Isolating domain-specific and domain-general contributions to global confidence

Sophie Bavard, Andrew Mcwilliams, Flora Chartier, Karim N'Diaye, Stephen Fleming, Marion Rouault

Title:

Isolating domain-specific and domain-general contributions to global confidence

Authors:

Sophie Bavard*[1], Andrew McWilliams*[2,3], Flora Chartier[1], Karim N'Diaye[1], Stephen M. Fleming°[2,4,5], Marion Rouault°[1]

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France

[2] Department of Experimental Psychology, University College London, London, UK

[3] Child and Adolescent Mental Health Services, Royal Free London NHS Foundation Trust, Pond Square, London, UK

[4] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

[5] Institute of Cognitive Neuroscience, University College London, London, UK

* equal contribution as first authors

° equal contribution as last authors

Correspondence:

sophie.bavard@gmail.com

marion.rouault@gmail.com

Keywords:

metacognition; confidence; memory; perception; self-performance estimates; decision-making; self-evaluation

ORCID numbers:

SB: 0000-0002-9283-2976

SF: 0000-0003-0233-4891

KN: 0000-0003-2142-0576

MR: 0000-0001-6586-3788

Abstract (145 words)

While metacognition - our ability to evaluate our own cognitive processes - has been extensively studied by measuring trial-by-trial self-assessments of task performance (local confidence), real-world decisions often require a broader perspective, drawing on evaluations of performance over wider timespans (global confidence). Despite its pervasive influence on decision-making and mental health, it remains unknown whether global confidence is formed through similar or distinct processes across cognitive domains. Here we employ a novel gamified approach to compare how global confidence is formed in the domains of memory and perception. In memory, we found that both local accuracy and confidence contributed to global confidence, whereas in perception, global confidence was predicted by local confidence alone. By comparing the formation of global confidence across domains, our study provides new insights into the mechanisms underpinning self-evaluation, paving the way for the development of metacognitive interventions in education and psychiatry.

Introduction

Metacognition - our ability to monitor and evaluate our own cognitive processes – has a pervasive role in adaptive learning and decision-making (Fleming, 2024; Metcalfe & Shimamura, 1994). For instance, confidence judgments play a crucial role in guiding how we allocate cognitive resources and adjust decision strategies (Fleming et al., 2012; Risko & Gilbert, 2016; Son & Metcalfe, 2000). Recent models have proposed that confidence is formed across a range of hierarchical timescales. “Local” confidence refers to trial-specific judgments of performance elicited around the time of a particular decision (e.g., Rouault et al., 2023) whereas “global” confidence tracks overall success aggregated across multiple trials (Lee et al., 2021; Rouault et al., 2019). Local confidence, also referred to as “decision” confidence, is a self-estimate of the probability that a choice is correct (Pouget et al., 2016). Previous studies across species (Kepecs & Mainen, 2012; Sanders et al., 2016) and cognitive domains (for reviews, see Mazancieux et al., 2023; Rouault et al., 2018) have allowed the development of rich cognitive and computational models of confidence formation and metacognitive sensitivity, revealing that local confidence formation depends on evidence (Gherman & Philiastides, 2015; Zylberberg et al., 2012), response times (Kiani et al., 2014) and the integration of post-decision evidence (Hilgenstock et al., 2014; Murphy et al., 2015; van den Berg et al., 2016). Moreover, the distinct factors influencing local confidence formation have been studied across different domains such as general knowledge (e.g., Lund et al., 2025), memory (e.g., Mazancieux et al., 2020), time estimation (e.g., Rouault et al., 2023), and perception (e.g., Sanders et al., 2016; Spence et al., 2016; Wilimzig et al., 2008). In contrast, how global confidence (also referred to as global self-performance estimates) is generated is still unclear (though see Katyal et al., 2025; Lee et al., 2021; Rouault & Fleming, 2020).

Specifically, it remains to be determined how local metacognitive computations contribute to self-evaluations over longer timescales, and how these mechanisms may vary across domains. For example, imagine a person trying to cook a new dish and bake a new cake on different days while assessing how confident they are about how it will turn out. At the end of the month, they realize they consistently feel more successful in cooking compared to baking, leading them to generate a broader self-evaluation of the form “I am better at cooking than baking”. Recent studies indicate that local and global metacognition rely on overlapping but distinct neural substrates (Rouault & Fleming, 2020), suggesting a hierarchy of metacognitive processes in the human brain (Purcell & Kiani, 2016; Seow et al., 2021). While decision confidence informs immediate adjustments in decision-making, global self-performance estimates are thought to shape our behavior and mental health over extended timescales (Bandura, 1977), determining the goals we choose to pursue, and the effort we put into our endeavors (Elliott et al., 1996; Zacharopoulos et al., 2014;) Orth et al., 2008; Seow et al., 2021). In other words, individuals may rely more on this broader sense of competence when making strategic choices (Bandura, 1977). Characterizing how individuals form global self-performance estimates is therefore critical for capturing aspects of metacognitive processes that may be particularly relevant to everyday decision-making, and to the subjective and functional symptoms experienced by psychiatric and neurological patients (Seow et al., 2021). While local confidence has

been extensively studied across domains, much less is known about the factors shaping self-performance estimates and whether these factors are similar or distinct across domains.

The current study builds on a previously developed gamified suite of web-based experiments designed to measure local metacognition in different cognitive domains (McWilliams et al., 2023). We extended this platform to incorporate measures of global confidence and tested human participants across two cognitive domains (memory and perception) and two sessions (test and retest). We found that participants integrated local confidence and accuracy to form task-level judgments but did so differently for memory and perception. To the extent that metacognition is hierarchically organized (from local to task-level to domain-level to self-evaluations), we expected that metacognitive estimates would be more stable across time points than across domains. We found some evidence that local metacognitive bias was indeed more reliable between sessions than between domains. Our study provides new insights into the architecture of global confidence across cognitive domains and paves the way for longitudinal studies of local and global metacognitive capacities in patient and developmental cohorts.

Results

Experimental protocol assessing local and global metacognition

We employed novel measures of local and global metacognition across (i) two domains: short-term memory and visual perception and (ii) two sessions: test and retest, separated by three weeks. In the memory domain, participants were asked to memorize a set of items and afterwards select the familiar stimulus when paired with a distractor. In the perception domain, participants were presented with an array of multiple identical red and blue shapes and then asked whether there were more red or blue stimuli. For both memory and perception, retrospective ("local") confidence judgments were elicited after each trial using a horizontal visual sliding scale, the ends of which were labelled "complete guess" and "absolutely certain". Within each domain, participants completed 6 blocks of 20 trials each, for a total of 120 trials per domain. Within each block, trials from two different categories were interleaved. For example, participants were asked to memorize strange artifacts (category 1) in half of the trials and alien insects (category 2) in the other half. At the end of each block, participants were asked to choose between the two categories of the block in which category they thought they performed best (in our example, strange artifacts or alien insects). This category choice is a measure proposed to reflect task-level ("global") confidence in previous work (Rouault et al., 2019; **Figure 1**).

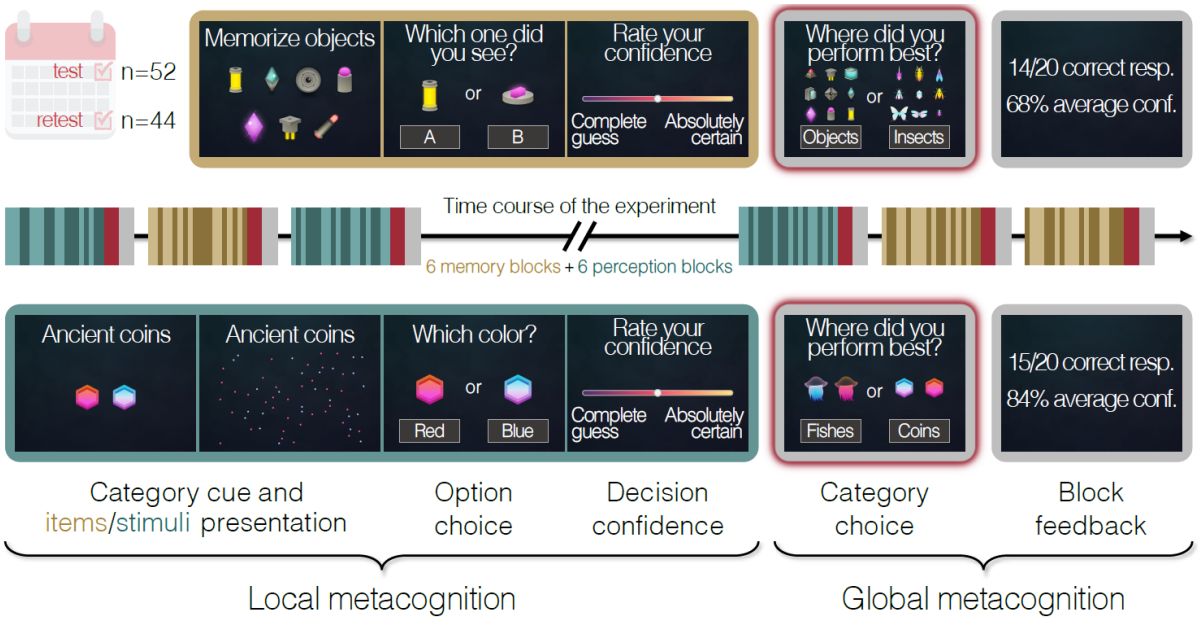


Figure 1. Experimental design assessing local and global metacognition in perception and memory. In the memory domain (yellow), participants were presented with a set of items and then were asked to select which of two items was previously seen. In the perception domain (green), participants were presented with a cloud of red and blue items and were asked to select whether there were more blue or red items. In both domains, participants were then asked to indicate their confidence in their decision being correct on a continuous scale from 'Complete guess' to 'Absolutely certain'. Participants completed mini blocks comprising 10 decisions of each category. At the end of each block, they had to select the category in which they thought they performed best, reflecting their estimation of self-performance at the task level ('Category choice'). End-of-block feedback provided participants with their average accuracy and their average local confidence over the block ('Block feedback'). To estimate the reliability of local and global metacognition metrics, the same participants were invited to complete a retest session three weeks later.

To control for variation in first-order task performance, a continuous adaptive staircase procedure dynamically adjusted task difficulty (García-Pérez, 1998; Levitt, 1971). We extracted trial-level difficulty, accuracy, response time (RT) and local confidence as candidate predictors of global confidence. These measures were subsequently analyzed as a function of self-performance estimates, i.e., whether a category was chosen or unchosen at the end of the block.

Construction of global self-performance estimates in each domain

At the end of each block, participants indicated in which category they believed they performed best. First, we investigated the factors contributing to the formation of self-performance estimates in each domain and in each session. We fitted four generalized linear mixed models (GLMM) predicting end-of-block category choices from local difficulty, accuracy, RT and confidence (**Figure 2; Figure S1**):

$$\text{chosen category} \sim \text{z(difficulty)} + \text{z(accuracy)} + \text{z(RT)} + \text{z(confidence)} + (1 \mid \text{participant})$$

We report here the results based on only the four main effects of local variables, as adding interaction terms did not significantly improve the model fit (BIC main vs. complete, memory test: 25814 vs. 25923, perception test: 25662 vs. 25783, memory retest: 21869 vs. 21976, perception retest: 21751 vs. 21859). (**Table S1-S2-S3-S4**). In both sessions, we found that participants relied on different factors to estimate their self-performance between domains (**Figure 2C-2D-2G-2H, Table S1-S3**). In the memory domain, we found a positive contribution of local accuracy (test session: $\text{beta} = 0.075, \text{SE} = 0.028, t = 2.7, p = 0.0075$; retest session: $\text{beta} = 0.063, \text{SE} = 0.030, t = 2.1, p = 0.039$) and local confidence (test session: $\text{beta} = 0.23, \text{SE} = 0.034, t = 6.8, p < 0.0001$; retest session: $\text{beta} = 0.24, \text{SE} = 0.036, t = 6.7, p < 0.0001$) to self-performance estimates. This was in the absence of significant contributions from other local indices (all $|\text{beta}| < 0.06, p > 0.05$). We note that, when considered in isolation (model-free analysis), RTs were longer (test session: $t(51) = 2.1, p = 0.0037$; retest session: $t(43) = 3.5, p = 0.0011$) and difficulty was harder in the test session ($t(51) = 2.6, p = 0.011$; retest session: $t(43) = 1.5, p = 0.13$) for the unchosen as compared to chosen category (**Figure 2A-2E**). However, their contribution was not significant when these metrics were put in competition with all local variables to explain end-of-block category choice in a GLMM (**Figure 2C-2G**). In contrast, in the perception domain, we found a positive contribution of local confidence (test session: $\text{beta} = 0.095, \text{SE} = 0.028, t = 3.4, p = 0.00064$; retest session: $\text{beta} = 0.076, \text{SE} = 0.030, t = 2.5, p = 0.012$) to self-performance estimates. This was found in the absence of contributions from other local cues (all $|\text{beta}| < 0.05, p > 0.08$).

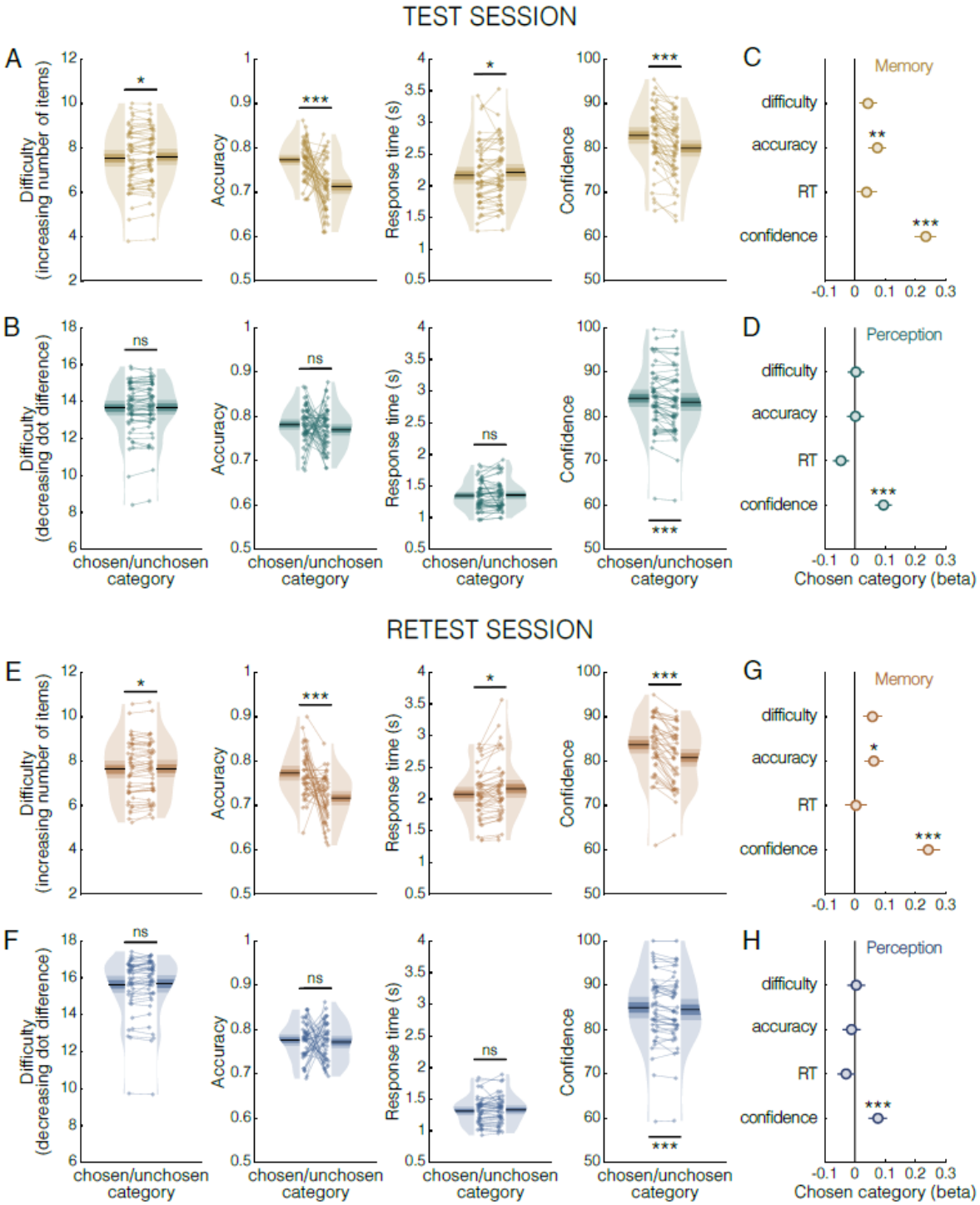


Figure 2. Factors contributing to the construction of global self-performance estimates on task. A, B, E, F. Local difficulty, accuracy, response time, confidence as a function of chosen vs. unchosen category at the end of blocks, reflecting high vs. low self-performance estimates in the memory (**A, E**) and perception (**B, F**) tasks respectively. **A-D**: test session (N=52); **E-H**: retest session (N=44). Points and thin lines indicate individual averages, shaded areas indicate probability density function, 95% confidence interval, and SE. Stars indicate statistical significance of a paired t-test between chosen and unchosen categories. Difficulty corresponds to the number of items in memory, and to the difference in red and blue shapes in perception. In memory, the significant difference in difficulty does not hold when all local variables are considered at the same time (see panels **C, G**). **C, D, G, H**. GLMM predicting end-of-block task choices in the memory (**C, G**) and perception (**D, H**) domains as a function of the difference in local difficulty, accuracy, response times and confidence between categories (see Methods for details). Circles and error bars indicate mean and SE over regression coefficients. Stars indicate statistical significance of a one sample t-test against zero at the group level for each coefficient. In all panels, ns: $p>0.05$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

To formally assess the extent to which the local predictors of self-performance estimates vary across domains, we added domain as an additional fixed effect in the GLMM (pooling both sessions). We further specified interaction terms between domain and each of the four predictors: local difficulty, accuracy, RT, and confidence (see **Methods**). First, we found a significant interaction between domain and confidence ($\text{beta} = 0.075, \text{SE} = 0.017, t = 4.5, p < 0.0001$, **Figure 3**, **Table S5**), indicating a larger effect of local confidence on global self-performance estimates in memory as compared to perception. We also identified a significant interaction between domain and accuracy ($\text{beta} = 0.041, \text{SE} = 0.015, t = 2.8, p = 0.0058$, **Figure 3, Table S5**). This is also consistent with within-domain results and means that local accuracy contributed more strongly to global self-performance estimates in memory than in perception (**Figure 2C-2D-2G-2H**). Taken together, these results confirm both domain-general and domain-specific patterns of local contributions to global self-performance estimates. Notably, local decision confidence emerged as a key domain-general contributor to global self-performance across both perception and memory tasks.

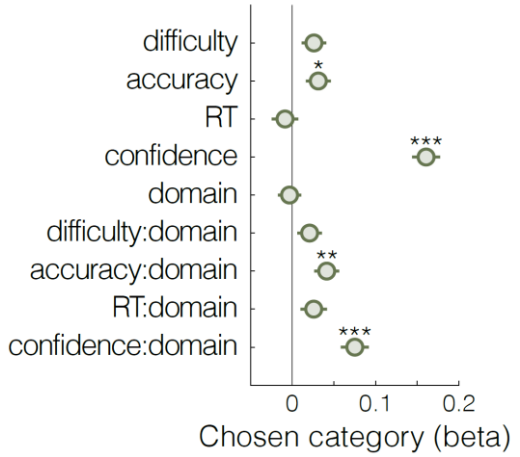


Figure 3. Different local cues contribute to global self-performance estimates between domains. GLMM predicting end-of-block category choices, pooled over sessions, as a function of the difference in the four local factors between categories and their respective interaction with domain. Circles and error bars indicate mean and SE over regression coefficients. Stars indicate statistical significance of a one sample t-test against zero at the group level for each coefficient. In all panels, * p<0.05, ** p<0.01, *** p<0.001.

Between-domain comparison of local metrics

We also examined between-domain differences in local metacognition. To do so, we focused on the session with the largest sample size (test session, N=52, **Figure 2-4**), but we note that virtually identical results are observed in the retest session (N=44, **Figure 2-S2**). Although the overall average accuracy in both domains was slightly higher than the expected target level (memory: $\text{mean} \pm \text{SD} = 0.74 \pm 0.026$; perception: $\text{mean} \pm \text{SD} = 0.78 \pm 0.020$), this was mostly driven by earlier blocks (final block accuracy vs. target level, memory: $t(51) = -0.091, p = 0.93$; perception: $t(51) = 1.4, p = 0.18$). Overall, all participants performed within a narrow range (**Figure 2A-2B-2E-2F**), indicating that the staircase procedure worked appropriately. Additionally, we found no correlation in accuracy between memory and perception (Spearman's $\rho(50) = 0.057, p = 0.69$), suggesting that the staircase procedure effectively adjusted difficulty level for the two tasks independently. Given that accuracy was

staircase-controlled, average difficulty served as a proxy indicator of first-order performance capacity. We found no correlation across domains in difficulty level (Spearman's $\rho(50) = 0.11, p = 0.45$), suggesting that the first-order abilities required to complete the perception and memory tasks were distinct. In contrast, we found significant positive correlations between domains for both RT (Spearman's $\rho(50) = 0.38, p = 0.0056$) and local confidence (Spearman's $\rho(50) = 0.72, p < 0.0001$), indicating that these behavioral measures may be influenced by trait-level variations.

Between-domain comparison of metacognitive bias and metacognitive efficiency

Metacognitive bias, defined as the discrepancy between mean confidence and mean accuracy, was positive in both memory ($\text{mean} \pm \text{SD} = 0.071 \pm 0.073,\ t(51) = 6.98, p < 0.0001$) and perception ($\text{mean} \pm \text{SD} = 0.061 \pm 0.076,\ t(51) = 5.81, p < 0.0001$), indicating overconfidence in both domains of about 7% and 6% on a 50-100% scale (**Figure 4A**). We observed equivalent levels of overconfidence across the two domains (memory vs. perception, $t(51) = 1.2, p = 0.23$), and these levels were significantly correlated across participants (Spearman's $\rho(50) = 0.69, p < 0.0001$; **Figure 4B**). This suggests that the degree of overconfidence may be a stable individual trait transcending domains, in line with prior work (Ais et al., 2016; Binnendyk et al., 2024).

We also examined the evolution of metacognitive bias across blocks in each domain. Although participants did not receive feedback at the trial level, at the block level participants received feedback on their confidence and their accuracy (**Figure 1**). This allowed us to examine whether participants were able to recalibrate their confidence throughout the experiment, if necessary. We found evidence against recalibration, as overconfidence kept increasing over trials (memory: Spearman's $\rho(118) = 0.40, p < 0.0001$; perception: Spearman's $\rho(118) = 0.62, p < 0.0001$) and over blocks (memory: Spearman's $\rho(4) = 0.94, p = 0.017$; perception: Spearman's $\rho(4) = 0.94, p = 0.017$; **Figure 4C**). We also examined the effect of block feedback on overconfidence in a GLMM predicting block-averaged metacognitive bias as a function of the previous block-averaged accuracy and previous block-averaged confidence (see Methods). For memory, we found a significant positive effect of previous accuracy ($\text{beta} = 0.011, \text{SE} = 0.0053, t = 2.0, p = 0.043$) and a significant negative effect of previous confidence ($\text{beta} = -0.014, \text{SE} = 0.0053, t = -2.7, p = 0.0074$) on metacognitive bias. In contrast, for perception, we found no significant effect of accuracy ($\text{beta} = -0.00071, \text{SE} = 0.0058, t = -0.12, p = 0.90$) or previous confidence ($\text{beta} = -0.0061, \text{SE} = 0.0058, t = -1.1, p = 0.29$) on metacognitive bias. These results indicate that while participants in the memory domain showed some sensitivity to previous block-level feedback, no such effect was observed in the perception domain. Together, this suggests a limited, domain-specific use of feedback in adjusting metacognitive judgments, though the format of feedback may also have played a role (see **Discussion**).

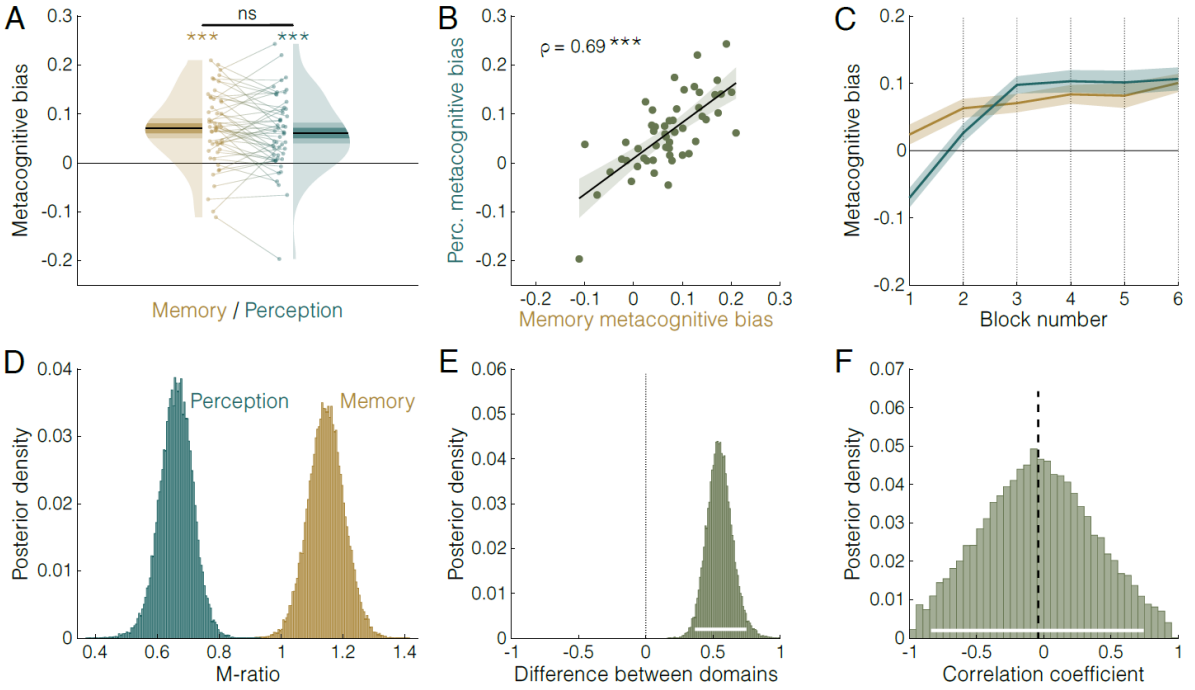


Figure 4. Between-domain comparison of metacognitive bias and metacognitive efficiency. A. Average metacognitive bias across participants in memory (yellow) and perception (green) tasks. Points and grey lines indicate individual average, shaded areas indicate probability density function, 95% confidence interval, and SE. N=52. **B.** Scatterplot of metacognitive bias between domains indicating a significant positive correlation (p < 0.0001, see main text). Shaded area represents 95% confidence intervals and black line is a linear regression fit. N=52. **C**. Trajectories of metacognitive bias measured as the discrepancy between local confidence and accuracy throughout the six blocks in memory (yellow) and perception (green) domains. Shaded areas indicate SE over participants. N=52. **D.** Group-level metacognitive efficiency (M-ratio) distribution estimated hierarchically in the memory (yellow) and perception (green) domains (see Methods). **E.** Group-level difference (in log units) between the group posteriors. The white bar represents the 95% HDI which excludes zero (dotted line), indicating a significantly higher metacognitive efficiency in the memory than the perception domain. **F.** Group-level correlation coefficient (dotted line) between metacognitive efficiencies in the two domains. The white bar represents the 95% HDI which includes zero, indicating no correlation between domains.

Second, metacognitive efficiency, defined as how well local confidence discriminates between correct and incorrect decisions with respect to objective accuracy, was quantified here using the hierarchical meta-d'/d' (M-ratio) framework (see Methods; Fleming, 2017; Maniscalco & Lau, 2012). At the group level, we found above-zero M-ratios of 1.2 ± 0.16 (mean ± SD) in memory and 0.69 ± 0.13 (mean ± SD) in perception (**Figure 4D**), driven by participants giving higher confidence on correct than incorrect decisions (memory: $t(51) = 17, p < 0.0001$; perception: $t(51) = 13, p < 0.0001$; **Figure S3A,B**). Metacognitive efficiency was significantly higher in memory compared to perception (95% HDI on difference in M-ratio: [0.37, 0.75]; **Figure 4E**). However, no significant correlation in metacognitive efficiency was observed between domains (correlation coefficient = -0.042; 95% HDI on correlation coefficient: [-0.84, 0.75]; **Figure 4F**). Model-free measures of metacognitive sensitivity derived from the type-2 area under the receiver operating curve (see **Methods**) gave very similar results (**Figure S3C,D**).

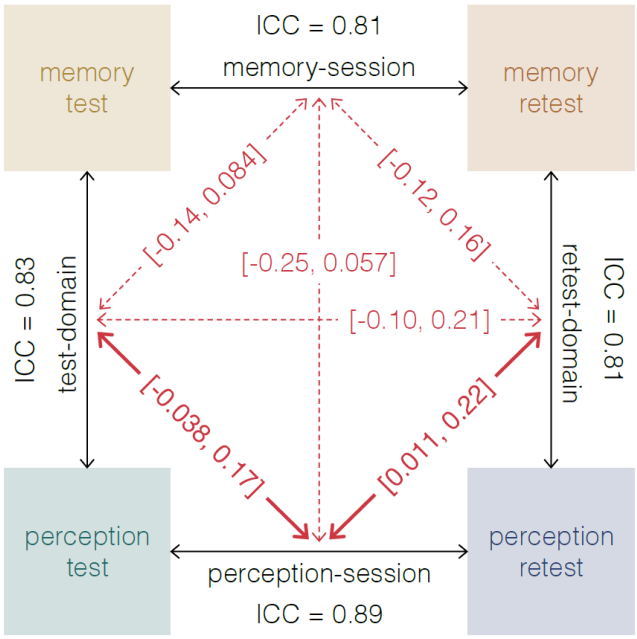
Exploratory analysis: assessing a hierarchical architecture for local metacognitive bias

In an exploratory analysis, we sought to assess whether metacognitive judgements may follow a hierarchical structure from local to global self-evaluations across domains, as has been previously theorized (Seow et al., 2022). To pursue this hypothesis, our logic was to compare the strength of between-domain and between-session associations. More precisely, we reasoned that any changes in metacognitive measures between test and retest sessions could be attributable to a true change in the metacognitive capacity of a person (i.e., a meaningful psychological shift), to noise in measurement, or a combination of both. Here, under the assumption of no measurement noise, if self-performance estimates emerge within a hierarchical architecture of metacognitive judgements, we expect between-session associations to be stronger than between-domain associations (**Figure 5**). In other words, if metacognition is hierarchical, then within a given domain, the same lower-level mechanisms should be reused across sessions for a given individual. This should lead to relatively more stable measures within a domain across time points.

To formally assess this, we first identified the most reliable metric across test and retest sessions by computing intraclass correlation coefficients (ICC, Koo & Li, 2016) for all local contributors to global self-performance estimates and local measures of metacognition described above. From these analyses, local metacognitive bias stood out as the most reliable measurement in both domains (detailed results are provided in **Figure S4** and **Table S7-S8**).

Then, to compare the strengths of associations, we used a two-step procedure. First, we computed ICCs to estimate the following four associations (**Figure 5**, black arrows): memory-session (ICC test vs. retest), perception-session (ICC test vs. retest), test-domain (ICC memory vs. perception), retest-domain (ICC memory vs. perception). Second, we compared the strength of the four associations (see **Methods**), leading to a total of six confidence intervals (CI) measuring if each association is stronger than the three others (**Figure 5**, red arrows).

We report here the results for each domain in turn. In memory, the effect sizes of domain vs. session correlations were similar ($95\%\ \mathrm{CI} = [-0.14, 0.084]$ and $95\%\ \mathrm{CI} = [-0.12, 0.16]$; **Figure 5**, top diagonal arrows). However, in perception, the test-retest association was significantly stronger than both the retest-domain reliability ($95\%\ \mathrm{CI} = [0.011, 0.22]$), and the test-domain reliability (marginally, $95\%\ \mathrm{CI} = [-0.038, 0.17]$; **Figure 5**, bottom diagonal arrows). This pattern of findings for perception is consistent with a hierarchical architecture of metacognition, in which local metacognitive bias exhibits a greater stability within- than across-domains.



298

299 **Figure 5. Testing the hierarchical structure of metacognitive bias.** Black arrows represent associations
300 between metacognitive bias estimated in each domain and each session separately. Red arrows represent the
301 relative strength of one association as compared to the three others, with the 95% CI of bootstrapped difference
302 in ICCs (see **Methods**). Dashed arrows: non-significant difference in association strength. Bold arrows:
303 significance and marginal significance. ICC: intraclass correlation coefficient of metacognitive bias.

Discussion

Recent years have witnessed growing interest in characterizing the mechanisms underlying metacognition and how confidence is formed at different levels of abstraction (Seow et al., 2021). Despite a strong focus on computational models of local confidence (decision confidence), the mechanisms underlying global confidence (evaluations of one's own performance over longer timescales) remain unclear. Here, we investigated which local factors contribute to global self-performance estimates and whether they are the same in different domains. We designed a gamified experiment measuring both local and global confidence across memory and perception domains (McWilliams et al., 2023). We established that both local accuracy and confidence influenced the formation of global self-performance estimates, but differently so across domains. We further found that local metacognitive bias was the most reliable measure both across domains and sessions, whereas metacognitive efficiency was not reliable. Finally, we observed some evidence that the test-retest association of metacognitive bias was stronger than the between-domain association, supporting the view that metacognition operates across multiple hierarchical levels. We discuss each of these findings in turn.

Previous work has revealed that local confidence and external feedback about performance influence global self-performance estimates (Haddara & Rahnev, 2022; Katyal et al., 2025; Lee et al., 2021; Rouault et al., 2019; Wittmann et al., 2016). We found that in the memory domain, both local accuracy and confidence positively predicted self-performance estimates, whereas in the perception domain, local confidence alone was a significant predictor. Critically, this was found in the context of a lack of influence of response times (RT) or objective difficulty on self-performance estimates, cues previously identified to be tightly linked to both local (Kiani et al., 2014; Zylberberg et al., 2012) and global (Lee et al., 2021) confidence. Whether the domain-specific contributions of local cues to global confidence reflect a common mechanism integrating information from specific sources, or whether they arise from completely distinct, domain-specific architectures, remains to be investigated. The consistent contribution of local confidence in both domains confirms and extends prior findings of a key role for local confidence over and above accuracy and RT in the formation of perceptual self-performance estimates (Rouault et al., 2019).

At the local level, we found that confidence, RT, and metacognitive bias were positively correlated across domains, suggesting the existence of stable, trait-like features of local metacognition. Moreover, the consistency of metacognitive bias was supported by a comparable degree of overconfidence across domains, in line with prior findings (Ais et al., 2016; Binnendyk et al., 2024; Seow et al., 2025; West & Stanovich, 1997). Interestingly, we did not find evidence for a better calibration in the retest as compared to the test session, nor across successive blocks within each session. This may at first appear surprising, since we provided participants with feedback regarding their average accuracy and average confidence at the end of each block. However, several factors may have limited any feedback effect. Feedback format (e.g., 14/20 for accuracy and 37% for confidence) may not have been the most appropriate or easiest to interpret, as the units of confidence did not match the fact that chance level was 50% correct. Moreover, it remains possible that too few

blocks are present for recalibration to occur. Finally, it is possible that perceived performance might have been category-dependent rather than domain-dependent, leading to a potential 'reset' between categories, limiting opportunities for gradual adjustment. Together, these results suggest that metacognitive bias may reflect a trait characteristic.

In contrast to metacognitive bias, metacognitive efficiency was higher for memory than for perception and did not show across-domain or across-session reliability. However, we note that since the time of data collection, several studies have indicated that much larger sample sizes are needed to obtain reliable estimates of cross-task correlations in metacognitive efficiency. Therefore, the lack of correlation across domains we observed may reflect limited statistical power (Lund et al., 2025).

To probe the hierarchical structure of metacognition, we performed an exploratory analysis comparing the effect sizes of between-domain and between-session associations for local metacognitive bias. We found no significant difference between these two associations for memory. However, the between-session reliability of metacognitive bias within perception was significantly higher than its between-domain association. This latter pattern provides additional support for a hierarchical organization of metacognitive evaluation. Although a similar pattern might also exist in the memory domain, it is possible that the present design did not offer enough variability in accuracy and confidence due to the staircase procedure to observe it. Furthermore, the limited number of blocks could have prevented the observation of consistent contributions of local factors to self-performance estimates across sessions. Future work should aim to replicate and extend these findings by increasing the number of blocks and sessions (e.g., Fox et al., 2024).

Several limitations of our study should be acknowledged. First, although we did our best to match the psychometric properties of the two domains, we noticed that the stimuli in memory were more salient. This potentially helped participants better track the current category as compared to the perceptual domain, despite the category being reminded in text on top of the perceptual stimulus (**Figure 1**). Moreover, the difficulty (number of items) may have been more evident in the memory domain, compared to perception where the dot difference is not immediately countable. These small differences may have helped participants to self-evaluate more easily in the memory domain. Second, the lack of influence of RT on the formation of self-performance estimates in perception may be due to relatively long stimulus presentation times during which participants could not respond, which may have impacted the RT collected during the following response window if the decision had already been made. Finally, while it can be argued that the binary nature of our self-performance estimates measurements is more ecological as confidence-based choices are ubiquitous, recent work suggests that self-performance estimates reported on a continuous scale provide a more fine-grained measure of global metacognitive estimates (Katyal et al., 2025).

Our results open several promising avenues for future research. First, it remains unknown how participants track more than two self-performance estimates at the same time, which is presumably relevant to real-life contexts such as educational settings. Another avenue involves the development of computational models to characterize the mechanisms underlying global self-performance estimate

formation. For example, reinforcement learning could formalize how participants update global confidence based on trial-by-trial feedback when it is available or local confidence when it is not, whereas Bayesian models could capture the probabilistic integration of accuracy and confidence with prior beliefs about self-performance in different domains or with domain-general self-beliefs (Heilbron & Meyniel, 2019; Katyal et al., 2025; Rouault et al., 2019). These approaches could offer a principled way to test competing hypotheses about the sources of global confidence variations across domains, be it differences in information sampling, learning rates, or strength of prior expectations about self-performance (Fleming & Daw, 2017; Rouault et al., 2019; Van Marcke et al., 2024). Finally, neuroimaging in other domains could extend our prior results showing that both local and global metacognitive variables modulated ventromedial prefrontal cortex and precuneus activity in a perceptual task (Rouault & Fleming, 2020) using designs adapted to contrast brain activity across domains (Morales et al., 2018).

To conclude, our study provides novel insights into the domain-specificity of the construction of global confidence. By identifying which factors are shared or distinct across domains, our findings pave the way for formal modeling approaches to characterize the mechanistic interplay between components of confidence.

Materials and methods

Ethics statement

The research was approved by the INSERM Ethical Review Committee (approval number IRB00003888) and carried out following the principles and guidelines for experiments including human participants provided in the Declaration of Helsinki (1964, revised in 2013).

Participants

We recruited 53 participants (17 females, 36 males, aged 34.2±1.6 years old) via the Prolific platform (www.prolific.co). All included participants provided online informed consent prior to their inclusion and were invited to perform the experiment a second time three weeks later. A total of 46 participants (16f/30m, 34.8±1.7 years old) completed the experiment during the second (retest) session (**Figure 1**). After exclusions (see below), the final samples were 52 participants for the test session and 44 participants for the retest session. The study was advertised to last for an hour (per session) and paid £8. In the end, participants obtained an average of £9.62 an hour for the test session and £9.78 for the retest session (median completion time was 49 minutes for both sessions).

Exclusion criteria

Trial exclusion. We performed the trial exclusion analyses separately for each domain (memory/perception) and each session (test/retest). We excluded trials for which the response time (RT) was below/above the within-participant mean ± 2.5 std in either domain or either session (Berger & Kiefer, 2021; Van Selst & Jolicoeur, 1994), leading to an average exclusion of 2.7% of the trials (per participant: minimum 1, maximum 7).

Participant exclusion. After performing trial exclusion, we excluded participants whose average RT was below/above the across-participant mean ± 2.5 std in either domain and either session, leading to an exclusion of 1 participant in the first session and 2 participants in the retest session. The participant excluded in the test session was also one of the excluded participants in the retest session. Therefore, a final sample of 44 participants of the retest session completed both domains and both sessions.

Behavioral tasks

We employed a design using novel measures of local and global metacognition in two domains (short-term memory and visual perception). To ensure participants remained motivated and engaged throughout our experiment, we extended the original setup of “Metacogmission”, a gamified web-based environment, to deliver performance-controlled perception and memory tasks (McWilliams et al., 2023). We collected both trial-by-trial (“local”) confidence judgments and end-of-block category choices (“global confidence”) in two domains (https://fr.metacogmission.com). Metacogmission has a gamified storyline, placing participants on an alien planet where they can explore and complete metacognitive tasks. Short-term memory trials consist of a memorization set of several items presented for 2 seconds, followed by a two-alternative forced-choice (2-AFC) requiring participants to

select the familiar stimulus previously seen over a new unseen item presented beside it. Perceptual discrimination trials involved presentation of an array of multiple identical red and blue shapes for 3 seconds, followed by a 2-AFC of whether there were more red or blue stimuli. For both memory and perception, retrospective (local) confidence judgments were elicited after each choice using a horizontal visual sliding scale, the ends of which were labelled “complete guess” and “absolutely certain” (**Figure 1**). The pointer was placed initially in the center and there were no visible divisions on the slider, generating confidence rating data on a near-continuous 201-point scale (coded in arbitrary units as 50 to 100). Trials were presented in blocks of 20, with each block having different thematic contents. Within each block, participants were presented with trials from two different item categories (e.g., fish and plant in the memory domain; coins and minerals in the perception domain). At the end of each block, participants were asked in which category they thought they did best (**Figure 1**), which has been proposed to reflect task (“global”) confidence (Lee et al., 2021; Rouault et al., 2019; Rouault & Fleming, 2020). Each block was preceded by additional optional practice trials. To aid engagement and motivation, at the end of each block participants received feedback on their average accuracy and confidence in that block. Participants were required to complete 6 blocks in each of the memory and perception domains to complete the task. If they submitted extra attempts at blocks, these trials were not analyzed. Importantly, in both domains, first-order task performance was controlled using a 2-down-1-up staircase procedure, which in the limit ensures first-order task performance converges to ~71% correct (García-Pérez, 1998; Levitt, 1971). The memorization set consisted initially of 3 stimuli, with the staircase increasing the set size by 1 (after 2 consecutive correct trials) or decreasing by 1 (after 1 incorrect trial). The first perceptual discrimination trial showed a difference of 15 between the numbers of shapes of the 2 colors, with the staircase increasing (after 1 incorrect trial) or decreasing (after 2 consecutive correct trials) this difference by 1 to make the task easier or harder. The staircase ran throughout blocks without resetting at each new block, separately for each domain. We acknowledge that, under the null hypothesis of random responding, the staircase procedure would maintain an average accuracy around 70% correct. In contrast, confidence ratings, reported uniformly, would average around 75%, leading to a spurious overconfidence of 5%. However, uniform random responding is very unlikely given that most participants rated their confidence higher on correct than incorrect decisions (**Figure S3A,B**).

Statistical analyses of behavior

For both domains and in both sessions, we were interested in 4 local variables (trial-level difficulty, accuracy, RT, confidence) and 1 global variable reflecting participant’s self-performance estimates (block-level category choice). Within-domain correlations between these 4 local variables across participants are presented in **Figure S5**. Despite the staircase procedure, there were natural fluctuations in objective difficulty and accuracy across trials, hence across short blocks. These variations allowed us to examine whether participants considered variability in their own performance when they form global confidence estimates for each category. To investigate which factors contribute to the formation of self-performance estimates, we conducted a Generalized Linear Mixed Model (GLMM) on the category choice, with a Binomial distribution of the response variable

(chosen/unchosen category) and a Logit link function, with z-scored local difficulty, accuracy, RT and confidence as within-participant predictors (**Table S1-S3**):

$$\text{chosen category} \sim \text{z(difficulty)} + \text{z(accuracy)} + \text{z(RT)} + \text{z(confidence)} + (1 \mid \text{participant})$$

Moreover, we conducted the same GLMM including all the main effects of independent variables and all their interactions (double, triple, quadruple; **Table S2-S4**):

$$\text{chosen category} \sim \text{z(difficulty)} * \text{z(accuracy)} * \text{z(RT)} * \text{z(confidence)} + (1 \mid \text{participant})$$

Finally, to formally test for an effect of domain on the factors contributing to global confidence, we initially considered performing a GLMM where all predictors would be z-scored over the whole dataset for each participant, i.e., pooling domains. Since all predictors were z-scored separately within each domain, the interaction terms with domain capture between-domains differences in the within-domain slopes of these predictors. Therefore, we conducted a GLMM including domain as a predictor and domain-related interactions (double interactions only) with difficulty, accuracy, RT and confidence (**Table S5**):

$$\text{chosen category} \sim \text{z(diff)} * \text{dom} + \text{z(acc)} * \text{dom} + \text{z(RT)} * \text{dom} + \text{z(conf)} * \text{dom} + (1 \mid \text{participant})$$

An analogous analysis was conducted to test for an effect of session on the factors contributing to global confidence, with a GLMM including session as a predictor and session-related interactions (double interactions only) with difficulty, accuracy, RT and confidence (**Table S6**):

$$\text{chosen category} \sim \text{z(diff)} * \text{ses} + \text{z(acc)} * \text{ses} + \text{z(RT)} * \text{ses} + \text{z(conf)} * \text{ses} + (1 \mid \text{participant})$$

For all GLMMs, we report the estimates (beta), standard errors, *t* statistics, and *p*-values. In **Figure 2A-B-E-F**, we report the significance for local metrics of one-sample *t* tests between chosen vs. unchosen category.

At the local level, metacognition is traditionally studied using two measures. Metacognitive bias is the discrepancy between mean confidence and mean accuracy, with positive and negative values indicating overconfidence and underconfidence, respectively. Metacognitive sensitivity is the capacity for confidence to distinguish between correct and incorrect responses. Hence, both metacognitive measures rely on accuracy and confidence.

First, we estimated metacognitive bias by subtracting accuracy from confidence ratings that had been rescaled to match the meaning of the confidence scale (from Guessing to Certain). Since participants received feedback on their accuracy and performance at the end of each block, we hypothesized that participants might re-calibrate their metacognitive bias from one block to the next. To quantify the effect of end-of-block feedback on metacognitive bias, we conducted a GLMM on block-averaged metacognitive bias with a Normal distribution of the response variable and an Identity link function, with z-scored previous block-averaged accuracy and previous block-averaged confidence as within-participant predictors:

$$\text{block metacognitive bias} \sim \text{z(prev. block accuracy)} + \text{z(prev. block confidence)} + (1 \mid \text{participant})$$

Second, to investigate participants' metacognitive sensitivity, we estimated participant-specific type-2 area under the receiver operating curve (AUROC2, Maniscalco & Lau, 2012). To calculate the AUROC2, confidence ratings were first derived from the near-continuous 201-point confidence scale were binned into 3 quantiles for low, medium and high confidence levels. Then, the conditional probabilities P(confidence = y | incorrect) and P(confidence = y | correct) are calculated for each confidence level. Cumulating these conditional probabilities and plotting them against each other produces the type 2 ROC function. A type 2 ROC that bows sharply upwards indicates that confidence has a high degree of sensitivity to correct/incorrect decisions (good metacognitive sensitivity) (**Figure S3**). The AUROC2 itself is a useful non-parametric measure of metacognitive sensitivity but is expected to be affected by participants' performance. To minimize this potential confound, we explicitly modeled the connection between performance and metacognition by calculating the metacognitive efficiency meta-d'/d', or "M-ratio", which quantifies metacognitive sensitivity (meta-d') relative to task performance (d'). To do so, we used the Hierarchical Bayesian modelling within the HMeta-d toolbox for inference on these parameters at the group level (Fleming, 2017), allowing direct domain/session comparisons while avoiding reliance on noisy point estimates of single-participant parameters. Certainty on these parameters (the group-level M-ratio) was determined by computing the 95% highest-density interval (HDI) from the posterior samples. We report the posterior density and average of M-ratio, posterior density and 95% High Density Intervals (HDI) on M-ratio difference and correlation between domains.

Finally, to assess the stability of local and/or global metrics over time, we conducted test-retest analyses across the two sessions using Intraclass Correlation Coefficients (ICC) estimates and their 95% confident intervals, calculated based on a mean-rating (k = 120 trials maximum, after trial exclusion), absolute-agreement, 2-way mixed-effects model (Koo & Li, 2016; **Figure 5; Figure S4**). We report estimates, 95% confidence intervals, *F* statistics, and *p*-values. To compare between-domains ICC and between-sessions ICC, we used a bootstrap approach with replacement to estimate confidence intervals. Bootstrapped ICC differences were estimated on 10,000 samples of metacognitive bias values in each condition. We report the 95% confidence interval of the difference in ICC. Finally, following the approach from Lund et al., 2025, we did not use single-measurement M-ratios to assess the test-retest stability of the metacognitive efficiency, but instead the 95% HDI on the posterior distribution of the correlation coefficient of M-ratios between sessions.

All statistical analyses were performed using MATLAB (www.mathworks.com).

Data and code availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the **Supplementary Materials** are available from the repository:
https://github.com/sophiebavard/metaCog_domains_time. All custom scripts are available.

Acknowledgements

SB was supported by the French National Research Agency (ANR-23-CE37-0018). The research leading to these results has received funding from the national program “Investissements d’avenir” ANR-10- IAIHU-0006. We thank DamnFine Ltd. The salary of AM was paid by a Wellcome Trust grant 203376/2/16/Z during the development of the experiment.

Conflicts of interest

The authors have declared that no competing interests exist.

Author contributions

Conceptualization: S.F., A.M., K.N., M.R.;

Data Collection: F.C.;

Methodology: S.F., A.M., M.R.;

Formal Analysis: S.B., F.C., M.R.;

Investigation: S.B., F.C., S.F., A.M., K.N., M.R.;

Writing – Original Draft: S.B., M.R.;

Writing – Review & Editing: F.C., S.F., A.M., K.N.;

Supervision: S.F., M.R.;

Funding Acquisition: S.F., M.R.

References

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377‑386. https://doi.org/10.1016/j.cognition.2015.10.006

Bandura, A. (1977). Self-efficacy : Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191‑215. https://doi.org/10.1037/0033-295X.84.2.191

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods : A Simulation Study. *Frontiers in Psychology*, *12*, 675558. https://doi.org/10.3389/fpsyg.2021.675558

Binnendyk, J., Li, S., Costello, T., Hale, R., Moore, D. A., & Pennycook, G. (2024). *Is Overconfidence a Trait? An Adversarial Collaboration*. OSF. https://doi.org/10.31234/osf.io/awugz

Elliott, R., Sahakian, B. J., McKay, A. P., Herrod, J. J., Robbins, T. W., & Paykel, E. S. (1996). Neuropsychological impairments in unipolar depression : The influence of perceived failure on subsequent performance. *Psychological Medicine*, *26*(5), 975‑989. https://doi.org/10.1017/s0033291700035303

Fleming, S. M. (2017). HMeta-d : Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M. (2024). Metacognition and Confidence : A Review and Synthesis. *Annual Review of Psychology*, *75*(Volume 75, 2024), 241‑268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making : A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91‑114. https://doi.org/10.1037/rev0000045

588 Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition : Computation, biology and function.
589 *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280‑ 1286.
590 https://doi.org/10.1098/rstb.2012.0021
591 Fox, C. A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R. J., Hanlon, A. K., Gallagher, E.,
592 Rouault, M., & Gillan, C. M. (2024). Reliable, rapid, and remote measurement of
593 metacognitive bias. *Scientific Reports*, *14*(1), 14941. https://doi.org/10.1038/s41598-024-
594 64900-0
595 García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes : Asymptotic and small-
596 sample properties. *Vision Research*, *38*(12), 1861‑ 1881. https://doi.org/10.1016/s0042-
597 6989(97)00340-4
598 Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the
599 process of decision formation during perceptual choices. *NeuroImage*, *106*, 134‑ 143.
600 https://doi.org/10.1016/j.neuroimage.2014.11.036
601 Haddara, N., & Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-Making and
602 Metacognition : Reduction in Bias but No Change in Sensitivity. *Psychological Science*, *33*(2),
603 259‑ 275. https://doi.org/10.1177/09567976211032887
604 Heilbron, M., & Meyniel, F. (2019). Confidence resets reveal hierarchical adaptive learning in
605 humans. *PLOS Computational Biology*, *15*(4), e1006972.
606 https://doi.org/10.1371/journal.pcbi.1006972
607 Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd better think twice : Post-decision perceptual
608 confidence. *NeuroImage*, *99*, 323‑ 331. https://doi.org/10.1016/j.neuroimage.2014.05.049
609 Katyal, S., Huys, Q. J., Dolan, R. J., & Fleming, S. M. (2025). Distorted learning from local
610 metacognition supports transdiagnostic underconfidence. *Nature Communications*, *16*(1),
611 1854. https://doi.org/10.1038/s41467-025-57040-0

612 Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in
613 humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*,
614 *367*(1594), 1322‑1337. https://doi.org/10.1098/rstb.2012.0037
615 Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and
616 decision time. *Neuron*, *84*(6), 1329‑1342. https://doi.org/10.1016/j.neuron.2014.12.015
617 Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation
618 Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155‑163.
619 https://doi.org/10.1016/j.jcm.2016.02.012
620 Lee, A. L. F., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin*
621 *& Review*, *28*(4), 1233‑1242. https://doi.org/10.3758/s13423-020-01869-7
622 Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical*
623 *Society of America*, *49*(2B), 467‑477. https://doi.org/10.1121/1.1912375
624 Lund, A. E., Corrêa, C. M. C., Fardo, F., Fleming, S. M., & Allen, M. G. (2025). Domain generality in
625 metacognitive ability : A confirmatory study across visual perception, episodic memory, and
626 semantic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.
627 https://doi.org/10.1037/xlm0001462
628 Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive
629 sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422‑430.
630 https://doi.org/10.1016/j.concog.2011.09.021
631 Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for
632 metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal*
633 *of Experimental Psychology: General*, *149*(9), 1788‑1799.
634 https://doi.org/10.1037/xge0000746
635 Mazancieux, A., Pereira, M., Faivre, N., Mamassian, P., Moulin, C. J., & Souchay, C. (2023). Towards a
636 common conceptual space for metacognition in perception and memory. *Nature Reviews*
637 *Psychology*, *2*(12), 751‑766.

638 McWilliams, A., Bibby, H., Steinbeis, N., David, A. S., & Fleming, S. M. (2023). Age-related decreases
639 in global metacognition are independent of local metacognition and task performance.
640 *Cognition*, *235*, 105389. https://doi.org/10.1016/j.cognition.2023.105389
641 Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition : Knowing about knowing* (p. xiii, 334). The
642 MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001
643 Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and Domain-Specific Patterns of
644 Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal of Neuroscience:*
645 *The Official Journal of the Society for Neuroscience*, *38*(14), 3534‑3546.
646 https://doi.org/10.1523/JNEUROSCI.2360-17.2018
647 Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation
648 persists after choice to inform metacognitive judgments. *eLife*, *4*, e11946.
649 https://doi.org/10.7554/eLife.11946
650 Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low self-esteem prospectively predicts depression
651 in adolescence and young adulthood. *Journal of Personality and Social Psychology*, *95*(3),
652 695‑708. https://doi.org/10.1037/0022-3514.95.3.695
653 Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency.
654 *Consciousness and Cognition*, *28*, 151‑160. https://doi.org/10.1016/j.concog.2014.06.007
655 Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty : Distinct probabilistic
656 quantities for different goals. *Nature Neuroscience*, *19*(3), 366‑374.
657 https://doi.org/10.1038/nn.4240
658 Purcell, B. A., & Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales
659 underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*,
660 *113*(31), E4531‑E4540. https://doi.org/10.1073/pnas.1524685113
661 Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal
662 detection task. *Psychology and Aging*, *16*(2), 323‑341.

663 Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676‑ 688.
664 https://doi.org/10.1016/j.tics.2016.07.002
665 Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from
666 local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-
667 09075-3
668 Rouault, M., & Fleming, S. M. (2020). Formation of global self-beliefs in the human brain.
669 *Proceedings of the National Academy of Sciences*, *117*(44), 27268‑ 27276.
670 https://doi.org/10.1073/pnas.2003094117
671 Rouault, M., Lebreton, M., & Pessiglione, M. (2023). A shared brain system forming confidence
672 judgment across cognitive domains. *Cerebral Cortex (New York, N.Y.: 1991)*, *33*(4),
673 1426‑ 1439. https://doi.org/10.1093/cercor/bhac146
674 Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across
675 Domains : Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*,
676 *1*, e17. https://doi.org/10.1017/pen.2018.16
677 Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human
678 Sense of Confidence. *Neuron*, *90*(3), 499‑ 506.
679 https://doi.org/10.1016/j.neuron.2016.03.025
680 Seow, T. X. F., Fleming, S. M., & Hauser, T. U. (2025). Metacognitive biases in anxiety-depression and
681 compulsivity extend across perception and memory. *PLOS Mental Health*, *2*(3), e0000259.
682 https://doi.org/10.1371/journal.pmen.0000259
683 Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global
684 Metacognition Shape Mental Health. *Biological Psychiatry*, *90*(7), 436‑ 446.
685 https://doi.org/10.1016/j.biopsych.2021.05.013
686 Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation.
687 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204‑ 221.
688 https://doi.org/10.1037/0278-7393.26.1.204

689 Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual
690 perception. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(5),
691 671‐682. https://doi.org/10.1037/xhp0000179
692 van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016).
693 A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*,
694 e12192. https://doi.org/10.7554/eLife.12192
695 Van Marcke, H., Denmat, P. L., Verguts, T., & Desender, K. (2024). Manipulating Prior Beliefs Causally
696 Induces Under- and Overconfidence. *Psychological Science*, *35*(4), 358‐375.
697 https://doi.org/10.1177/09567976241231572
698 Van Selst, M., & Jolicoeur, P. (1994). A Solution to the Effect of Sample Size on Outlier Elimination.
699 *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631‐650.
700 https://doi.org/10.1080/14640749408401131
701 West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence :
702 Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*(3),
703 387‐392. https://doi.org/10.3758/BF03210798
704 Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases
705 performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*(5), 7.
706 https://doi.org/10.1167/8.5.7
707 Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. S. (2016). Self-
708 Other Mergence in the Frontal Cortex during Cooperation and Competition. *Neuron*, *91*(2),
709 482‐493. https://doi.org/10.1016/j.neuron.2016.06.022
710 Zacharopoulos, G., Binetti, N., Walsh, V., & Kanai, R. (2014). The Effect of Self-Efficacy on Visual
711 Discrimination Sensitivity. *PLOS ONE*, *9*(10), e109392.
712 https://doi.org/10.1371/journal.pone.0109392
713 Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual
714 decision. *Frontiers in Integrative Neuroscience*, *6*. https://doi.org/10.3389/fnint.2012.00079

Supplementary Materials

Supplementary Results: Bridging domain, global, and local confidence to assess a hierarchical architecture of metacognitive judgements

To assess the stability of all our metrics over time, we conducted test-retest analyses across the two sessions using intraclass correlation coefficients (ICC, **Table S7-S8**; Koo & Li, 2016). We first looked at the test-retest reliability of the factors contributing to global self-performance estimates. To do this, we performed a generalized linear model (GLM), separately for each participant, predicting category choice as a function of local difficulty, accuracy, response time (RT), and confidence. We note that, since participants performed one category choice at the end of each of the six blocks, these GLMs are performed on a lower number of data points compared to group-level GLMMs. Despite this caveat, this approach yields one coefficient per predictor per participant per domain per session. We then estimated the test-retest reliability of these coefficients using ICC to examine whether the same factors contributed to self-performance estimates between test and retest sessions. Overall, we found low reliability across all measures (**Figure S4A; Table S7**), except for the contribution of local confidence to global confidence in the perceptual domain ($\mathrm{ICC} = 0.28, 95\%\ \mathrm{CI} = [-0.0083, 0.53]$). As noted previously, a limiting factor in establishing the reliability of these estimates is statistical power with N=44 participants in the retest session, given that each session contained six blocks.

We then turned to the test-retest reliability of local metrics (**Figure S4B; Table S8**). For metacognitive efficiency, we used the 95% HDI on the posterior distribution of the correlation coefficient between metacognitive efficiencies in the two sessions (see **Methods**; Fleming, 2017; Lund et al., 2025). First, these analyses revealed a very high test-retest reliability for local confidence (memory: $\mathrm{ICC} = 0.87, 95\%\ \mathrm{CI} = [0.77, 0.93]$; perception: $\mathrm{ICC} = 0.91, 95\%\ \mathrm{CI} = [0.83, 0.95]$) and metacognitive bias (memory: $\mathrm{ICC} = 0.81, 95\%\ \mathrm{CI} = [0.65, 0.90]$; perception: $\mathrm{ICC} = 0.89, 95\%\ \mathrm{CI} = [0.80, 0.94]$). In contrast, metacognitive efficiency showed no test-retest reliability (memory: $95\%\ \mathrm{HDI} = [-0.98, 0.182]$; perception: $95\%\ \mathrm{HDI} = [-0.46, 0.99]$). We found no evidence for better calibration in the retest session, where participants were just as overconfident as in their test session ($t(43) = 1.6, p = 0.11$). Consistent with our previous analyses, these results support the idea that the degree of overconfidence may be a stable, domain-general, individual trait.

Supplementary Results: Age-related associations in local metrics

Finally, even though the study was not designed or powered to examine age effects (N=52), based on the identified relationships between metacognition and age in previous work (N>300) (McWilliams et al., 2023), we investigated here the associations between age (age range: 20-73 years, mean ± SD: 34.4±11.4) and difficulty, accuracy, RT and local confidence, as well as metacognitive bias and metacognitive efficiency (**Figure S6**) in each domain. We only found significant positive correlations between age and RT (memory: Spearman's $\rho(50) = 0.29, p = 0.037$; perception: Spearman's $\rho(50) = 0.32, p = 0.020$), in line with previous work (Ratcliff et al., 2001). No other measure correlated with age in our sample (all Spearman's $|\rho(50)| < 0.24$, all $p > 0.090$). However, we note that our

753 study was not optimized for age-related analyses, which could explain why we do not observe the
754 previously documented decrease in metacognitive bias (McWilliams et al., 2023) and in (perceptual)
755 metacognitive efficiency (Palmer et al., 2014) with age, presumably due to lack of statistical power.

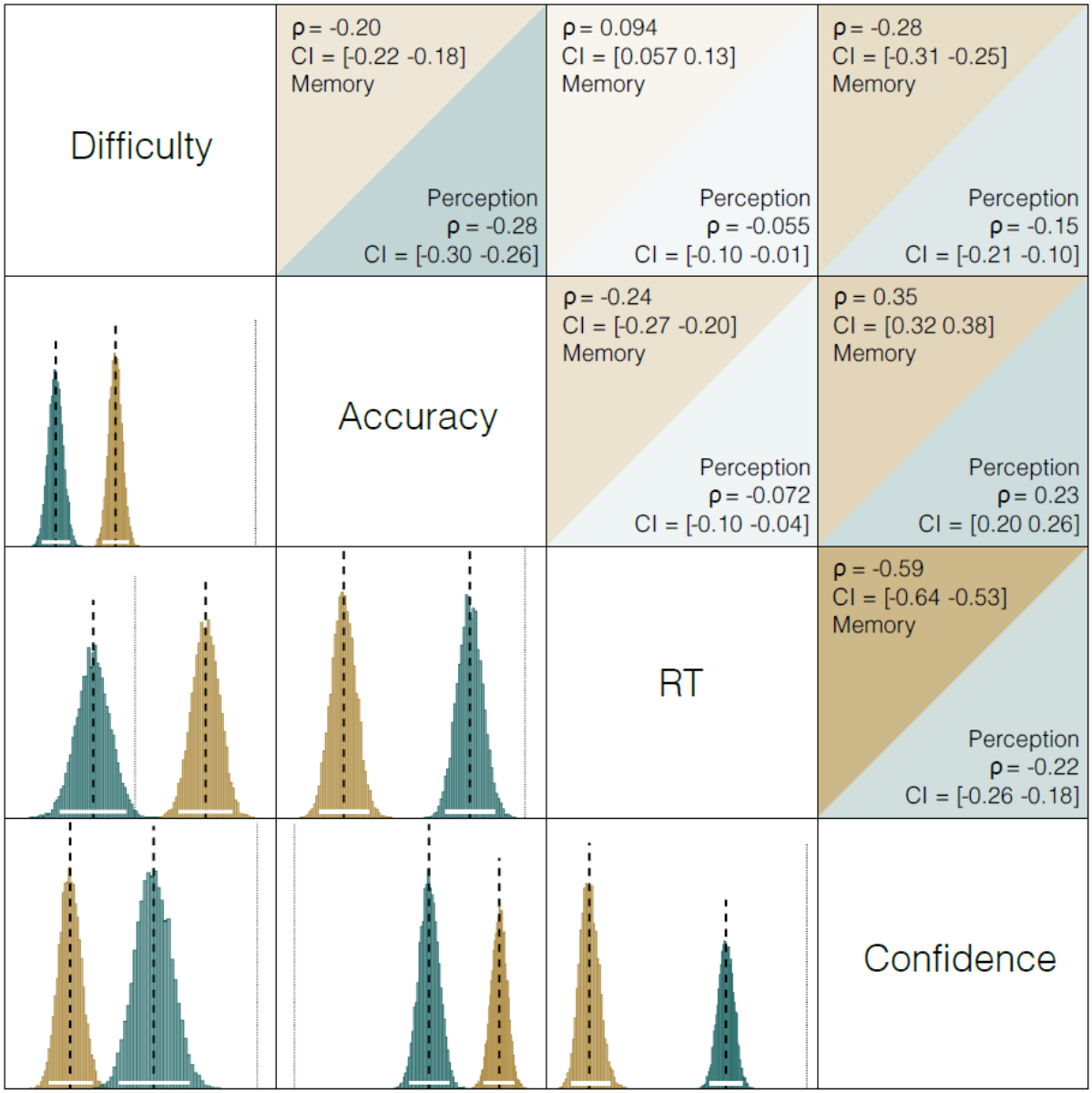


Figure S1. Within-participant correlations for local metrics in the test session. Distributions in the bottom-left half represent bootstrapped Spearman's correlation coefficients between each pair of local variables in memory (brown) and perception (green). The thin dashed line represents zero, the bold dashed line represents the average coefficient, and the horizontal white line represents 95% CI which all exclude zero, indicating significant group-level trends toward positive or negative correlations between our variables (see upper-right half for numerical values). N=52.

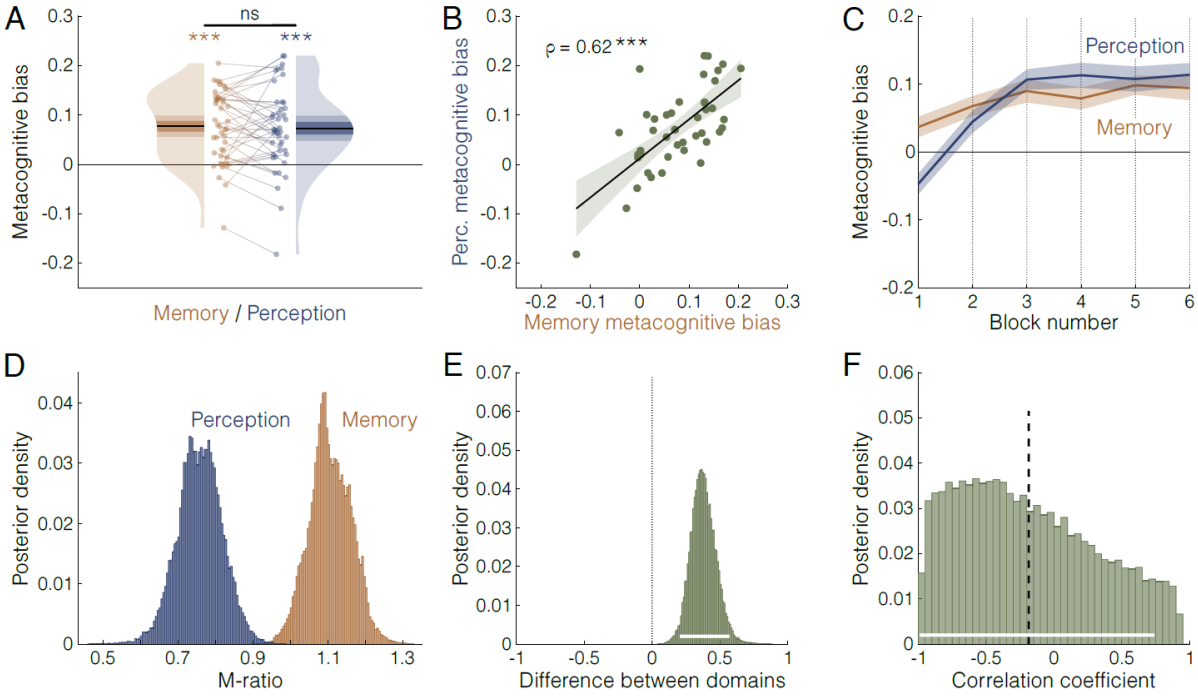
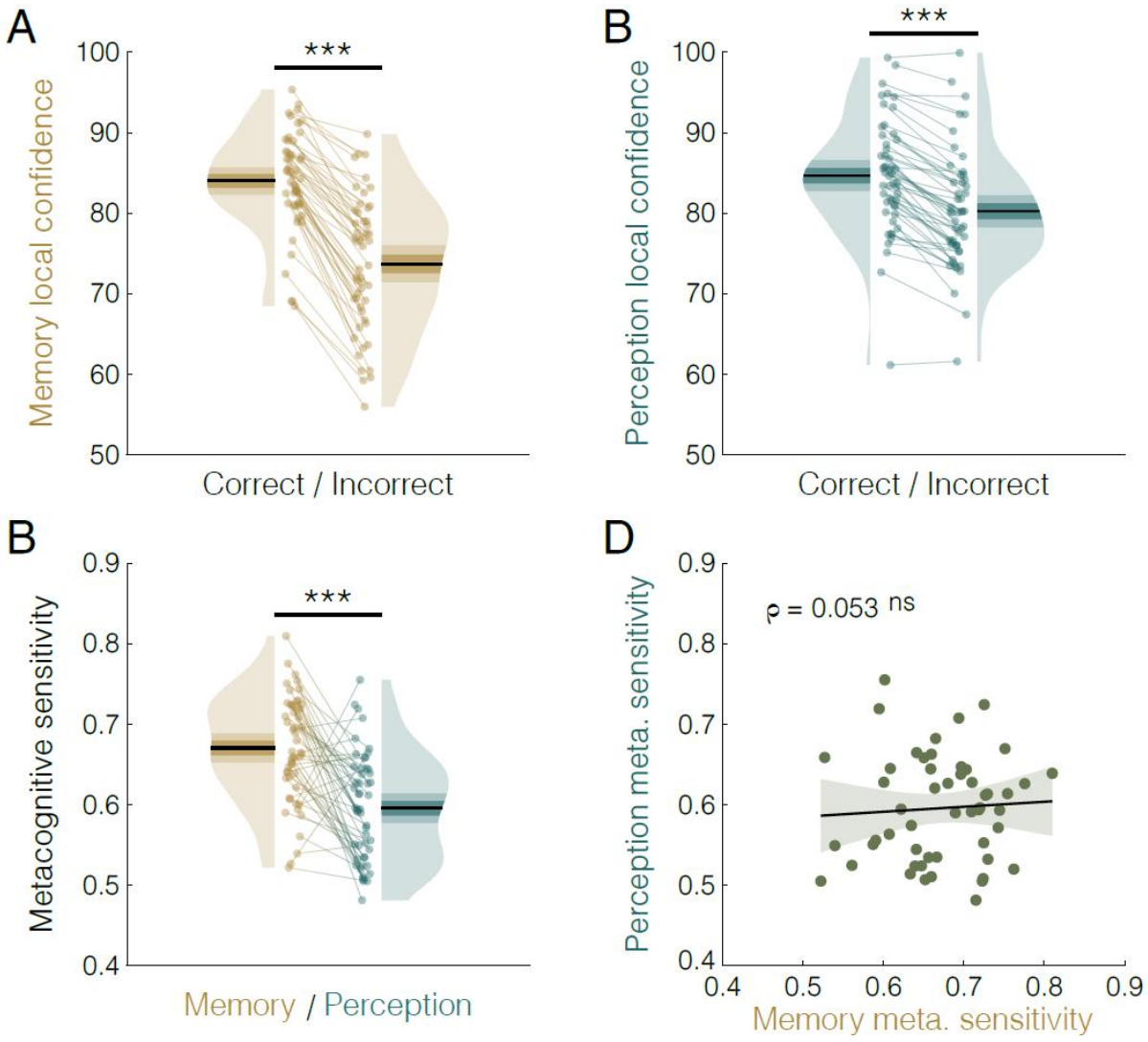


Figure S2. Between-domain comparison of metacognitive bias and metacognitive efficiency in the retest session. A. Average metacognitive bias across participants in memory (orange) and perception (blue) tasks. Points and grey lines indicate individual average, shaded areas indicate probability density function, 95% confidence interval, and SE. N=44. **B.** Scatterplot of metacognitive bias between domains indicating a significant positive correlation (p < 0.0001). Shaded area represents 95% confidence intervals and black line is a linear regression fit. N=44. **C**. Trajectories of metacognitive bias measured as the discrepancy between local confidence and accuracy throughout the six blocks in memory (orange) and perception (blue) domains. Shaded areas indicate SE over participants. N=44. **D.** Group-level metacognitive efficiency (M-ratio) distribution estimated hierarchically in the memory (orange) and perception (blue) domains (see Methods). **E.** Group-level difference (in log units) between the group posteriors. The white bar represents the 95% HDI which excludes zero (dotted line), indicating a significantly higher metacognitive efficiency in the memory than the perception domain. **F.** Group-level correlation coefficient (dotted line) between metacognitive efficiencies in the two domains. The white bar represents the 95% HDI which includes zero, indicating no correlation between domains.



777

778 **Figure S3**. **A, B.** Local confidence as a function of correct vs. incorrect decisions, reflecting metacognitive
779 abilities in the memory ($t(51)=17$, $p<0.0001$) **(A)** and perception ($t(51)=13$, $p<0.0001$) **(B)** domains respectively.
780 Points and grey lines indicate individual averages, shaded areas indicate probability density function, 95%
781 confidence interval, and SE. Stars indicate statistical significance of a paired t-test between correct and incorrect
782 decisions. **C.** Between-domain comparison of metacognitive sensitivity calculated with AUROC2 (see Methods),
783 indicating domain-specificity (between-domain t-test: $t(51)=6.0$, $p<0.0001$). **D.** Scatterplot of metacognitive
784 sensitivity between domains indicating no significant correlation ($\rho(50)=0.053$, $p=0.71$). Shaded area represents
785 95% confidence intervals and black line is a linear regression fit. In all panels, ns $p>0.05$, *** $p<0.0001$, $N=52$.

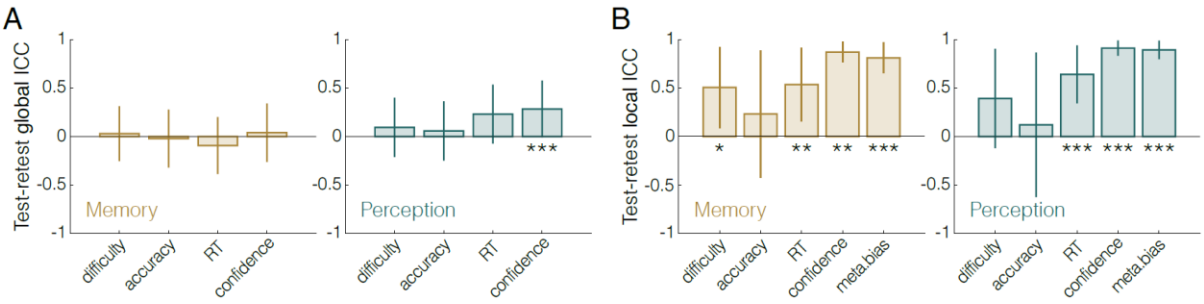


Figure S4. Intraclass correlation coefficients. A. Intraclass correlation coefficients for regression weights of factors contributing to self-performance estimates between test and retest sessions, in the memory (left, yellow) and perception (right, green) domains (see Methods). **B**. Intraclass correlation coefficients for local indices between test and retest sessions, in the memory (left, yellow) and perception (right, green) domains. Error bars are 95% CI.

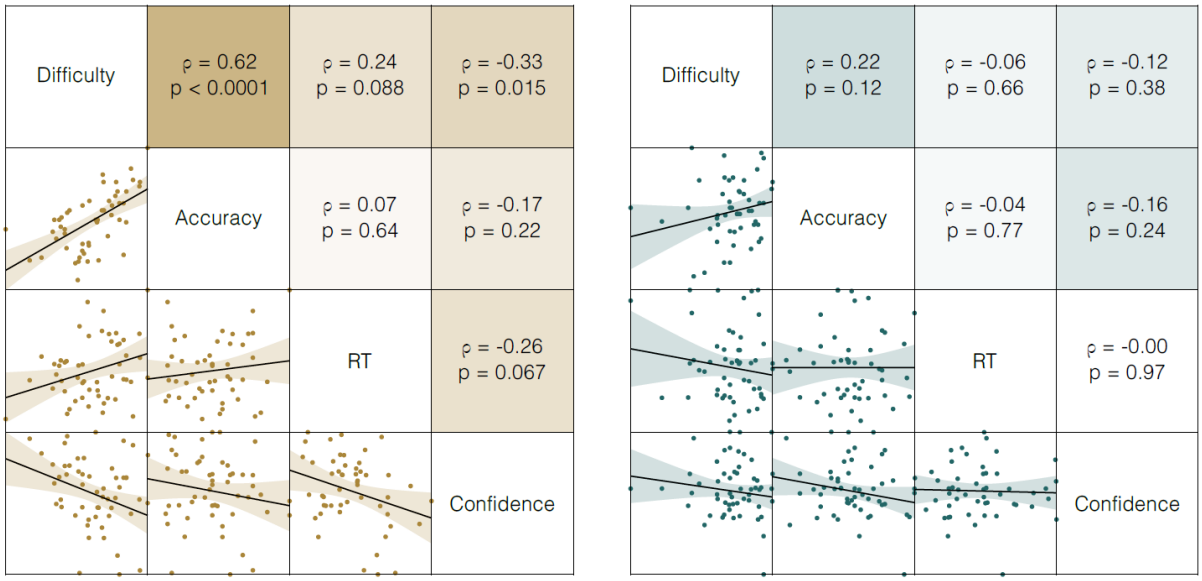
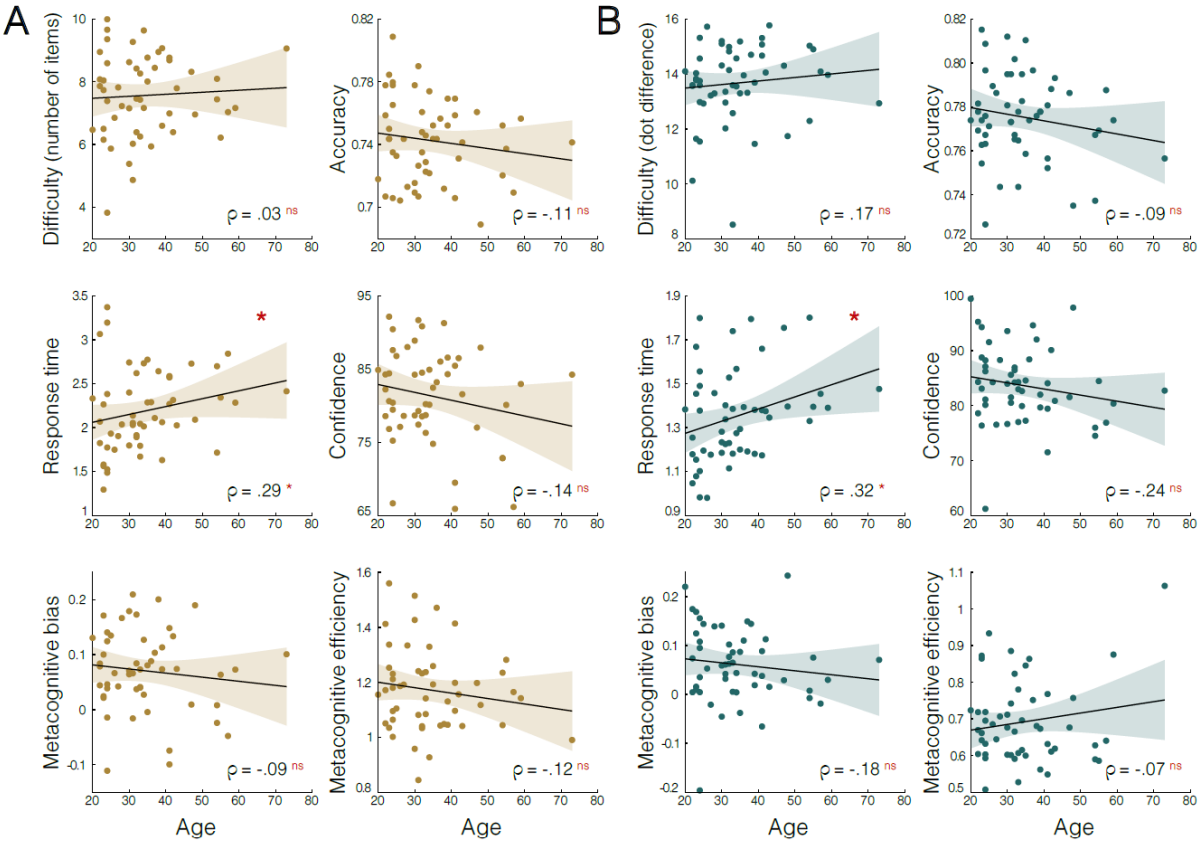


Figure S5. Between-participant correlations for local metrics in the test session. Scatter plots in the bottom-left corner represent pairwise correlations between local variables in memory (top panel, yellow) and perception (bottom panel, green). Shaded areas represent 95% confidence intervals on the predicted values from a linear regression fit. The upper-right halves indicate Spearman's coefficients and associated p-values. N=52.



797

798 **Figure S6. Correlations between age and local measures of metacognition** in memory (**A**) and perception
799 (**B**). Each dot in a participant (N=52). Shaded areas represent 95% confidence intervals on the predicted values
800 from a linear regression fit. *p<0.05, uncorrected for multiple testing. All others were not statistically significant.

	Memory						Perception					
Predictor	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**
Intercept	-0.00	0.026	-0.064	-0.052	0.049	0.95	0.010	0.026	0.38	-0.041	0.060	0.70
Difficulty	0.043	0.027	1.58	-0.010	0.097	0.11	0.00	0.027	0.13	-0.049	0.057	0.90
Accuracy	**0.075**	0.028	2.67	0.020	0.13	0.0075 **	0.00	0.027	0.076	-0.051	0.056	0.94
RT	0.039	0.032	1.21	-0.024	0.10	0.23	-0.046	0.027	-1.71	-0.098	0.0068	0.088
Confidence	**0.23**	0.034	6.84	0.17	0.30	<0.0001 ***	**0.095**	0.028	3.42	0.040	0.15	0.00064 ***

Table S1. Results from main GLMM fitted on category choice in the test session. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT and confidence as independent variables. All VIFs<1.8 indicating no multicollinearity issues, ****p*<0.001, ***p*<0.01, **p*<0.05.

	Memory						Perception					
Predictor	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**
Intercept	-0.0015	0.028	-0.054	-0.056	0.054	0.96	0.0019	0.028	0.068	-0.053	0.057	0.95
Difficulty	0.058	0.029	2.0	0.00	0.12	0.050	0.0046	0.030	0.15	-0.053	0.063	0.88
Accuracy	**0.063**	0.030	2.1	0.0039	0.12	0.039 *	-0.011	0.030	-0.37	-0.069	0.047	0.71
RT	0.0037	0.034	0.11	-0.064	0.071	0.92	-0.029	0.029	-1.0	-0.085	0.028	0.32
Confidence	**0.24**	0.036	6.7	0.17	0.31	<0.0001 ***	**0.076**	0.030	2.5	0.012	0.14	0.012 *

Table S2. Results from main GLMM fitted on category choice in the retest session. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT and confidence as independent variables. All VIFs<1.7 indicating no multicollinearity issues, ****p*<0.001, ***p*<0.01, **p*<0.05.

	Memory						Perception					
Predictor	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**
Intercept	-0.035	0.032	-1.11	-0.10	0.027	0.27	-0.019	0.030	-0.65	-0.078	0.039	0.52
Difficulty	0.044	0.033	1.34	-0.020	0.11	0.18	0.018	0.033	0.55	-0.046	0.082	0.58
Accuracy	**0.11**	0.034	3.14	0.040	0.17	0.0017 **	0.047	0.036	1.32	-0.023	0.12	0.19
RT	0.043	0.035	1.23	-0.025	0.11	0.22	-0.054	0.030	-1.81	-0.11	0.005	0.071
Confidence	**0.22**	0.036	6.07	0.15	0.29	<0.0001 ***	**0.11**	0.030	3.70	0.052	0.17	0.00022 ***
diff:acc	-0.010	0.035	-0.27	-0.078	0.059	0.79	-0.033	0.044	-0.75	-0.12	0.053	0.45
diff:RT	0.060	0.034	1.76	-0.007	0.13	0.079	0.029	0.031	0.94	-0.032	0.091	0.35
diff:conf	0.060	0.035	1.70	-0.009	0.13	0.088	0.005	0.030	0.17	-0.053	0.063	0.86
acc:RT	-0.025	0.037	-0.69	-0.10	0.047	0.49	-0.048	0.033	-1.44	-0.11	0.017	0.15
acc:conf	0.072	0.038	1.89	-0.003	0.15	0.060	0.063	0.032	1.93	-0.001	0.13	0.054
RT:conf	-0.026	0.029	-0.90	-0.083	0.031	0.37	-0.025	0.027	-0.91	-0.079	0.029	0.36
diff:acc:RT	-0.003	0.036	-0.093	-0.074	0.068	0.93	0.028	0.042	0.67	-0.054	0.11	0.50
diff:acc:conf	-0.019	0.039	-0.49	-0.10	0.058	0.63	0.013	0.039	0.34	-0.063	0.089	0.74
diff:RT:conf	-0.015	0.028	-0.55	-0.070	0.040	0.59	0.034	0.023	1.47	-0.011	0.080	0.14
acc:RT:conf	-0.014	0.032	-0.43	-0.076	0.048	0.66	0.018	0.027	0.68	-0.034	0.071	0.49
diff:acc:RT:conf	-0.013	0.030	-0.44	-0.073	0.046	0.66	0.011	0.030	0.37	-0.047	0.069	0.71

Table S3. Results from complete GLMM fitted on category choice in the test session. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT, confidence, and their interactions as independent variables. All VIFs<2.5 indicating no multicollinearity issues, ****p*<0.001, ***p*<0.01, **p*<0.05.

	Memory						Perception					
Predictor	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**	**Estimate**	**SE**	***t*-value**	**Lower**	**Upper**	***p*-value**
Intercept	-0.028	0.035	-0.80	-0.10	0.040	0.42	-0.0012	0.032	-0.037	-0.064	0.062	0.97
Difficulty	**0.081**	0.036	2.3	0.011	0.15	0.023 *	0.032	0.036	0.89	-0.038	0.10	0.37
Accuracy	**0.086**	0.036	2.4	0.015	0.16	0.018 *	-0.022	0.038	-0.59	-0.10	0.052	0.55
RT	-0.016	0.037	-0.42	-0.089	0.058	0.68	-0.042	0.032	-1.3	-0.11	0.020	0.18
Confidence	**0.246**	0.038	6.4	0.17	0.32	<0.0001 ***	0.059	0.032	1.8	-0.0042	0.12	0.067
diff:acc	-0.020	0.037	-0.53	-0.093	0.053	0.60	0.035	0.046	0.76	-0.055	0.13	0.45
diff:RT	-0.039	0.038	-1.0	-0.11	0.035	0.30	-0.00016	0.035	-0.0048	-0.068	0.068	1.00
diff:conf	-0.0030	0.040	-0.077	-0.081	0.075	0.94	0.011	0.036	0.31	-0.059	0.081	0.75
acc:RT	-0.022	0.036	-0.61	-0.093	0.049	0.54	-0.030	0.031	-0.96	-0.092	0.032	0.34
acc:conf	-0.0066	0.040	-0.16	-0.085	0.072	0.87	0.014	0.034	0.42	-0.052	0.080	0.67
RT:conf	-0.034	0.032	-1.07	-0.10	0.028	0.28	0.0023	0.029	0.08	-0.054	0.058	0.93
diff:acc:RT	-0.064	0.042	-1.53	-0.15	0.018	0.13	-0.054	0.042	-1.3	-0.14	0.029	0.20
diff:acc:conf	-0.012	0.040	-0.29	-0.090	0.067	0.77	**-0.081**	0.038	-2.2	-0.16	-0.0077	0.031
diff:RT:conf	0.051	0.031	1.6	-0.011	0.11	0.11	0.028	0.026	1.1	-0.023	0.078	0.28
acc:RT:conf	0.046	0.034	1.4	-0.020	0.11	0.17	0.012	0.026	0.48	-0.039	0.064	0.63
diff:acc:RT:conf	0.012	0.035	0.35	-0.056	0.080	0.73	-0.015	0.030	-0.50	-0.073	0.043	0.62

Table S4. Results from complete GLMM fitted on category choice in the retest session. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT, confidence, and their interactions as independent variables. All VIFs<2.1 indicating no multicollinearity issues, ****p*<0.001, ***p*<0.01, **p*<0.05.

Predictor	Estimate	SE	*t*-value	Lower	Upper	*p*-value
Intercept	0.0034	0.014	0.24	-0.024	0.031	0.81
Difficulty	0.026	0.015	1.8	-0.0029	0.055	0.077
Accuracy	**0.031**	0.015	2.1	0.0019	0.061	0.037 *
RT	-0.0086	0.016	-0.54	-0.040	0.023	0.59
Confidence	**0.16**	0.017	9.5	0.13	0.19	<0.0001 ***
Domain	-0.0032	0.014	-0.23	-0.031	0.024	0.82
domain:diff	0.021	0.015	1.4	-0.0081	0.050	0.16
domain:acc	**0.041**	0.015	2.8	0.012	0.071	0.0058 **
domain:rt	0.026	0.016	1.6	-0.0055	0.057	0.11
domain:conf	**0.075**	0.017	4.5	0.042	0.11	<0.0001 ***

Table S5. Results from GLMM with domain-interactions fitted on category choice. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category, both sessions pooled, N=88) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT, confidence, and their interactions with domain (memory vs. perception) as independent variables. All VIFs<1.5 indicating no multicollinearity issues, ***p<0.001, **p<0.01, *p<0.05, °p<0.1.

Predictor	Estimate	SE	*t*-value	Lower	Upper	*p*-value
Intercept	0.0035	0.014	0.25	-0.024	0.031	0.80
Difficulty	0.025	0.015	1.7	-0.0041	0.054	0.092
Accuracy	**0.033**	0.015	2.2	0.0040	0.063	0.026 *
RT	-0.028	0.015	-1.8	-0.058	0.00	0.076
Confidence	**0.15**	0.016	9.1	0.12	0.18	<0.0001 ***
Session	0.0033	0.014	0.23	-0.024	0.031	0.82
session:diff	-0.0055	0.015	-0.37	-0.034	0.023	0.71
session:acc	0.0053	0.015	0.36	-0.024	0.035	0.72
session:rt	0.0046	0.015	0.30	-0.026	0.035	0.77
session:conf	0.00036	0.016	0.022	-0.032	0.032	0.98

Table S6. Results from GLMM with session-interactions fitted on category choice. The generalized linear mixed model (GLMM) with a Binomial distribution of the response variable (chosen/unchosen category, both domains pooled, N=88) and a Logit link function was fitted on end-of-block category choice, with z-scored local difficulty, accuracy, RT, confidence, and their interactions with session (test vs. retest) as independent variables. All VIFs<1.5 indicating no multicollinearity issues, ***p<0.001, **p<0.01, *p<0.05, °p<0.1.

845

846

	Memory							Perception						
	Interclass Correlation	95% Confidence Interval		F Test with true value 0				Interclass Correlation	95% Confidence Interval		F Test with true value 0			
		Lower Bound	Upper Bound	Value	df1	df2	*p*-value		Lower Bound	Upper Bound	Value	df1	df2	*p*-value
Difficulty	0.029	-0.25	0.31	1.06	43	43.27	0.42	0.09	-0.21	0.38	1.20	43	43.04	0.27
Accuracy	-0.021	-0.32	0.28	0.96	43	42.98	0.55	0.06	-0.25	0.35	1.12	43	43	0.36
RT	-0.093	-0.39	0.21	0.83	43	42.99	0.72	0.23	-0.073	0.49	1.59	43	43.04	0.07
Confidence	0.040	-0.26	0.33	1.08	43	43.07	0.40	**0.28**	-0.0083	0.53	1.80	43	43.74	0.028 *

847

848 **Table S7. ICC estimates for regression weights of factors contributing to self-performance estimates**
849 **between test and retest sessions.** Interclass Correlation Coefficients (ICC) and their 95% confident intervals
850 were calculated based on a single measurement, absolute-agreement, 2-way mixed-effects model. *p<0.05.

	Memory							Perception						
	Interclass Correlation	95% Confidence Interval		F Test with true value 0				Interclass Correlation	95% Confidence Interval		F Test with true value 0			
		Lower Bound	Upper Bound	Value	df1	df2	*p*-value		Lower Bound	Upper Bound	Value	df1	df2	*p*-value
Difficulty	**0.51**	0.089	0.73	2.01	43	43.3	0.012 *	0.39	-0.12	0.67	1.64	43	43.3	0.055
Accuracy	0.22	-0.42	0.58	1.3	43	43.1	0.198	0.12	-0.63	0.52	1.14	43	43.2	0.34
RT	**0.54**	0.16	0.75	2.17	43	43.9	0.0059 **	**0.64**	0.34	0.80	2.77	43	43.6	0.00054 ***
Confidence	**0.87**	0.77	0.93	7.77	43	44	<0.0001 ***	**0.91**	0.83	0.95	12.2	43	36.3	<0.0001 ***
Meta. bias	**0.81**	0.65	0.90	5.20	43	43.3	<0.0001 ***	**0.89**	0.80	0.94	10.06	43	35.88	<0.0001 ***
Meta. efficiency	-0.074	-0.34	0.21	0.85	43	41.5	0.70	**0.41**	0.091	0.65	2.92	43	19.60	0.0063 **

851

852

853 **Table S8. ICC estimates for local metrics.** Interclass Correlation Coefficients (ICC) and their 95% confident
854 intervals were calculated based on a mean-rating (k=120 trials max.), absolute-agreement, 2-way mixed-effects
855 model. ***p<0.001, **p<0.01, *p<0.05.

856