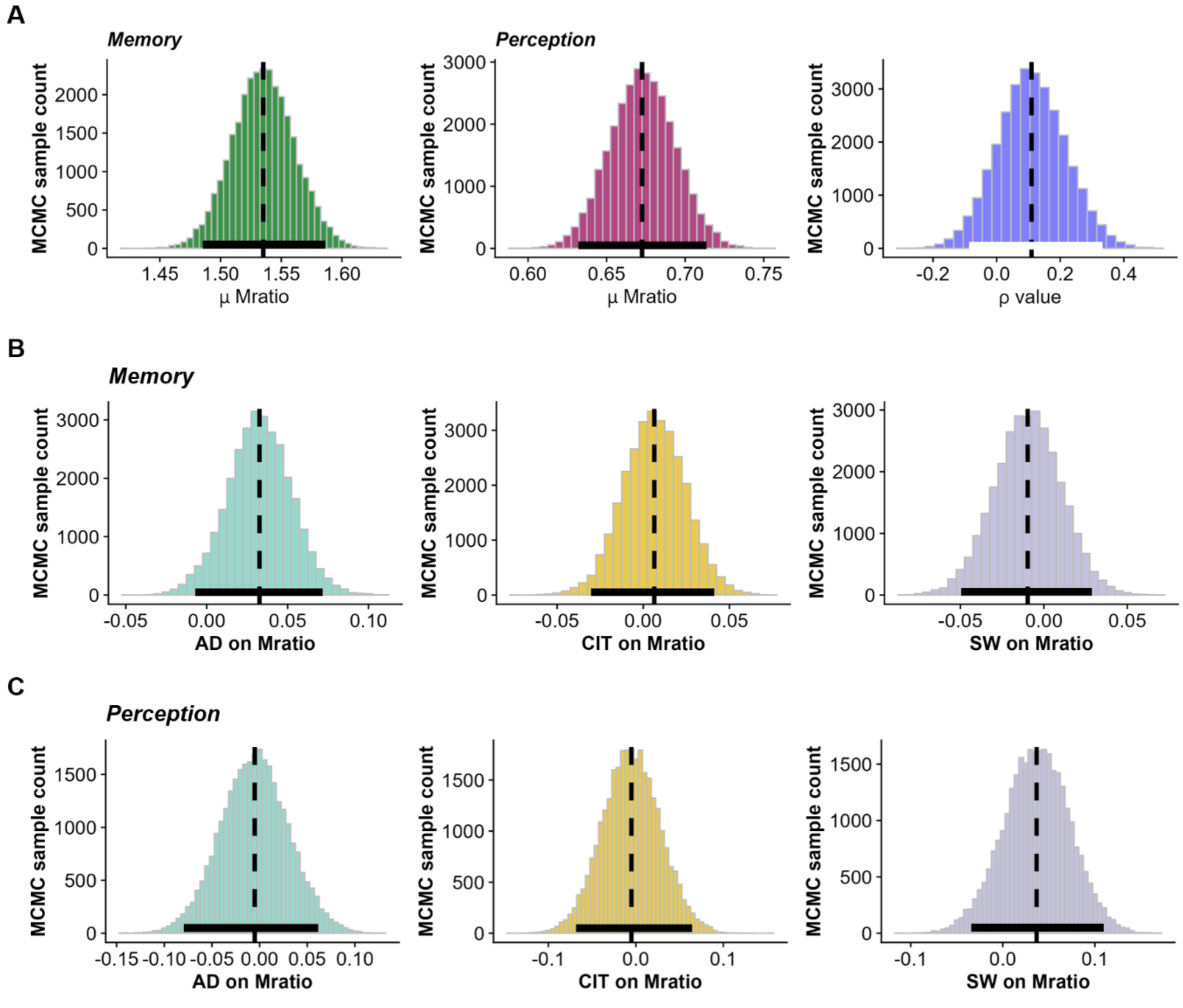
Supplementary Fig. 4. Transdiagnostic dimension score distributions and correlations. (A) Distributions of dimension scores. Each circle indicates an individual participant. **(B) Correlations between dimension scores.** All correlations are significant (p<0.001).



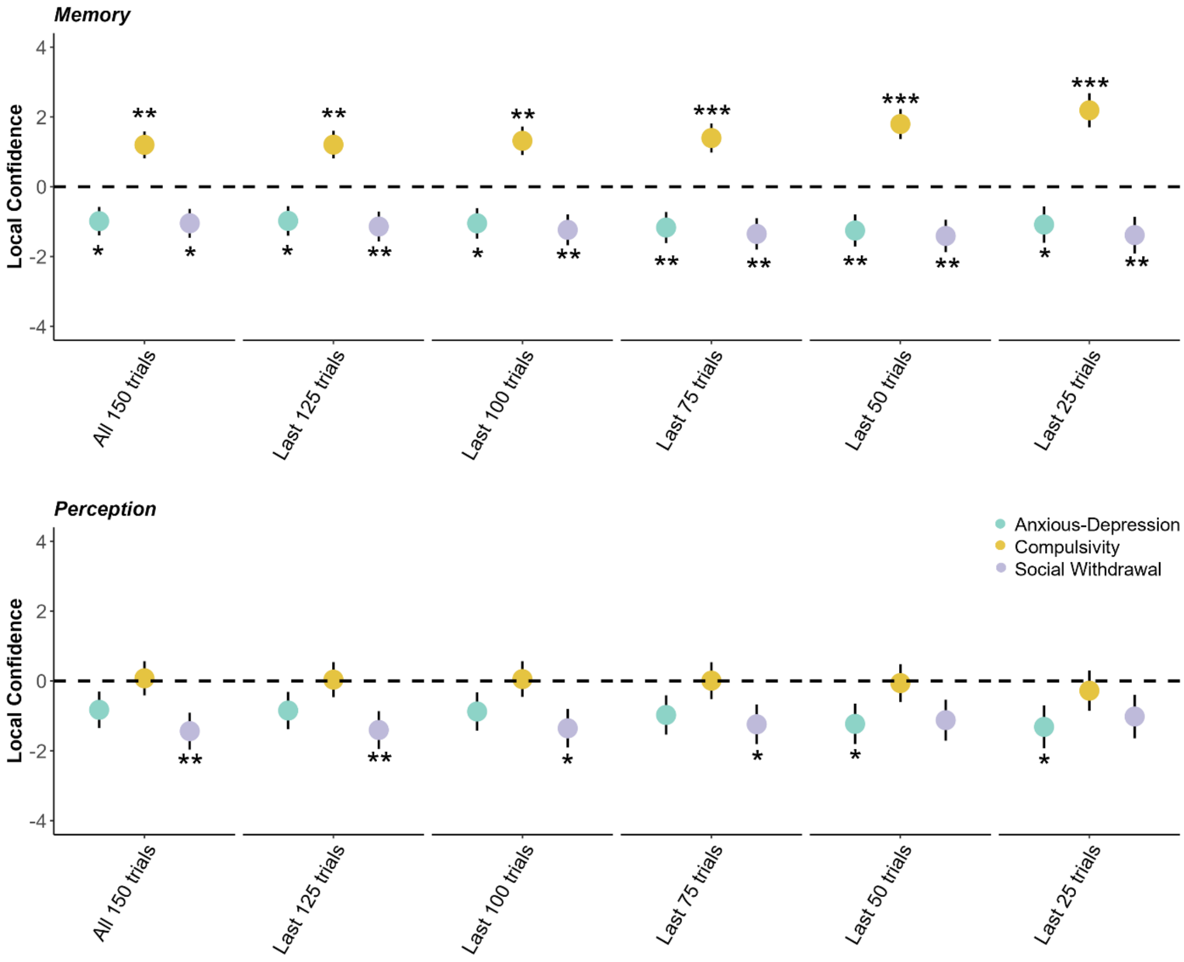
Supplementary Fig. 5. Correlations between item loadings for the three-factor structure from de novo factor analysis between the current sample (N=327) and Gillan et al. (2016) (N=1413). We find that even with a reduced sample size, we were able to replicate the same factor structure from the original analysis in Gillan et al. (2016). Each circle indicates an individual participant.



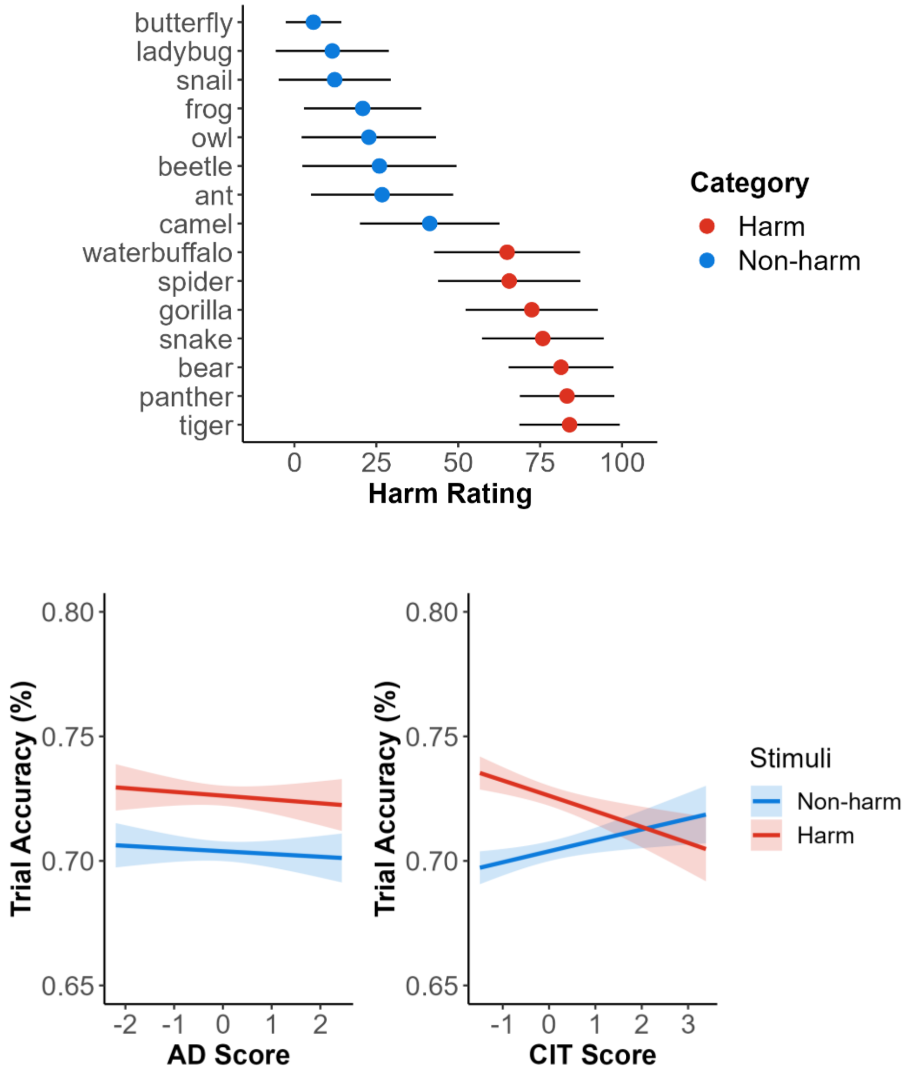
Supplementary Fig. 6. Individual meta-d' model parameter analyses and its relation to dimension scores. (A) Distributions of decision sensitivity (d'), metacognitive sensitivity (meta-d') and metacognitive efficiency (meta-d'/d'). We find lower d' (t(325)=-7.81, p<0.001, 95% CI=[-0.16 -0.10]), higher meta-d' (t(325)=23.41, p<0.001, 95% CI=[0.76 0.86]) and higher meta-d'/d' (t(325)=27.39, p<0.001, 95% CI=[0.77 0.89]) in the memory versus perception task. d' (r=-0.008, p=0.88, 95% CI=[-0.12 0.10]) and meta-d'/d' (r=0.07, p=0.19, 95% CI=[-0.04 0.18) were not correlated between tasks, with only a trending positive correlation of meta-d' (r=0.10, p=0.08, 95% CI=[-0.01 0.20]). However, we note that d' patterns are more distributed, with several participants 3SD away from the mean d' in perception (N=6) and memory (N=5). **(B) Associations of decision sensitivity, metacognitive sensitivity and metacognitive efficiency with dimension scores.** We observed no significant relationships between any of the metrics with any dimension. The regression models were controlled for task domain, task order, age, IQ and gender.



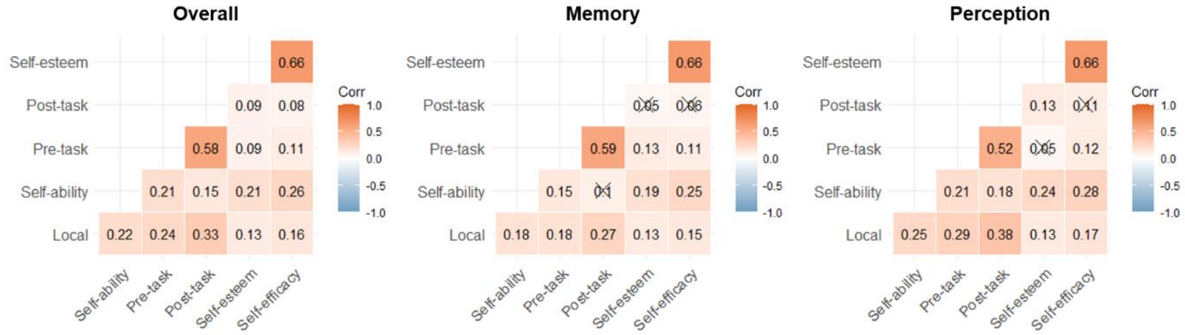
Supplementary Fig 7. HMeta-d’ model parameter analyses and their relation to dimension scores. (A) Markov chain Monte Carlo (MCMC) sample distributions for group level metacognitive efficiency (meta-d’/d’) for the perception and memory tasks, and the correlation between the estimates from both tasks. As the reliability of single-participant meta-d’/d’ estimation in Supplementary Fig. 6 can be low with small trial numbers, we ran the hierarchical model to get group level estimates. We found group level estimates were quite similar with the mean of the single-participant estimates for memory (mean meta-d/d’=1.54, 95% highest density interval (HDI)=[1.49 1.59]) and perception (mean meta-d/d’= 0.67, 95% HDI=[0.63 0.71]). We also found that the correlation between perception and memory group level meta-d’/d’ estimates was positive (mean ρ=0.11, 95% HDI=[-0.10 0.31]). **(B) Associations of group level metacognitive efficiency of the memory task with dimension scores.** We conducted the regression of the dimension scores on metacognitive efficiency, controlled for age, gender and IQ, within the hierarchical model. We found that memory metacognitive efficiency was not linked to any of the dimensions (AD: mean=0.03, 95% HDI=[-0.007 0.07]; CIT: mean=0.006, 95% HDI=[-0.03 0.04]; SW: mean=-0.01, 95% HDI=[-0.05 0.03]). **(C) Associations of group level metacognitive efficiency of the perception task with dimension scores.** Similarly, we found that perceptual metacognitive efficiency was not linked to any of the dimensions (AD: mean=-0.005, 95% HDI=[-0.08 0.06]; CIT: mean=-0.005, 95% HDI=[-0.07 0.06]; SW: mean=-0.04, 95% HDI=[-0.03 0.11]).



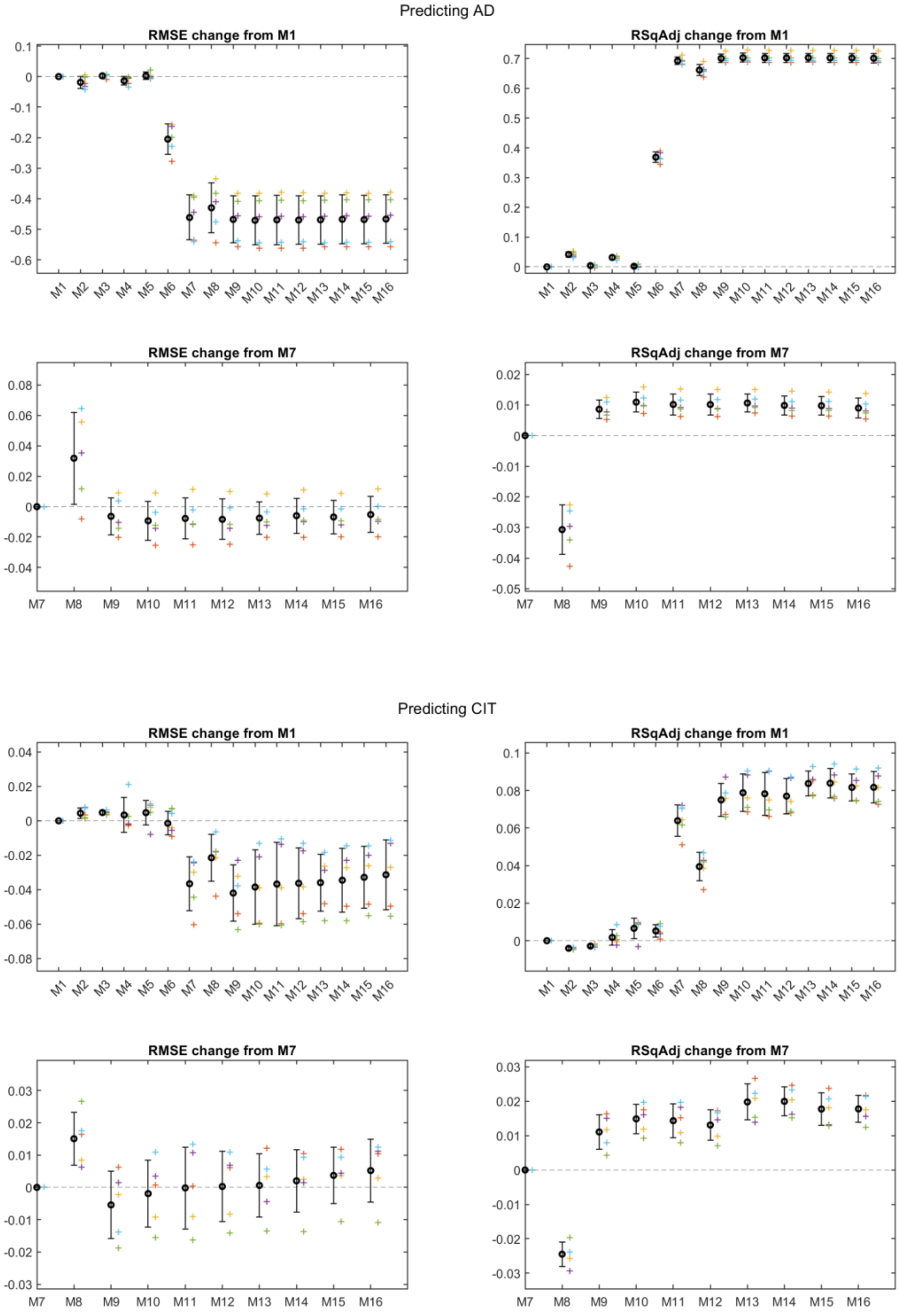
Supplementary Fig. 8. Regression of dimension scores on local confidence ratings for perception and memory over varying number of trials. The staircase procedure may have still required a burn-in for the beginning trials of the main task. We examined if associations of confidence with dimension scores would differ depending on the trials included in the analysis. We analysed the regression models separately for each task. All three dimension scores were included in the same model, which was controlled for task order, age, IQ and gender. We find similar patterns of results across each task even with varying trial numbers.



Supplementary Fig. 9. Harm levels of memory task stimuli and its relation to trial accuracy and dimension scores. In a prior affective rating study, we collected the harm ratings of the stimuli used in the memory task from N=80 participants (Seow & Hauser, 2024). Marker indicates the mean and standard deviation of the rating of each stimuli over two repeated presentations across participants. With the metamemory task, we assessed if the correct answer belonging to a harm versus non-harm category had an impact on the trial accuracy, and if it varied with dimension severity. Interaction effects show that high compulsive individuals do not show the accuracy enhancement effect for harm stimuli. AD: anxious-depression, CIT: compulsivity.



Supplementary Fig. 10. Correlations between metacognitive metrics. Pearson's correlations between the various metacognitive metrics obtained from the task—local (trial-by-trial task confidence), pre-task metacognition, post-task metacognition, self-ability metacognition as well as questionnaire scores of self-esteem and self-efficacy.

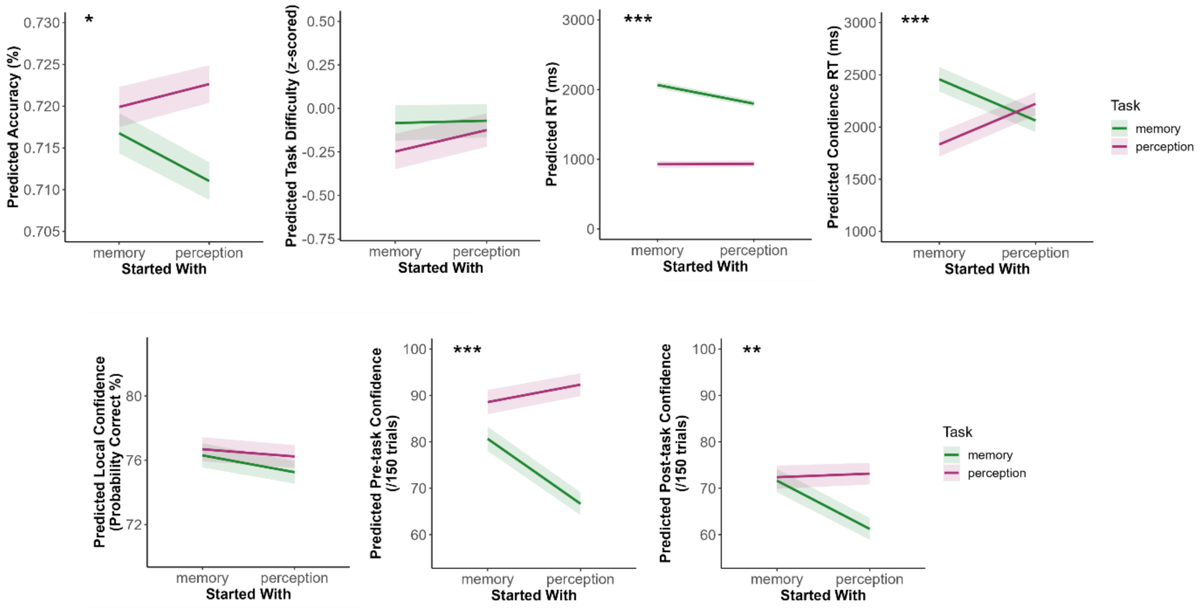


Supplementary Fig. 11. Cross validation of metacognitive models predicting anxious-depression (AD) or compulsivity (CIT) severity. We examined 16 different models (see below) to predict dimension scores with a 5-fold cross validation procedure. Each coloured cross indicates model fit from a fold run, while the black marker indicates the mean and standard deviation of the model fits across the folds. We assessed both root mean square error (RMSE) as well as the adjusted R^2 of the models. Model fits are depicted against a base model of demographics (age, gender and IQ; M1), as well as against a model of demographics with self-esteem (M7). For anxious-depression, the winning model with the highest adjusted R^2 is M10. For compulsivity, the winning model with the highest adjusted R^2 is M13.

Models examined with cross validation:

- M1 = Dimension ~ Age + Gender + IQ
- M2 = Dimension ~ Age + Gender + IQ + Domain Metacognition
- M3 = Dimension ~ Age + Gender + IQ + Pre-task Metacognition
- M4 = Dimension ~ Age + Gender + IQ + Local Confidence
- M5 = Dimension ~ Age + Gender + IQ + Post-task Confidence
- M6 = Dimension ~ Age + Gender + IQ + Self-efficacy
- M7 = Dimension ~ Age + Gender + IQ + Self-esteem
- M8 = Dimension ~ Age + Gender + IQ + mean(Self-efficacy+Self-esteem)
- M9 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem
- M10 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence
- M11 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Domain Metacognition
- M12 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Pre-task Metacognition
- M13 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Post-task Metacognition
- M14 = Dimension ~ Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Post-task Metacognition + Domain Metacognition
- M15 = Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Post-task Metacognition + Pre-task Metacognition
- M16 = Age + Gender + IQ + Self-efficacy + Self-esteem + Local Confidence + Post-task Metacognition + Pre-task Metacognition + Self-ability Metacognition

We chose the winning model as the one with the highest cross-validated (out-of-sample) adjusted R^2 and obtained the same results as the step-wise regression—the best model predicting anxious-depression included self-esteem, self-efficacy and local confidence (R^2=0.75, RMSE=0.50, LLR=0.50), while the best model predicting compulsivity included self-esteem, self-efficacy, post-task metacognition and local confidence (R^2=0.23, RMSE=0.89, LLR=0.10). Note that local confidence, pre-task metacognition, post-task metacognition and self-ability metacognition regressors were interacted with task domain to control for cognitive domain effects.



Supplementary Fig. 12. Task order effects. Participants were randomly assigned to complete either the metamemory or the metaperception task set first. Task order was taken as a control regressor in the main analyses, but showed some significant interaction effects with task domain in some analyses. Significance indicates significant interaction effect of task order with task domain on the dependent variable. *p<0.5, **p<0.01, ***p<0.001.

Supplemental Discussion

Local confidence and social withdrawal

We unexpectedly observed that social withdrawal exhibited similar associations with metacognition as those found for anxious-depression. Although we found that social anxiety questionnaire score distributions of the current sample were similar to those of our prior study (Rouault et al., 2018), we also found that correlations between the anxious-depression and social withdrawal factors were higher than expected (r=0.53 versus r=0.43 (Rouault et al., 2018) or r=0.33 (Hoven et al., 2023)). Indeed, when social withdrawal was not included in the regression models, our anxious-depression effects became much stronger (e.g. β=-0.12 to β=-0.20 when predicting local confidence). As social withdrawal has not shown significant relations with metacognition in any of the prior studies using the full three-factor transdiagnostic approach (Benwell et al., 2022; Hoven et al., 2023; Rouault et al., 2018) we refrain from interpreting these effects further and focus on anxious-depression and compulsivity.

Harm versus non-harm and compulsivity

Our metamemory task design allowed an exploratory analysis of the effect of harm versus non-harm answer stimuli on confidence. Prior work has implicated contextual effects of memory contents on its retrieval confidence in OCD (Tolin et al., 2001), with a stronger underconfidence effect for remembering “unsafe” objects. We found that “harm” stimuli boosted memory performance in the group as a whole, but this effect was blunted in high compulsive individuals. It might be that individuals high in compulsivity identify all items as high in “harm”. Although we selected stimuli that were norm rated as “harmful” and “not harmful” (Seow & Hauser, 2024), it might be more pertinent to use individual OCD-relevant stimuli for each participant in future work. While this is generally done in patients with OCD, it is more challenging in our dimensional approach with a general population where triggers are unknown and possibly less specific.

Supplemental References

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.

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health*, 1–9.

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.

Seow, T. X., & Hauser, T. U. (2024). What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli. *Wellcome Open Research*, *9*, 83.

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive–compulsive disorder. *Behaviour Research and Therapy*, *39*(8), 913–927.