nature human behaviour

Supplementary information

https://doi.org/10.1038/s41562-026-02443-2

Domain generality is an emergent, not inherent, property of metacognition

In the format provided by the authors and unedited

Supplementary Information for "Domain-Generality is an Emergent, not Inherent, Property of Metacognition"

Here we report several preregistered analyses to complement those presented in the main text.

Experiment 1

Preregistered Accuracy Correlations

We preregistered conducting correlations in Memory and Perceptual task accuracy separately for younger (4-5 years) and older (6-7 years) participants due to stimulus differences, but deviated from our preregistration as we had to exclude 4-year-olds from the sample (see Methods). These correlations are $r(91) = -0.14$, $p = .188$ for 6-7-year-olds, and $r(34) = 0.21$, $p = .213$ for 4-5-year-olds.

Task Accuracy Comparison

Adults were more accurate on Perceptual trials than Memory trials, exploratory paired $t(128) = -15.41$, $p < .001$, $d = 1.71$. Children were slightly more accurate on Perceptual trials than on Memory trials, exploratory paired $t(130), = 2.16$, $p = .033$, $d = 0.27$, and performance was significantly above chance for both tasks, all p's $< .001$, d's > 2.27. Children's accuracy did not change with age, all p's $> .125$.

Correlation of Children's Metacognitive Difference Scores

For children, we preregistered calculating the difference between their average confidence on correct and incorrect trials on each task, as meta-d' models had not been applied to young children's data at the time of preregistration. The average difference in confidence for Memory was 0.38 (SD = 0.30, out of 2), and 0.15 (SD = 0.21) for Perception. These differences were correlated, $r(129) = .27$, $p = .002$, $BF_{10} = 14.60$, see Figure S1. As highlighted in the main text,

we do not think this is the best measurement of metacognitive ability as it conflates metacognitive bias, sensitivity and task performance.

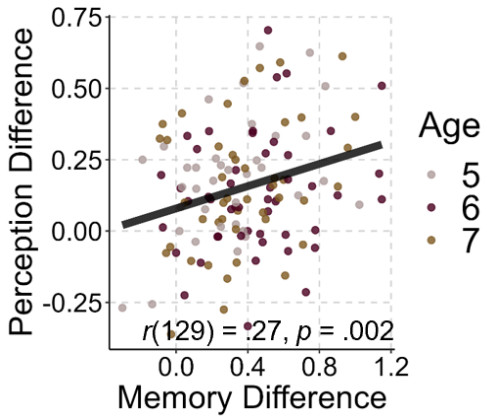

Figure S1: Difference Score Correlation. The correlation between children's (N = 132) difference scores (average confidence on accurate trials minus average confidence on inaccurate trials).

HMeta-d' Modelling

We used a hierarchical Bayesian model to estimate group-level MRatio as a complement to the main analyses. As discussed in Fleming (2017), estimates from the standard meta-d' model are influenced by low numbers of trials (e.g., 32 in each task in our study), and requiring edge-correction is not ideal for getting pure estimates of ability. In contrast, a hierarchical Bayesian model captures uncertainty around each single subject's estimate and accounts for this in the estimate of the group-level statistic. However, other recent work documents a higher false positive rate for HMeta-d' estimates, related to its dependent on the selection of a prior (Rausch & Zehetleitner, 2023). We further found that the HMeta-d' model strongly changed children's data: the large number of negative MRatio values (which occur when a participant uses the

confidence scale incorrectly) are completely absent from the HMeta-d' estimates. Further, the posterior distributions for correlation coefficient ρ were extremely left-skewed. We are therefore skeptical of these model outputs, but present this data for posterity. Our data, analyses, and outputs are on OSF for those interested in exploring this further.

For adults, the HMeta-d' model estimated Memory MRatio as 1.19, 95% HDI[1.12, 1.27], and Perceptual MRatio as 0.5, 95% HDI[0.43, 0.56], very similar to the MRatio estimates from the standard model in the main text. The 95% Highest-Density Interval (HDI) for the correlation overlapped 0, consistent with no correlation between Memory and Perceptual MRatios, ρ = .17, 95% HDI[-0.26, 0.68]. See Figure S2 for a plots of point estimates and the posterior distribution of ρ.

For children, the HMeta-d' model estimated Memory MRatio as 0.66, 95% HDI[0.51, 0.81], and Perceptual MRatio as 0.53, 95% HDI[0.40, 0.69], lower than the MRatio estimates from the standard model in the main text. The 95% Highest-Density Interval (HDI) for the correlation overlapped 0, consistent with no significant correlation between Memory and Perceptual MRatios, ρ = 0.48, 95% HDI[-0.15, 0.97].

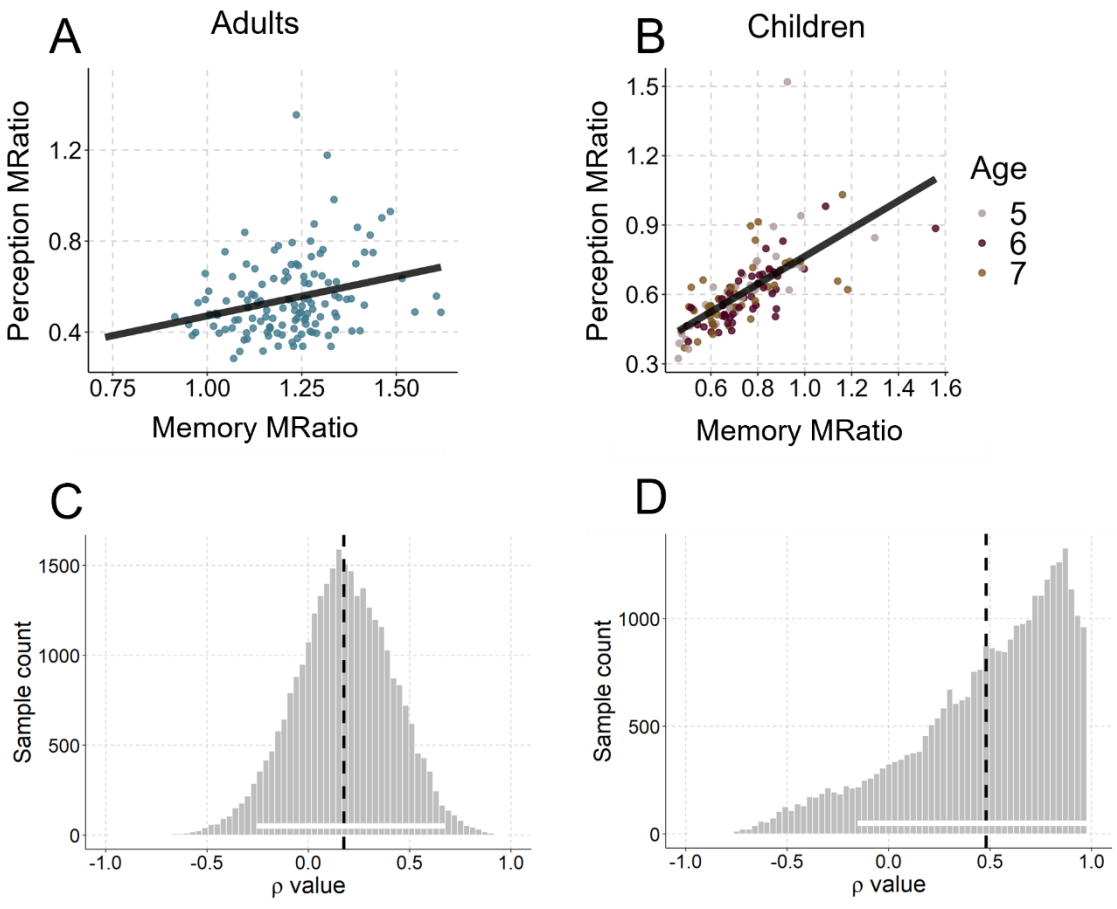

Figure S2. Metacognitive efficiency (MRatio) estimates from the HMeta-d' framework. A) and B) show individual estimates for adults (N = 129) and children (N = 132), respectively. C) and D) show the posterior distribution of the correlation rho. Both distributions overlap 0, suggesting non-significant values.

Robustness Test

There were many children in Experiment 1 who were excluded per our preregistered exclusion criteria. To confirm that our results are robust to these exclusions, we re-ran our analyses with all children who completed the task, including 4-year-olds and those with less than 55% accuracy. We found all the same patterns of correlations as in the main text (see Table S1 below). We also replicated the main effects of accuracy predicting children's confidence scores,

though now also found several significant interactions and main effects with age (see Table S2). As shown in Figure S3, these effects appear driven by the lack of metacognitive differentiation in 4-year-olds.

Table S1: Correlations with fewer exclusion criteria

Measure	Correlation	*p*
Accuracy (controlling for age)	r(207) = .06	.410
Meta-d'	ρ(208) = .13	.063
MRatio	ρ(208) = .08	.263
Bias	r(208) = .33	< .001

Table S2: ANOVA of children's average confidence by their accuracy and age, split by task.

Task	Effect	df	*F*	*p*	Partial η²
Memory	Age	3, 212	1.12	.343	.02
	Correct	1, 212	88.35	< .001	.29
	Age × Correct	3, 212	7.65	< .001	.10
Perception	Age	3, 209	4.30	.006	.06
	Correct	1, 209	36.96	< .001	.15
	Age × Correct	3, 209	2.49	.062	.03

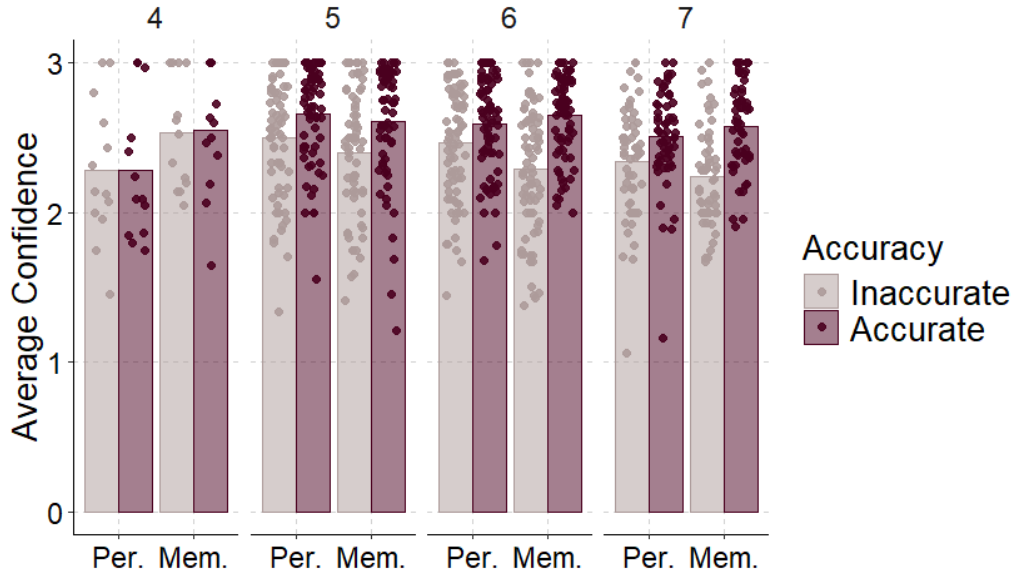

Figure S3. Children's average confidence as predicted by their age, accuracy, and task.

Experiment 2

Memory and Perceptual Accuracy

Adults' Perceptual accuracy was similar to Memory accuracy, paired $t(106) = 1.33$, $p = .180$, $d = 0.17$. Children's accuracy was above chance of 50% for both Perceptual, $t(131) = 18$, $p < .001$, $d = 1.5$, and Memory trials, $t(131) = 12$, $p < .001$, $d = 1.04$, with higher accuracy on the Perceptual task, paired $t(131) = 3.91$, $p < .001$, $d = 0.45$. See Table 1 for descriptive statistics.

Side and Task Preferences

Adults had a slight preference for the right-side answer $M = 53\%$ right side, $SD = 8\%$, one-sample $t(106) = 4.11$, $p < .001$, $d = 0.40$, and a preference for Memory answers, 67% ($SD = 15\%$) of Across-Domain trials, one-sample $t(106) = 11.44$, $p < .001$, $d = 1.11$. Children had no side preference, $M = 51\%$ right side, SD = 16%, one-sample $t(131) = 0.69$, $p = .491$, $d = 0.06$, but a preference for Perceptual answers, 62% (SD = 18%) of Across-Domain trials, one-sample $t(131) = 8.06$, $p < .001$, $d = 0.70$.

Confidence Choices Given the Difference in Expected Accuracies

On half of the trials, we presented participants with question pairs that differed in their expected accuracy, as estimated from the data of Experiment 1. However, we also wanted to know if children could compare their confidence when the expected accuracies were similar, in case some items had objective properties that participants used to make their judgments rather than reasoning about their own perceived chances of answering the questions accurately. Therefore, on the other half of trials, we carefully paired trials that at the group level in Experiment 1 had the same expected accuracy.

A Confidence x Condition x Difference (Same, Different) ANOVA found that adults were more accurate on Chosen than Discarded trials, as described above, $F(1, 106) = 198.64$, $p <$

.001, η_p^2 = .65, and this did not interact with Difference, *p*'s > .111. Participants could distinguish high from low confidence regardless of the expected difference in difficulties.

A Confidence x Condition x Difference (Same, Different) ANOVA found that children were more accurate when the expected accuracies were different, $F(1, 131) = 7.47$, $p = .007$, $\eta_p^2 = .05$, and more accurate on Chosen than Discarded trials, as described earlier, $F(1, 131) = 20.77$, $p < .001$, $\eta_p^2 = .14$. Difference did not interact with Confidence, *p*'s > .453, indicating that children were able to compare their confidence efficiently even when expected accuracies were similar. Difference did interact with Comparison Type (Within or Across-Domains), $F(1, 131) = 3.99$, $p = .048$, $\eta_p^2 = .03$, but as this was not predicted and did not impact our central question about Confidence choices, we did not explore this further.

Global Confidence Judgments

As an exploratory measure (preregistered), we asked participants to indicate which domain they felt they were more accurate in over the course of the study. We then further examine if these judgments aligned with a) their local confidence judgments and b) their accuracy on the tasks.

Most adults (72%) chose Memory as their best overall domain, significantly above chance, $p < .001$ (binomial test). Global confidence judgments corresponded to trial-level confidence choices and accuracy. Participants who chose Memory as their best domain were more likely to choose Memory answers in Across-Domain-Trials (chose Memory on 72%, SD = 14%) compared to participants who chose Perception as their best domain (chose Memory on 54%, SD = 12%), paired $t(60.64) = 6.62$, $p < .001$, $d = 1.34$. Participants who chose Perception as their best domain were somewhat more accurate on Perception items relative to Memory items

(mean difference = 8%, SD = 16%), while participants who chose Memory as their best domain were slightly more accurate on Memory relative to Perception items (mean difference = 1%, SD = 10%), independent $t(36.59) = 2.79$, $p = .008$, $d = 0.75$. Together, this suggests that participants overestimated their performance on the Memory items, but their global confidence judgments still corresponded to their confidence choices and relative accuracy.

Forty-four percent of children chose Perception as their best overall task, not significantly different from chance, $p = .136$. Global confidence judgments corresponded to trial-level confidence choices and accuracy. Children who chose Memory as their best domain chose Memory over Perception in 68% of Across-Domain trials (SD = 18%), while children who chose Perception as their best domain selected Memory on only 55% of Across-Domain trials (SD = 14%), paired $t(129.75) = 4.68$, $p < .001$, $d = 0.8$. Children who chose Perception as their best domain were more accurate on Perception items relative to Memory items (mean difference = 10%, SD = 16%), while children who chose Memory as their best domain were only slightly more accurate on Perception relative to Memory items (mean difference = 2%, SD = 15%), independent $t(119.20) = 3.14$, $p = .002$, $d = 0.55$. Like adults, children overestimated their performance on the Memory items, but their global confidence judgments still corresponded to their confidence choices and relative accuracy.