Supplemental Information for
The metacognitive paradox of OCD: confidence is globally reduced but shows increased sensitivity to local evidence

Alisa M. Loosen[1-3], Brian A. Zaboski[3], Avalon Moore[3], Calvin Bohner[3-4], Helen Pushkarskaya[3], Christopher Pittenger[3,5*], Tobias U. Hauser[1-2, 6-7*]

[1] Max Planck UCL Centre for Computational Psychiatry and Ageing Research

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

[3] Department of Psychiatry, Yale University School of Medicine, New Haven, CT USA

[4] Frank H. Netter MD School of Medicine at Quinnipiac University

[5] Departments of Neuroscience and Psychology, Yale Child Study Center, Yale Center for Brain and Mind Health, and Wu-Tsai Institute, Yale University, New Haven, CT USA

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

[7] German Center for Mental Health (DZPG), Tübingen, Germany

*The authors contributed equally to the work

Corresponding author:

Alisa M. Loosen, PhD

alisa.loosen@yale.edu

Keywords: Obsessive-Compulsive Disorder, Rule-Shifting, Decision Making, Learning, Metacognition, Confidence, Cognitive Flexibility

Supplemental Methods

Extended Participant Description

All diagnoses of the 31 patients with OCD (16 identified as female, 13 as male and two as binary) recruited for this study were established by a licensed psychologist and/or psychiatrist using an unstructured clinical diagnostic interview and the Mini International Neuropsychiatric Interview (MINI)[1].To be eligible for the study, patients had to be unmedicated or stably treated with selective serotonin reuptake inhibitors (SSRIs).

Within approximately two weeks of behavioral testing, the severity of their obsessions and compulsions were assessed using the Yale-Brown Obsessive Compulsive Scale (Y-BOCS)[2]. Patients had to have a minimum total score of 16 (moderate severity of symptoms) to be included in the study. One patient with OCD was unable to complete the Y-BOCS interview but was still included in the analyses as the checklist and other clinical interviews recorded moderate to severe symptoms. Another patient was excluded from the analyses because they began taking SSRIs during the testing phase.

The 30 participants recruited as controls (15 identified as female and 15 as male) had never been diagnosed with any mental health disorder in the past. The group was matched with the group of participating patients with OCD on gender, age (Wilcoxon rank-sum test: $z = 0.327, p = 0.744$) and cognitive ability using the ICAR assessment tool (Wilcoxon rank-sum test: $z = 0.715, p = 0.471$; cf. below). Controls were first screened for any psychiatric history using a telephone interview. After this screening, they underwent an additional clinical online assessment, including the MINI and an interview version of the self-report tool PI-WSUR measuring obsessive-compulsive (OC) symptoms[3]. They were enrolled in the study as controls if they did not meet the criteria for any psychiatric diagnoses on the MINI and did not endorse any OC symptoms on the PI-WSUR at a level judged clinically significant by the Yale OCD Research Clinic staff (since the PI-WSUR does not have an official clinical cutoff score).

Table S1

Final Sample Characteristics

Variable	Participant Group	
	Controls	Participants with OCD
N	29	29
Gender (f/m/nb)	(14/15/0)	(15/12/2)
Age (M ± SD)	31 ± 11.43	32 ± 10.99
CA (M ± SD)	11.79 ± 3.49	11.38 ± 2.72
Education (count)		
Some college	9	4
Associate's degree	0	1
Bachelor's degree	11	10
Master's degree	6	7
Doctoral degree	3	7

Note. Shown are demographic characteristics. CA denotes cognitive ability, which was estimated using the international cognitive ability resource[4] and served as a proxy for IQ. The demographics of both groups were matched on a group level. F denotes female, m male and nb non-binary.

Table S2

Clinical Characteristics of Participants With OCD

Variable		Mean ± Standard deviation
Y-BOCS Score (M ± SD)*		
	Total	24.76 ± 4.07
	Obsessions	12.28 ± 2.17
	Compulsions	12.48 ± 2.53
Medication		
	Stable SSRIs	10
	Unmedicated	19
Comorbidities (count)		
Total number of participants with a comorbid diagnosis		10
	Agoraphobia	1
	Eating Disorder - NOS	1
	Body Dysmorphism	5
	Gender Dysphoria	5
	Depression - NOS	1
	Major Depressive Disorder	1
	Paranoid Personality Disorder	2
	Post-Traumatic Stress Disorder	4
	General Anxiety Disorder	1
	Illness Anxiety Disorder	3

Note. Shown are the clinical characteristics of participants with OCD. *One patient with OCD could not complete the Y-BOCS interview but was still included in the analyses as the checklist and other clinical interviews recorded moderate to severe symptoms. SSRIs stands for selective serotonin reuptake inhibitors and NOS stands for not otherwise specified.

Variable Definitions and Supplemental Behavioral Analyses

To measure cognitive flexibility, we analyzed general accuracy (cf. main manuscript) and the occurrence of perseverative errors. Perseverative errors, a specific subset of incorrect responses indicating cognitive inflexibility, were defined as choices made after a dimensional rule shift that incorrectly adhered to the previously rewarded dimension (e.g., choosing the previously rewarding blue stimulus instead of the newly rewarding round one). For our statistical analyses reported here, we applied standard corrections to ensure methodological rigor: p-values were adjusted for multiple comparisons using the Bonferroni method where applicable (denoted as p^*) and the Greenhouse-Geisser correction was used for repeated measures ANOVAs when the assumption of sphericity was violated (denoted as p_{GG}).

Model Comparisons

To determine the best fitting form of behavioral adaptation after rule shifts, we compared three mixed-effects models for each condition: a linear, a quadratic, and a logarithmic model. All competitive models consistently showed a significant effect in the same direction, confirming a robust overall pattern of adaptation and reported the winning model in the main manuscript and fort reaction time (RT) findings below. The BIC captures the fit of a statistical model and is based on the likelihood function, quantifying the probability of the observed data given the model, and has an additional penalty term for the number of model parameters, which takes model complexity into account. The BIC is defined as follows:

$$BIC = log(n) \times k - 2 \times log(L)$$

where n is the number of observations, k is the number of parameters in the model, and L is the maximum value of the likelihood function. For other mixed-effects analyses, we likewise used BIC scores to compare fixed- and random-effects structures and selected the most parsimonious specification.

Table S3

DV	Shift Type	Model	BIC
Accuracy	All Shifts	Linear	12429.648
		Quadratic	12348.629
		Logarithmic	**12257.745**
	IDs	Linear	5012.112
		Quadratic	4898.328
		Logarithmic	**4778.105**
	EDs	Linear	7368.123
		Quadratic	7381.918
		Logarithmic	**7363.941**
Confidence	All Shifts	Linear	104777.762
		Quadratic	**104742.758**
		Logarithmic	104972.913
	IDs	**Linear**	**47358.151**
		Quadratic	47372.334
		Logarithmic	47426.362
	EDs	Linear	57339.703
		Quadratic	**57264.454**
		Logarithmic	57462.035
Reaction Times (log-RTs)	All Shifts	**Linear**	**20361.694**
		Quadratic	20381.207
		Logarithmic	20453.624
	IDs	**Linear**	**8653.537**
		Quadratic	8660.154
		Logarithmic	8687.507
	EDs	**Linear**	**11698.594**
		Quadratic	11714.620
		Logarithmic	11756.008

Note. Reported are Bayesian Information Criterion (BIC) scores (lower = better fit; best in bold) for candidate models of post-shift dynamics in each behavioral measure; DV = dependent variable, IDs/EDs = intra-/extra-dimensional shifts.

Supplemental Results

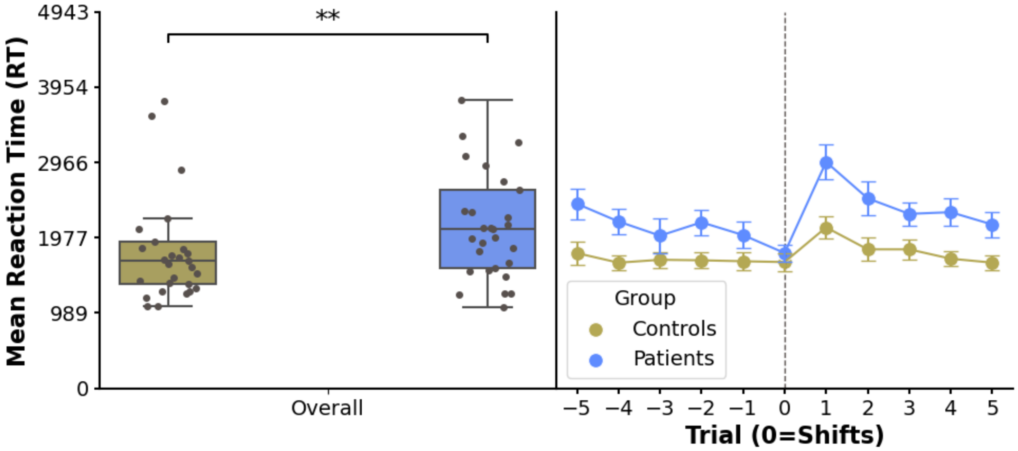
Reaction Times Increase With Difficulty Task Changes and are Longer in Participants with OCD

We investigated how reaction times (RTs) developed across task difficulty levels in participants with OCD compared to healthy controls. To address the right-skewed distribution of RTs, all values were log-transformed prior to statistical tests. We ran a mixed ANOVA with group as a between-participant factor (participants with OCD vs. controls) and task level (difficulty level 1-3) as the within-participant factor. This showed a significant main effect of difficulty level ($F(2{,}112) = 21.353, p_{GG} < 0.001$; Greenhouse-Geisser correction was applied) and group ($F(1{,}56) = 5.330, p < 0.05$) on average log-RT. However, the interaction between difficulty level and group (patients versus controls) was not significant ($F(2, 112) = 1.766, p = 0.176$), thus, the relationship between difficulty level and RT did not differ significantly between the patient and control groups. Between-group comparison on subject-mean log-RT confirmed slower responses in the OCD group ($t(56) = -2.570, p = 0.013$). A Mann–Whitney U test was performed to compare raw RTs between patients with OCD and controls; results showed a significant difference ($U = 154.000, N^1 = 29, N^2 = 29, z = -4.144, p < 0.01$; Supplemental Figure 1, left panel). Pairwise comparisons (within-subject, Bonferroni-corrected) on log-RT showed that level 1 was not statistically different from level 2 (*t*(57)=2.384, *p*=0.061*), while level 3 responses were slower than both level 1 (*t(*57*)*=3.642, *p*=0.002*) and level 2 (*t(*57*)*=7.167, *p*<0.001*). For interpretability, raw RT (ms) by level were: level 1 ($M = 1940, SD = 1900$), level 2 ($M = 1671, SD = 717$), and level 3 ($M = 2709, SD = 3174$).

We also examined RTs relative to changes in the task (i.e. intra-dimensional [ID] and extra-dimensional [ED] shifts) and preceding accuracy. Mixed-effects models (on log-RT) showed significantly longer RTs on trials following shifts ($\beta = 0.111, \mathrm{SE} = 0.033, p < 0.01$). There was a main effect of group ($\beta = 0.153, \mathrm{SE} = 0.075, p < 0.05$), with individuals with OCD responding more slowly, and a significant group × prior-shift interaction ($\beta = 0.166, \mathrm{SE} = 0.047, p < 0.001$), indicating a larger post-shift increase in RTs in the patient group. Consistent with this pattern, correct choices on the preceding trial predicted shorter RTs ($\beta = -0.222, SE =$

$0.057, \mathrm{p} < 0.001$); there was a significant main effect of group ($\beta = 0.272, SE = 0.094, p < 0.01$) but only a trend for a group × preceding-accuracy interaction ($\beta = -0.146, SE = 0.080, p = 0.075$).

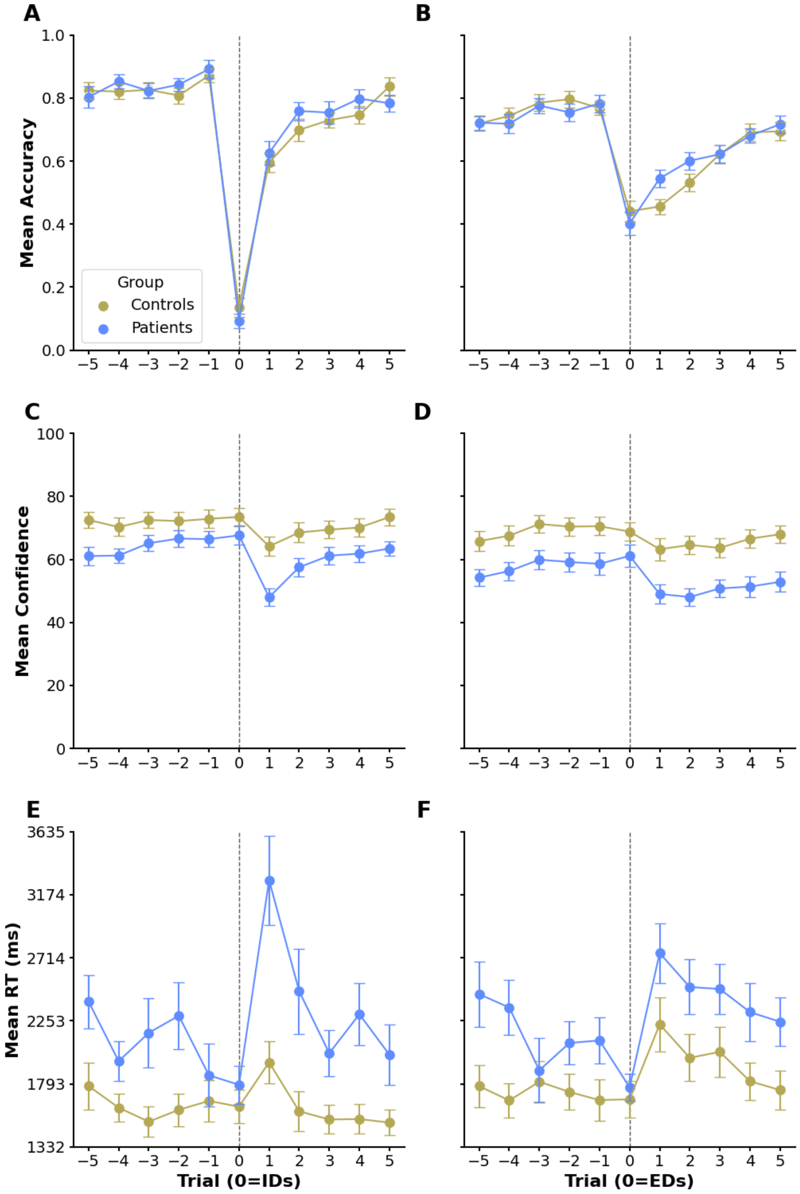
To characterize RT recovery after rule shifts, we modeled log-RT as a function of trial distance from the shift (covarying age, gender, IQ; cf. Supplemental Methods). Pooling across groups, log-RT decreased with distance, consistent with recovery to baseline (ID shift: $\beta = -0.026, SE = 0.003, p < 0.001$; ED shifts: $\beta = -0.022, SE = 0.004, p < 0.001$; cf. Supplemental Figure 1 illustrating learning patterns in the right panel and Table S3 and accompanying text for model comparison). Adding the group variable did not yield significant group × distance interactions (ID shifts: $\beta = -0.007, SE = 0.007, p = 0.319$; ED shifts: $\beta = -0.010, SE = 0.008, p = 0.222$), indicating comparable post-shift recovery in individuals with OCD and controls. Main effects in the group-augmented models showed slower overall responding in the OCD group (ID model: $\beta = 0.225, SE = 0.083, p < 0.01$; ED model: $\beta = 0.189, SE = 0.081, p < 0.05$) and distance effects that remained significantly negative (ID model: $\beta = -0.022, SE = 0.005, p < 0.001$; ED model: $\beta = -0.018, SE = 0.006, p < 0.01$).



Supplemental Figure 1. **Average reaction times.** The box plots display the mean reaction times (RTs) across all trials for participants with OCD ($N_{OCD} = 29$) and controls ($N_{Controls} = 29$; **left panel**). RTs for both groups increased after a rule-shift happened and decreased again with following trials (**right panel**). The right panels display the mean behavioral measures from five trials before to five trials after each shift and error bars depict the standard error of the mean (SEM). While raw RTs are plotted here for visualization, statistical comparisons between groups were performed using log-transformed RT data to better meet statistical assumptions (***p<0.01*).

Behavioral Patterns Around Shifts

Overall, participants exhibited similar behavioral adaptation dynamics following both intra-dimensional (ID) and extra-dimensional (ED; cf. Supplemental Figure 2). Specifically, immediately following both types of shifts mean accuracy and mean confidence typically showed a marked decrease, followed by gradual recovery in subsequent trials. Mean RTs mirrored this pattern, showing a sharp increase (slowing) immediately post-shift, which subsequently decreased towards baseline levels.



Supplemental Figure 2. **Behavioral performance dynamics relative to intra-dimensional (ID) and extra-dimensional (ED) shifts.** Mean performance metrics participants with OCD ($N_{OCD} = 29$) and controls ($N_{Controls} = 29$) plotted for the five trials preceding and the five trials following rule shifts. Trial 0 indicates the trial a shift occurred. **(A-B)** Mean accuracy **(C-D)** mean confidence and (**E-F**) mean reaction times (RTs; in milliseconds [ms]) and followed similar patterns across relative to intra-dimensional (IDs; left column; **A, C, & E**) and extra-dimensional shifts (EDs; right column; **B, D, & F**) as described in the main manuscript (i.e. relative to shifts in general). The displays performance relative to data points represent the group mean, and error bars depict the standard error of the mean (SEM).

To gain additional understanding of the effect of rule shifts on behavioral variables (accuracy, confidence, and RTs), we examined how average performance varied relative to Intra-Dimensional (ID) and Extra-Dimensional (ED) shifts. For this analysis, trials were grouped relative to the shift event (Trial 0) into three distinct time bins based on their position:

1. Bin 1 ('Before/During Shift'): Included the five trials immediately preceding the shift plus the shift trial itself (covering trials -5, -4, -3, -2, -1, and 0 relative to the shift).
2. Bin 2 ('First Trial After'): Consisted solely of the first trial immediately following the shift (Trial +1 relative to the shift).
3. Bin 3 ('Recovery Phase'): Included the second, third, fourth, and fifth trials following the shift (covering trials +2, +3, +4, and +5 relative to the shift).

We then calculated the mean value for each behavioral variable within each of these bins for every participant. These binned averages were subsequently analyzed using mixed-design ANOVAs, with Trial Bin (Bin 1, Bin 2, Bin 3) as the within-participant factor and Group (participants with OCD vs. controls) as the between-participant factor. Analyses were performed separately for ID and ED shifts.

When predicting average accuracy for ID shifts, we found a significant main effect of trial bin ($F(2, 112) = 27.184, p_{GG} < 0.001$), indicating accuracy changed across the bins relative to the shift. However, there was no significant main effect of group ($F(1, 56) = 0.617, p = 0.435$) and no significant interaction between trial bin and group ($F(2, 112) = 0.205, p = 0.815$). Paired t-tests showed that the average accuracy of controls significantly dropped from bin 1 to bin 2 ($t(28) = 3.897, p < 0.01$), while for participants with OCD this drop did not survive correction for multiple comparison ($t(28) = 2.400, p < 0.05, p^{*} = 0.069$). Both groups significantly recovered in average accuracy from bin 2 to bin 3 (OCD: $t(28) = 4.213, p^* < 0.001$; controls: $t(28) = 4.884,\ p^* < 0.001$). Accuracy in bin 3 was significantly higher than bin 1 only for participants with OCD ($t(28) = 2.850, p^* < 0.01$), while for controls it did not survive correction for multiple comparison ($t(28) = 2.104, p < 0.05, p^* = 0.133$).

For ED shifts, we found a significant main effect of trial bin ($F(2, 112) = 57.631, p_{GG} < 0.001$), no significant main effect of group ($F(1, 56) = 3.109, p = 0.083$), but a significant interaction between trial bin and group ($F(2, 112) = 3.873, p^* = 0.024$)). Paired t-tests showed that the average accuracy of controls significantly dropped from bin 1 to bin 2 (OCD: $t(28) = 4.216, p * < 0.001$; controls: $t(28) = 10.075, p^* < 0.001$) and significantly recovered from bin 2 to bin 3 (OCD: $t(28) = 3.486, p < 0.01$; controls: $t(28) = 6.392, p < 0.001$). Average accuracy in bin 3 was significantly higher than bin 1 for controls ($t(28) = 3.248, p^* = 0.003$), but not after correction for multiple comparison in the participant group with OCD ($t(28) = 2.185, p < 0.05, p^* = 0.037$). The significant interaction appeared potentially driven by the between-group difference in bin 2, where participants with OCD had lower accuracy than controls, however this difference did not survive Bonferroni correction ($t(56) = -2.420, p < 0.05, p^* = 0.057$). No significant group differences surviving correction were found in bin 1 ($t(56) = 0.732$, p=0.467) or bin 3 ($t(56) = -0.848, p = 0.400$).

We performed the same analysis for the average confidence ratings as the dependent variable. Relative to both ID shifts and ED shifts, we again found a significant main effect of trial bin (ID: $F(2, 112) = 51.287, p_{GG} < 0.001$; ED: $F(2, 112) = 20.419, p_{GG} < 0.001$) indicating that participants adapted their confidence ratings around them. We also found a main effect for group relative to both types of shifts (ID: $F(1,56) = 10.152, p < 0.01$; ED: $F(1,56) = 10.920, p < 0.01$). A significant interaction between trial bin and group was found for ID shifts (F(2,112)=6.089, p<0.01) but not for ED shifts ($F(2,112) = 1.625, p = 0.202$).

Follow-up paired t-tests within groups for ID shifts showed that both, participants with OCD and controls, experienced a significant drop in confidence from bin 1 (before/during) to bin 2 (first after shift; OCD: $t(28) = 6.763,\ p^* < 0.001$; controls: $t(28) = 4.269, p^* < 0.003$) and a significant recovery from bin 2 to bin 3 (second-fifth after; OCD: $t(28) = 6.225,\ p^* < 0.001$; controls: $t(28) = 3.652,\ p^* < 0.01$). Furthermore, patients' average confidence in bin 3 remained significantly lower than bin 1 ($t(28) = 3.590,\ p^* < 0.01$), while controls' average confidence returned to baseline levels ($t(28) = 1.436, p = 0.162$). For ED shifts, both participants with OCD and controls showed a significant drop from bin 1 to bin 2 (OCD: $t(28) = 4.463,\ p^* < 0.001$; controls: $t(28) = 2.563, p\ ^* < 0.05$). Confidence did not significantly recover between bin 2 and bin 3 for either group (OCD: $t(28) = 1.167, p = 0.253$; Controls: $t(28) = 1.542, p\ = 0.134$), and bin 3 confidence remained significantly lower than bin 1 for both groups (OCD: $t(28) = 5.266, p^* < 0.001$; controls: $t(28) = 2.776,\ p^* < 0.05$).

Between-group independent t-tests showed that for ID shifts, patients had significantly lower confidence than controls in bin 2 ($t(56) = 3.928,\ p^* < 0.001$) and bin 3 ($t(56) = 2.585,\ p^* < 0.05$), but the difference in Bin 1 was not significant after correction ($t(56) = 2.145,\ p^* = 0.108, p < 0.05$). For ED shifts, patients consistently had significantly lower confidence than controls across all three bins (bin 1: $t(56) = 2.667,\ p^* < 0.05$; bin 2: $t(56) = 3.005,\ p^* < 0.05$; bin 3: $t(56) = 3.756, p^* < 0.001$).

We predicted the averages of log-transformed RTs based on the described ANOVAs (separately for the different shifts), we found a significant main effect of trial bin relative to ID shifts ($F(2, 112) = 33.012, p_{GG} < 0.001$) as well as ED shifts ($F(2,112) = 17.852,\ p_{GG} < 0.001$). We found a main effect of group for the analysis of both shift types (ID shifts: $F(1, 56) = 13.390, p < 0.01$; ED shifts: $F(1, 56) = 5.806, p = 0.019$). We also found a significant interaction effect between trial bin and group for ID shifts ($F(2,112) = 4.804, p < 0.05$) but not ED shifts ($F(2, 112) = 0.051, p = 0.950$). This suggests that patients with OCD and controls displayed significantly different adaptation patterns after the intra-dimensional shifts but not extra-dimensional shifts.

The results of follow-up paired t-tests showed that participants' average log-RTs on the first trial after the shifts (bin 2) were significantly longer than their average log-RTs on the trials ranging from five trials before to the shift trial (bin 1) for both groups and both shift types (participants with OCD ID: $t(28) = 5.643,\ p^* < 0.001$; controls ID: $t(28) = 2.893,\ p^* < 0.05$; participants with OCD ED: $t(28) = 4.033,\ p^* < 0.001$; controls ED: $t(28) = 3.015,\ p^* < 0.05$). The average log-RTs on the first trial after the shifts (bin 2) were also significantly longer than the average log-RTs on the second to fifth trials after the shifts (bin 3) for both groups during ID shifts (participants with OCD: $t(28) = 5.443,\ p^* < 0.001$; controls: $t(28) = 3.733,\ p^* < 0.01$) and for participants with OCD during ED shifts ($t(28) = 2.821,\ p^* < 0.05$); this comparison was not significant after correction for multiple comparison for controls during ED shifts ($t(28) = 2.090,\ p^* = 0.138, p\ = 0.046$). Additionally, the average log-RTs on the second to fifth trial after shifts (bin 3) were not significantly different from the trials before/at the shifts (bin 1) for ID shifts (participants with OCD: $t(28) = 0.673,\ p^{*} < 1.000, p\ = 0.506$; controls: $t(28) = 1.376, p^{*} = 0.540, p\ = 0.180$). However, for ED shifts, bin 3 log-RTs remained significantly longer than bin 1 log-RTs for participants with OCD ($t(28) = 2.919,\ p^{*} < 0.05$), while this comparison was not significant after correction for controls ($t(28) = 2.257,\ p^{*} = 0.096, p\ = 0.032$). This pattern indicates potentially faster log-RT recovery relative to baseline after ID shifts compared to ED shifts, particularly for controls.

Conducting follow-up independent t-tests, we saw significant differences in average log-RTs between participants with OCD and controls during ID shifts across all bins (bin 1: $t(56) = 2.532, p *< 0.05$; bin 2: $t(56) = 3.921, p *< 0.001$; bin 3: $t(56) = 3.201, p *< 0.01$), with participants with OCD consistently showing longer log-RTs. For ED shifts, however**,** no significant group differences survived Bonferroni correction in any bin (bin 1: $t(56) = 2.144, p *= 0.108, p < 0.05$; bin 2: $t(56) = 1.898, p *= 0.189, p = 0.063$; bin 3: $t(56) = 2.391, p *= 0.060, p < 0.05$).

Confidence Differences Prevail After Controlling for Depressive Symptoms and Medication Status

As noted in the main manuscript, participants with OCD showed significantly lower confidence ratings than controls. Given the high rate of comorbid major depressive disorder (MDD) and OCD[5] and accounts (primarily based on general-public studies investigating obsessive-compulsive symptoms on a transdiagnostic level) in the literature suggesting, lowered confidence in OCD can be attributed to heightened depression scores[6,7], we ran an additional control analysis to test this possibility in our sample. We measured depressive symptoms in both the OCD and healthy control groups using the Zung Self-Rating Depression Scale (SDS)[8]. These SDS scores were then included as a covariate in our regression analyses predicting confidence. The results of this analysis showed that while depression scores in our sample showed a negative link with confidence that did not reach significance ($\beta = -0.052, SE = 0.194, p = 0.792$), the negative main effect of group on confidence prevailed ($\beta = -8.934, SE = 3.931, p < 0.05$).

This indicates that participants with OCD remained significantly less confident than controls, even when statistically controlling for depressive symptoms in both groups.

Additionally, we examined whether medication status within the patient group was a confounding factor. A direct comparison of average confidence scores between medicated ($N = 10, M = 60.077, SD = 11.378$) and unmedicated ($N = 19, M = 57.344, SD = 14.448$) patients, using an independent samples t-test, revealed no significant difference ($t(27) = 0.518, p = 0.609$).

We further examined whether confidence in the OCD group was influenced by psychiatric comorbidities. For most diagnosed comorbidities, sample sizes were insufficient ($N < 3$; see Supplementary Table S2) to permit meaningful statistical comparison. However, for comorbidities with a larger representation ($N \geq 3$), independent samples t-tests revealed no significant differences in mean confidence. Specifically, confidence levels were comparable between patients with and without Body Dysmorphic Disorder ($M_{with} = 58.494, SD = 13.434\ vs.\ M_{without} = 58.243, SD = 13.592; t(27) = -0.038, p = 0.970$), Gender Dysphoria ($M_{with} = 60.022$,$SD = 8.580$ *vs.* $M_{without} = 57.925, SD = 14.232;\ t(27) = -0.315, p = 0.755$), Illness Anxiety Disorder ($M_{with} = 64.757, SD = 4.713$ *vs.* $M_{without}$ =57.540, *SD*=13.837*;* $t(27) = -0.885, p = 0.384$), and Post-Traumatic Stress Disorder (PTSD; M_{with} =65.448, *SD*=9.243 *vs.* $M_{without} = 57.141, SD = 13.658;\ t(27) = -1.165, p = 0.254$*)*. These results, in conjunction with prior analyses of depression and medication status, demonstrate that the primary group difference in confidence is a robust finding.

Exploratory Associations Between OCD Symptom Severity (Y-BOCS) and Behavioral/Metacognitive Indices

Within the OCD group, and consistent with the main manuscript, we observed a significant negative correlation between the per-participant regression coefficient linking confidence to Bayesian choice certainty (CC) and Y-BOCS Total ($r_s = -0.398, p < 0.001$). Breaking this down by subscales (Spearman, due to non-normality), the overall effect was primarily driven by the Y-BOCS compulsions score ($r_s = -0.468, p *< 0.01$), whereas the correlation with the obsessions score showed a non-significant negative trend ($r_s = -0.209, p * = 0.104, p = 0.052$). Complementary Pearson analyses indicated that higher symptom severity was also associated with weaker alignment between confidence ratings and CC ($r_p = -0.421, p < 0.05$) and with lower overall task accuracy ($r_p = -0.379, p < 0.05$) but not with GC ($r_s = -0.220, p = 0.25$) or average confidence level ($r_p = -0.221, p = 0.250$). These correlations should be interpreted with caution given potential clinical confounders, but the parallel and specific negative associations with both CC–confidence alignment and accuracy suggest that higher symptom load may broadly impair efficient task engagement rather than selectively altering metacognitive tracking independent of performance.

Supplemental References

1. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* 59 Suppl 20, 22-33;quiz 34-57 (1998).
2. Steketee, G., Frost, R. & Bogart, K. The Yale-Brown Obsessive Compulsive Scale: interview versus self-report. *Behav. Res. Ther.* 34, 675–684 (1996).
3. Shams, G., Kaviani, H., Esmaili, Y., Ebrahimkhani, N. & Manesh, A. A. Psychometric Properties of the Persian Version of the Padua Inventory: Washington State University Revision (PI-WSUR). *Iran. J. Psychiatry* 6, 12–18 (2011).
4. Condon, D. M. & Revelle, W. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* 43, 52–64 (2014).
5. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. (American Psychiatric Association, 2013). doi:10.1176/appi.books.9780890425596.
6. Gillan, C. M., Fineberg, N. A. & Robbins, T. W. A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychol. Med.* 47, 1528–1548 (2017).
7. 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).
8. Zung, W. W. A Self-Rating Depression Scale. *Arch. Gen. Psychiatry* 12, 63–70 (1965).