Supplemental Materials

The benefits of a metacognitive lesson on children's understanding of mathematical equivalence, arithmetic, and place value

In the manuscript, we report on children's accuracy, monitoring scores, and control scores across three mathematics topics. The monitoring scores represent the calibration of children's accuracy and certainty ratings (e.g., giving a high certainty rating when solving the item correctly), and the control scores represent the calibration of children's study selections and certainty ratings (e.g., giving a high certainty rating and opting not to re-study an item). Here, we report two sets of supplemental analyses. The first set focuses on children's raw certainty ratings and study selections across the three mathematics topics. Certainty ratings were completed at pretest, posttest, and retention test. Study selections were completed at pretest and retention test only. The second set focuses on a different way of measuring monitoring scores.

Quantifying Children's Certainty Ratings and Study Selections at Pretest

Across all pretest items, children's average certainty rating was very high (M = 3.43 out of 4.00, SE = 0.04). A repeated measures ANOVA with topic entered as a within-subject factor and condition entered as a between-subject factor revealed a significant main effect of topic, $F(2, 266) = 23.46$, $p < .001$, $\eta_p^2 = .15$. Children were most certain on arithmetic items (M = 3.59, SE = 0.04), followed by place value items (M = 3.43, SE = 0.05), and least certain on equivalence items (M = 3.26, SE = 0.06). All three pairwise comparisons with a Bonferroni correction were statistically significant (ps < .016). Conditions were well-matched in pretest certainty as there was not a significant main effect of condition, $F(1, 133) = 0.39$, $p = .532$, $\eta_p^2 = .00$, or a condition-by-topic interaction, $F(2, 266) = 0.05$, $p = .951$, $\eta_p^2 = .00$.

Across all pretest items, children opted to study 32% of problems (SE = 2%). A repeated measures ANOVA with topic as a within-subject factor and condition as a between-subject factor revealed a significant main effect of topic, $F(2, 266) = 27.72$, $p < .001$, $\eta_p^2 = .17$. Children opted to study fewer arithmetic items (M = 20%, SE = 2%) relative to equivalence items (M = 40%, SE = 3%) and place value items (M = 37%, SE = 3%). Pairwise comparisons with a Bonferroni correction revealed that study selections were significantly lower on arithmetic items relative to the other two topics (ps < .016), but the difference between equivalence items and place value items was not statistically significant (p = 1.00). Conditions were well-matched in pretest study selections as there was not a significant main effect of condition, $F(1, 133) = 0.12$, $p = .726$, $\eta_p^2 = .00$, or a condition-by-topic interaction, $F(2, 266) = 2.04$, $p = .132$, $\eta_p^2 = .01$.

Condition Differences in Children's Certainty Ratings and Study Selections on the Immediate Posttest and Delayed Retention Test

Table S1 presents the raw certainty ratings by condition at pretest, posttest, and retention test. Table S2 presents the study selections at pretest and retention test. To examine condition differences at posttest and retention test, we conducted repeated measures ANCOVAs with lesson condition (Metacognitive and Control) included as a between-subject effect and topic (Equivalence, Arithmetic, and Place Value) included as a within-subject effect. The models predicting certainty ratings also included pretest certainty ratings within each topic as covariates. The model predicting study selections also included pretest study selections within each topic as

covariates. However, the conclusions remained unchanged when no covariates were included or when additional covariates were included (e.g., pretest accuracy, monitoring, and control scores).

With average certainty ratings at posttest as the dependent variable, there was not a main effect of condition, $F(1, 130) = 0.55$, $p = .459$, $\eta_p^2 = .00$, as children in the Metacognitive Lesson had similar average certainty ratings at posttest ($M = 3.46$, $SE = 0.03$) relative to children in the Control Lesson ($M = 3.49$, $SE = 0.03$). There was a significant main effect of topic, $F(2, 260) = 9.61$, $p < .001$, $\eta_p^2 = .07$, but the condition-by-topic interaction was not statistically significant, $F(2, 260) = 0.68$, $p = .506$, $\eta_p^2 = .01$.

With average certainty ratings at retention test as the dependent variable, there was not a main effect of condition, $F(1, 130) = 0.09$, $p = .762$, $\eta_p^2 = .00$, as children in the Metacognitive Lesson had similar average certainty ratings at retention test ($M = 3.43$, $SE = 0.04$) relative to children in the Control Lesson ($M = 3.44$, $SE = 0.05$). There was a significant main effect of topic, $F(2, 260) = 4.87$, $p = .008$, $\eta_p^2 = .04$, but the condition-by-topic interaction was not statistically significant, $F(2, 260) = 0.02$, $p = .983$, $\eta_p^2 = .00$.

With the percentage of items selected for restudy at retention test as the dependent variable, there was not a main effect of condition, $F(1, 130) = 0.98$, $p = .325$, $\eta_p^2 = .01$, as children in the Metacognitive Lesson opted to study a similar percentage of items at retention test ($M = 32\%$, $SE = 2\%$) relative to children in the Control Lesson ($M = 28\%$, $SE = 3\%$). There was not a significant main effect of topic, $F(2, 260) = 2.66$, $p = .072$, $\eta_p^2 = .02$, and the condition-by-topic interaction was not significant, $F(2, 260) = 0.74$, $p = .479$, $\eta_p^2 = .01$.

See tables at the end of this document.

Operationalizing Children's Monitoring Skills as Gamma Correlations

In the manuscript, we report on a measure of absolute monitoring accuracy, which assessed the match between children's certainty ratings and their accuracy on each item. Here, we report descriptive information on a measure of relative monitoring accuracy – the Goodman-Kruskal gamma correlation between certainty ratings and accuracy scores across a set of items. The gamma correlation assesses children's ability to discriminate between problems on which they are successful and problems on which they are unsuccessful. These correlations can vary between –1 and +1, and the closer to +1 the higher the monitoring accuracy.

For each child, we calculated 9 gamma correlations – one for each of the three topics (arithmetic, equivalence, and place value) at pretest, posttest, and retention test. Then, we averaged across children to obtain an average gamma correlation for each topic at each time point. This measure was somewhat problematic. Each correlation was based on only 6 items (e.g., correlating children's scores on the six arithmetic items at pretest with their certainty ratings on the six arithmetic items at pretest). Given the small number of items, the gamma correlations were often exactly –1 or +1. For example, as shown in Table S3 below, the median gamma correlation was +1 for 7 of the 9 variables. Further, calculating correlations requires variability in the two measures being correlated. Many children in our sample were invariant in their accuracy or in their certainty within a particular topic (e.g., giving a certainty rating of 4 on all six arithmetic items, or answering all six arithmetic items correctly), which resulted in a lot of missing data on these gamma variables. For each of the nine variables considered separately, we had missing data on 48-83 children. See Table S3 at the end of this document.

Table S1: Children's certainty ratings at pretest, posttest, and retention test (operationalized as the mean rating on a scale from 1 to 4)

	Control Lesson *M (SD)*	**Metacognitive Lesson** *M (SD)*	**Total** *M (SD)*
	N = 65	N = 70	N = 135
Pretest Certainty Ratings			
Arithmetic	3.62 (0.46)	3.57 (0.43)	3.59 (0.45)
Equivalence	3.29 (0.69)	3.23 (0.64)	3.26 (0.66)
Place Value	3.45 (0.54)	3.42 (0.52)	3.43 (0.53)
Posttest Certainty Ratings			
Arithmetic	3.56 (0.45)	3.54 (0.43)	3.55 (0.44)
Equivalence	3.46 (0.64)	3.34 (0.65)	3.40 (0.64)
Place Value	3.52 (0.53)	3.46 (0.49)	3.49 (0.51)
Retention Test Certainty Ratings			
Arithmetic	3.54 (0.54)	3.47 (0.50)	3.51 (0.52)
Equivalence	3.38 (0.74)	3.31 (0.76)	3.34 (0.75)
Place Value	3.47 (0.59)	3.43 (0.50)	3.45 (0.54)

Table S2: Children's study selections at pretest and retention test (operationalized as the percentage of items selected for re-study)

	Control Lesson *M (SD)*	**Metacognitive Lesson** *M (SD)*	**Total** *M (SD)*
	N = 65	N = 70	N = 135
Pretest Study Selections			
Arithmetic	18% (23%)	23% (29%)	20% (26%)
Equivalence	43% (34%)	38% (33%)	40% (33%)
Place Value	40% (36%)	35% (34%)	37% (35%)
Retention Test Study Selections			
Arithmetic	23% (28%)	26% (29%)	25% (28%)
Equivalence	32% (34%)	31% (33%)	31% (33%)
Place Value	31% (34%)	36% (34%)	33% (34%)

Table S3: Descriptive statistics for gamma correlations across topic and time point

	Pretest Arithmetic Gamma	Pretest Equiv. Gamma	Pretest Place Val. Gamma	Posttest Arithmetic Gamma	Posttest Equiv. Gamma	Posttest Place Val. Gamma	Retention Arithmetic Gamma	Retention Equiv. Gamma	Retention Place Val. Gamma
N valid	81	79	81	85	52	67	87	55	66
N missing	54	56	54	50	83	68	48	80	69
Mean	0.65	0.85	0.19	0.73	0.53	0.36	0.77	0.69	0.38
Median	1.00	1.00	0.50	1.00	1.00	1.00	1.00	1.00	0.69
SD	0.66	0.43	0.84	0.55	0.78	0.79	0.56	0.64	0.78
Min	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00