Supplemental Analyses

Robustness Tests

Below, we report 5 additional Repeated-Measures ANOVAs, all with Confidence Choice and Comparison Type as predictors. The first test replicates the analysis in the main text, but with age as a covariate. The next 3 tests use the accuracy data from each of the three dimensions independently (e.g., Area trials from the Within-Domain comparisons and Area trials only from the Across-Domain comparisons, a total of 24 trials). The final test uses RTs instead of accuracy data.

Accounting for Age. Because children's metacognitive abilities develop, we confirmed that these effects held throughout the age group tested. When treating age as a covariate, we found that children did get marginally more accurate overall with age, $F(1, 46) = 3.74$, $p = .059$, $\eta_p^2 = .08$, but it did not interact with any other variables, all F's < 1. Importantly, we still found that children's confidence choices predicted their accuracy, $F(1,46) = 30.03$, $p < .001$, $\eta_p^2 = .40$, with no effect of comparison type or an interaction, F's < 2.

Area Trials. Children were more accurate on Chosen Area trials than Discarded ones, $F(1, 47) = 17.06$, $p < .001$, $\eta_p^2 = .27$. Within versus Across-Domain comparisons didn't affect accuracy, $F(1, 47) = 0.14$, $p = .712$, $\eta_p^2 = .00$, though Comparison Type interacted with children's Confidence Choices, $F(1, 47) = 5.61$, $p = .022$, $\eta_p^2 = .11$. In particular, children had higher accuracy on Chosen Within-Domain trials than Discarded Within-Domain trials, $t(47) = 4.69$, $p < .001$, $d = .68$, but the difference in accuracy on Across-Domain trials was only marginally significant, $t(47) = 1.87$, $p = .067$, $d = .27$ (see Figure S1).

In contrast, an exploratory analogous Bayesian RM ANOVA found that children's accuracy was best predicted by their Confidence Choices alone, $p(\text{model}|\text{data}) = .67$, $BF_{10} > 19$

000; the model including the interaction between Confidence Choice and Comparison Type was one-third as likely, p(model|data) = .22, BF_{10} = 0.33, providing moderate evidence for no interaction.

Emotion Trials. As shown in Figure S1, children's accuracy on Emotion trials was higher on Chosen trials than Discarded trials, $F(1, 47) = 20.90$, $p < .001$, $\eta_p^2 = .31$. Unexpectedly, children were more accurate on Emotion trials that appeared in Across-Domain comparisons than Within-Domain comparisons, $F(1, 47) = 12.70$, $p < .001$, $\eta_p^2 = .21$. However, there was no interaction between the type of comparison and children's confidence predicting accuracy, $F(1, 47) = 0.48$, $p = .494$, $\eta_p^2 = .01$.

An exploratory analogous Bayesian RM ANOVA similarly found the most support for a model including both main effects, but not their interaction, p(model|data) = 0.70, BF_{10} > 40 000. Comparatively, the model containing the interaction term was anecdotally less likely, p(model|data) = 0.21, BF_{10} = 0.30.

Number Trials. As shown in Figure S1, children were more accurate on Chosen Number trials than Discarded ones, $F(1, 47) = 23.61$, $p < .001$, $\eta_p^2 = .33$. Within-Domain comparisons were marginally more accurate than Across-Domain comparisons, $F(1, 47) = 2.99$, $p = .090$, $\eta_p^2 = .06$, but there was no interaction, $F(1, 47) = 0.24$, $p = .628$, $\eta_p^2 = .01$.

In contrast, an exploratory analogous Bayesian RM ANOVA found the strongest support for Confidence Choices alone predicting Number accuracy, p(model|data) = .63, BF_{10} > 130 000; the model including the interaction term was moderately less likely, p(model|data) = .07, BF_{10} = 0.12.

Response Times. When examining RT in the RM ANOVA, Chosen trials were faster than Discarded trials, $F(1, 47) = 20.05$, $p < .001$, $\eta_p^2 = .30$. There was no difference between

Within and Across-Domain comparison types, $F(1, 47) = 0.21$, $p = .645$, $\eta_p^2 = .01$, and only a marginal interaction, $F(1, 47) = 3.87$, $p = .055$, $\eta_p^2 = .08$.

Summary. Across all additional tests, we find clear support for a link between Confidence Choice and accuracy, but very little support for differences in confidence reasoning between Comparison Types.

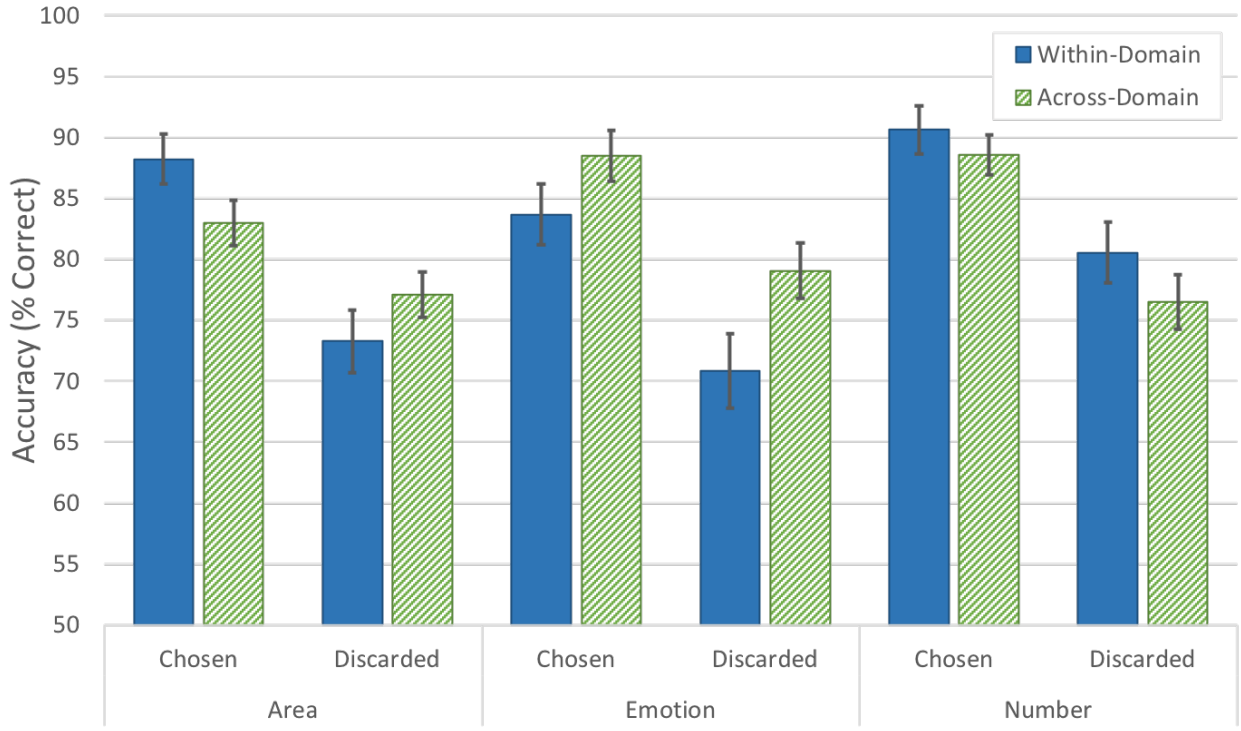

Figure S1. Children's perceptual accuracy on Area, Emotion, and Number trials as a function of Confidence Choice and Comparison Type. Error bars represent 1 Standard Error.

Replication of Baer, Gill & Odic, 2018

Given the similarities between the current study and that of Baer et al. (2018), we felt that the current study could serve as a valuable conceptual replication attempt for some of their key findings. In the analyses below (all preregistered), we focus on two individual differences

measures. First, to quantify perceptual discrimination abilities, we use children's accuracy on all perceptual decisions in each domain (Area, Emotion, and Number). Second, to quantify confidence reasoning abilities, we preregistered an individual differences measure of children's confidence choices by taking the difference of their accuracy on Chosen trials and Discarded trials, such that children with good confidence reasoning should have large accuracy differences between Chosen and Discarded trials. Upon further reflection after the preregistration was submitted and the data collected, we noticed that this difference score was fundamentally limited by children's perceptual accuracy (a child with 100% accuracy would necessarily have a difference score of 0, regardless of their confidence reasoning abilities). Therefore, we also report the analyses using a non-preregistered individual differences measure that captures whether children chose the trial designed to be higher in confidence (i.e., the higher ratio, as reported by Baer et al., 2018).

First, we examined whether children's perceptual accuracy differed between the three domains, as was reported by Baer et al. (2018). A one-way ANOVA found a main effect of Domain, $F(1.58, 74.32) = 4.33$, $p = .024$, $\eta_p^2 = .08$. Accuracy on Number trials ($M = 84\%$, $SD = 7\%$) was significantly higher than accuracy on Area trials ($M = 80\%$, $SD = 6\%$), $t(47) = 4.11$, $p < .001$, $d = 0.59$ (Bonferroni corrected). We also found, contrary to our hypothesis and inconsistent with Baer et al., that there were correlations between perceptual discrimination accuracy in the three domains (see Table S1).

Next, we found that children responded strategically to confidence: the pre-registered difference score was greater than chance on all three dimensions, Area: $M = 9.13$, $SD = 17.59$, $t(47) = 3.60$, $p < .001$, $d = 0.52$, Emotion: $M = 10.46$, $SD = 15.45$, $t(47) = 4.69$, $p < .001$, $d = 0.68$, Number: $M = 11.11$, $SD = 15.97$, $t(47) = 4.82$, $p < .001$, $d = 0.70$. Similarly, children chose

the High-Confidence trials more often than chance in all three dimensions, Area: M = 74%, SD = 25%, $t(47) = 6.64$, $p < .001$, $d = 0.96$, Emotion: M = 67%, SD = 23%, $t(47) = 5.20$, $p < .001$, $d = 0.75$, Number: M = 63%, SD = 24%, $t(47) = 3.80$, $p < .001$, $d = 0.55$. There was no difference between the three domains in difference scores, $F(2, 47) = 0.42$, $p = .660$, $\eta_p^2 = .00$, though there was a domain difference in High-Confidence trial choice, $F(2, 94) = 5.97$, $p = .004$, $\eta_p^2 = .11$, with Area choices more likely to reflect the designed confidence than Number choices, $t(47) = 3.50$, $p = .003$, $d = 0.51$ (Bonferroni corrected).

Consistent with Baer et al. (2018), we found that individual differences in confidence reasoning correlated between the three domains for both the difference score and the High-Confidence choices (see Table S1). However, these correlations could be driven by the correlated perceptual accuracies, so we re-ran the correlations between confidence measures while controlling for accuracy on all three perceptual tasks. As shown in Table S1, all correlations remained strong, indicating that children's confidence judgments were related even above and beyond potential perceptual similarities, replicating the findings of Baer et al.

Table S1. Correlations between perceptual dimensions.

	Perceptual Accuracy		Accuracy Difference		Choice of High Confidence		Accuracy Difference Controlling for All Perceptual Accuracies		Choice of High Confidence Controlling for All Perceptual Accuracies	
	Emotion	Number	Emotion	Number	Emotion	Number	Emotion	Number	Emotion	Number
Area	.42**	.47***	.48***	.59***	.50***	.62***	.52***	.66***	.54***	.70***
Emotion		.30*		.64***		.63***		.62***		.64***

Note: * denotes $p < .05$, ** denotes $p < .01$, and *** denotes $p < .001$