SOM-S1

Children's explanations following their initial judgment about the dolls' weight and their subsequent judgement post-testimony

The first author and a research assistant blind to the hypotheses of the study coded 25% of the total number of explanations children provided. Both coders were blind to children's age, condition, and judgments about the dolls. Agreement was 95%. Disagreements were resolved through discussion. The research assistant coded the remaining explanations.

We coded children's explanations following their initial and post-testimony judgments. Explanations were coded as: *Bigger = Heavier* if children described a positive association between size and weight (e.g., "It's the largest so it's the heaviest"); as *Smaller = Heavier* if they described a negative association between size and weight (e.g., "It's small so it's the heaviest) or referred to being told that the smallest was the heaviest (e.g., "Because you told me"); as *Size Sometimes Unrelated to Weight* if they described why the biggest doll might not be the heaviest (e.g., "The largest one is hollow, while the smallest one is solid"). Finally, explanations were coded as *Other* if they could not be coded into the other four categories (e.g., "It's the heaviest", "It's very heavy, as heavy as four water bottles", "It is just right") or if children did not provide an explanation. Table S1 shows the percentage of each type of explanation as a function of age, testimony type, and timing.

Table S1. Percentage of preschool and elementary school children's explanations coded into each category as a function of Testimony Type (Confirming vs. Counter-Intuitive), and Timing (Before vs. After Testimony).

Explanation Type	**Confirming Testimony**		**Counter-Intuitive Testimony**	
	Before	***After***	***Before***	***After***
Preschool Children				
Other	21%	28%	26%	34%
Bigger = Heavier	79%	72%	74%	8%
Smaller = Heavier	0%	0%	0%	55%
Size Sometimes Unrelated to Weight	0%	0%	0%	3%
Elementary School Children				
Other	0%	7%	0%	9%
Bigger = Heavier	95%	80%	96%	0%
Smaller = Heavier	0%	2%	0%	38%
Size Sometimes Unrelated to Weight	5%	11%	4%	53%

Both age groups mostly offered Bigger = Heavier explanations (e.g., "It's the largest so it's the heaviest") before and after confirming testimony. By contrast, both age groups mostly offered Bigger = Heavier explanations before but rarely after counter-intuitive testimony. Children who endorsed the experimenter's counter-intuitive claim justified their decision either by repeating the experimenter's claim (i.e., *Smaller = Heavier*) or by noting that size and weight are not always correlated (i.e., *Size is Sometimes Unrelated to Weight*). The latter explanation was common among elementary school children (53%) but was rare among preschool children (3%). Thus, although preschool and elementary school children endorsed the counter-intuitive testimony of the experimenter at similar rates, they often did so for different reasons.

SOM-S2

Replicating analyses of children's exploration of the dolls (i.e., picking up the dolls) when children who did not touch any of the dolls are excluded from the analyses.

In Figure S2, we display the number of times each child picked up each doll during the experimenter's absence for children who touched at least one doll. We conducted the same a 2x2x2x5 ANOVA with the between subject factors of Age Group (2: Elementary vs. Preschool), Testimony Type (2: Counter-Intuitive vs. Confirming), and Priming (2: Prime vs. no Prime), and the within subject factor of Doll (5: one (i.e., smallest), two, three, four, and five (i.e., biggest)) on the number of times children picked up a doll.

This analysis revealed a main effect of Priming, $F(1, 136) = 4.04$, $p = .046$, $\eta^2_p = .03$. Children picked up the dolls more often when they received a prime to explore than when they did not, $M = 8.36$, $SD = 6.87$ vs. $M = 5.80$, $SD = 5.24$. However, receiving a prime to explore did not interact with Age Group, Testimony Type, or Dolls. By implication, receiving a prime to explore increased children's general exploration of the dolls (i.e., whether they picked up the dolls) but it did not increase their targeted exploration of the dolls.

Our analyses also revealed a main effect of Doll, $F(4, 544) = 25.83$, $p < .001$, $\eta^2_p = .16$, a main effect of Age Group, $F(1, 136) = 4.67$, $p = .032$, $\eta^2_p = .03$, a significant Age Group X Testimony Type interaction, $F(1, 136) = 5.10$, $p = .026$, $\eta^2_p = .04$, a significant Doll X Testimony Type interaction, $F(4, 544) = 5.18$, $p < .001$, $\eta^2_p = .04$, and a significant Doll X Age Group interaction, $F(4, 544) = 4.08$, $p = .003$, $\eta^2_p = .03$. The Doll X Age Group X Testimony Type interaction was not significant, $F(4, 544) = 1.71$, $p = .15$, $\eta^2_p = .01$. To better understand these interactions, we computed simple effects using a Bonferroni correction. For preschool children, the frequency with which they picked each doll was not influenced by the type of testimony they received, $p > .09$ for all tests. By contrast elementary school children picked up the smallest and the biggest dolls significantly more often when they received counter-intuitive rather than confirming testimony (both $p < .007$). Testimony type did not influence the frequency with which elementary school children picked up the three intermediate dolls, $p > .25$ for all dolls. Thus, elementary school children but not preschool children sought evidence that could confirm or disconfirm the informant's testimony. Moreover, this targeted exploration cannot be attributed to a general tendency of elementary school children to pick up the dolls more often than preschool children. Preschool and elementary school children in the confirming testimony condition did not differ in the frequency with which they picked up all five dolls, $p > .33$ for all five dolls. Thus, we replicate the same pattern of result we report in the manuscript on this sub-sample of our data.

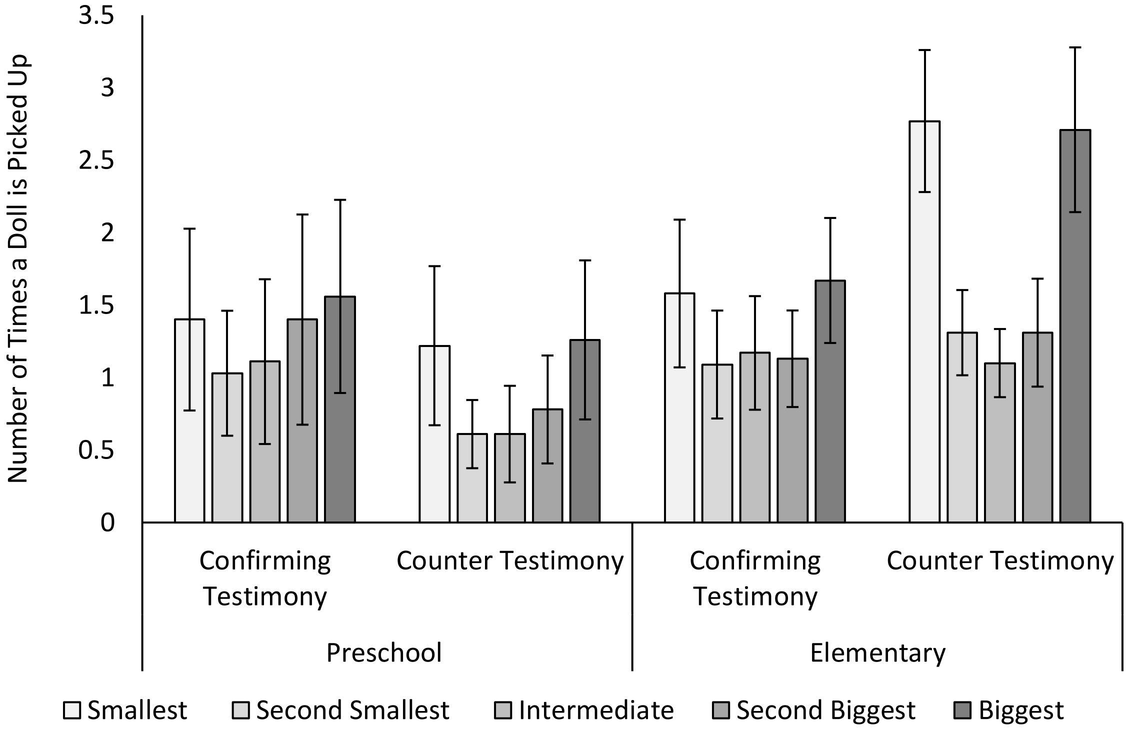


Figure S2. Number of dolls children picked up when the experimenter left the room. Error bars represent 95% confidence intervals.

SOM-S3

Did children pick up the smallest and the biggest doll *concurrently* (Figure S3)?

Preschoolers rarely picked up the smallest and the biggest doll concurrently and when they did it was unrelated to the type of testimony they received, $\chi^2(1, n = 81) = .81, p = .37$. Elementary school students picked up the smallest and the biggest doll concurrently significantly more often following counter-intuitive than confirming testimony, GLH Test: $\chi^2(1, n =109) = 5.12, p = .024$.

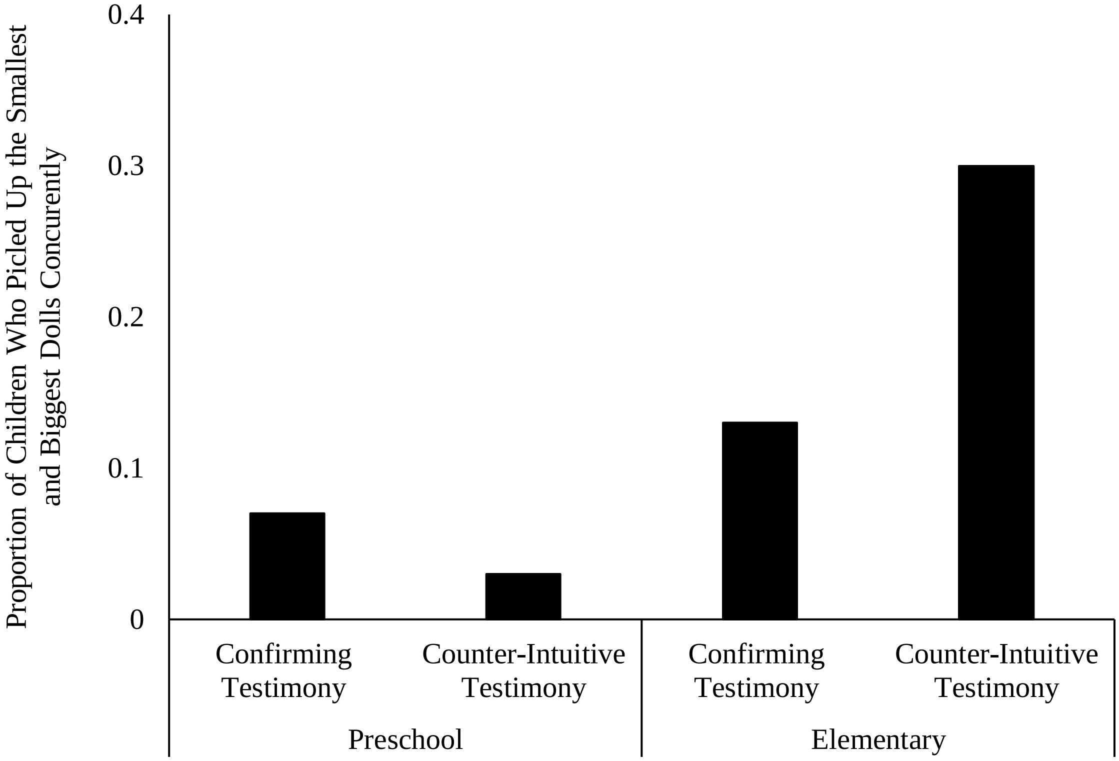


Figure S3. Proportion of children receiving confirming vs. counter-intuitive testimony who picked up the biggest and the smallest doll *at the same time* while E1 was out of the room.

SOM-S4

Children's decision to explore in the counter-intuitive testimony condition was unrelated to whether they endorsed or rejected the experimenter's testimony.

Whether children endorsed the experimenter's testimony that smallest = heaviest or stuck to their initial intuition that biggest = heaviest was unrelated to their decision to pick up the biggest and smallest doll during the experimenter's absence, 63.41 % vs. 55.56%, $\chi^2(1, n = 91) = 0.21$, $p = .64$. This was true for preschool and elementary school children, $\chi^2(1, n = 38) = 0.52$, $p = .47$, $\chi^2(1, n = 53) = 0.89$, $p = .35$, respectively.

Children's decision to explore in the counter-intuitive testimony was unrelated to the type of explanation they provided following their endorsement or rejection of the testimony.

Did children's explanation for their judgment following the receipt of counter-intuitive testimony predict their decision to explore the dolls? To answer this question, in Table S4, we provide the percentage of children in each age group who explored as a function of the type of explanation they provided. Inspection of Table S4 reveals no clear association between children's explanations and whether they picked up the smallest and the biggest doll. Preschool children were unlikely to pick up these dolls no matter what explanation they gave. Similarly, elementary school children very often picked up these two dolls no matter what explanations they gave.

Table S4. Percentage of children in each age group who explored as a function of the type of explanation they provided.

Explanation Type	*Explored*
Preschool Children (*n* = 38)	
Other (n = 13)	38%
Bigger = Heavier (n = 3)	33%
Smaller = Heavier (n = 21)	29%
Size Sometimes Unrelated to Weight (n = 1)	100%
Elementary School Children (*n* = 53)	
Other (n = 5)	40%
Bigger = Heavier	
Smaller = Heavier (n = 20)	80%
Size Sometimes Unrelated to Weight (n = 28)	93%

SOM-S5

Differences in children's judgments with E1 and E2

Children made a judgment about the weight of the dolls: (1) immediately after the opportunity to explore the dolls by E1; (ii) when explicitly asked by E2; and (iii) when invited by E2 to select the heaviest paperweight. In table S5, we display the percentage of preschool and elementary school children who endorsed the biggest doll as the heaviest at each time point in each condition. To assess the stability of children's judgments over these three points we regressed using a multi-level logistic regression model (Stata 14's –xtlogit- command) children's judgements on the timing of these judgments using two dummy variables (Explicit Judgement with E2 and Paperweight Task with E2, the reference category was Explicit Judgement with E1). This allowed us to compare whether children's judgements changed significantly across the three time points. We conducted these analyses separately for each type of testimony.

Table S5. Percentage of preschool and elementary school children who endorsed the biggest doll as the heaviest (1) immediately after the opportunity to explore the dolls when questioned by E1; (ii) when explicitly asked by E2; and (iii) when invited by E2 to select the heaviest paperweight.

	Following Opportunity to Explore with E1	**Initial Judgement with E2**	**Paperweight Task**
Confirming			
Preschool (n = 43)	98%	93%	77%
Elementary (n = 56)	96%	100%	91%
Total (n = 99)	97%	97%	85%
Counter-Intuitive			
Preschool (n = 38)	34%	37%	45%
Elementary (n = 53)	49%	53%	68%
Total (n = 91)	43%	46%	58%

Confirming Testimony: We found that when children received confirming testimony the proportion of children who stated that the biggest doll was the heaviest did not differ whether children were asked by E1 or by E2 in a direct manner, i.e., “Which doll do you think is the heaviest?”. However, when asked to select a heavy paperweight by E2 children were significantly less likely to select the biggest doll relative to when they were asked in a direct manner by E1 and E2, $z = 2.87$, $p = .004$. However, when we tested whether this pattern applied to both preschool and elementary school children, we found that Elementary school children’s judgement that biggest = heaviest did not change significantly across the three time points. In contrast, preschool children were significantly less likely to select the biggest doll on the paperweight task relative to when they were asked in a direct manner by E1 and E2, 98% vs. 77%, $z = 4.39$, $p < .001$, 93% vs. 77%, $z = 3.09$, $p < .01$, respectively (see Table S5). Thus, children’s judgements about the weight of the doll in the confirming testimony condition were generally stable. When asked directly about the weight of the dolls by a second experimenter they had never met before, most of the preschool and elementary school children provided answers that were similar to those they had given to E1 after they had had an opportunity to explore the dolls. Moreover, Elementary school children provided equivalent answers whether E2 asked them directly or indirectly (i.e., by asking them to select a heavy paperweight). Preschool children deviated from this pattern; they were less likely to select the largest doll as a suitable paperweight. However, even then, the vast majority of preschoolers continued to endorse the biggest doll as the heaviest.

Counter-Intuitive Testimony: We found that when children received counter-intuitive testimony the proportion of children who stated that the biggest doll was the heaviest did not differ whether children were asked by E1 or by E2 in a direct manner, i.e., "Which doll do you think is the heaviest?". However, when asked to select a heavy paperweight by E2 children were significantly less likely to select the biggest doll relative to when they were asked in a direct manner by E1 and E2, $z = 5.08$, $p > .001$, $z = 3.95$, $p > .001$, respectively.

We followed up on this interaction by investigating whether being asked about the doll's weight differed as a function of children's age and whether they had received a prime. In Table S5b, we display the proportion of children receiving counter-intuitive testimony who stated that biggest = heaviest at three successive time-points: immediately after having had the opportunity to explore the dolls (i.e., children's final judgment with E1); when asked by E2; and when invited by E2 to select the heaviest paperweight. Figure D1 displays these proportions for each of the four combinations of age and prime.

The timing of the three judgments questions did not explain a significant amount of variance in the probability that children in the counter-intuitive condition selected the biggest doll in response to the E1 and E2's questions for three groups of children: preschool children who did not receive a prime to explore, Model $\chi^2(2) = 1.18$, $p = .55$, elementary school children who did not receive a prime to explore, Model $\chi^2(2) = 4.73$, $p = .09$, and elementary school children who received a prime to explore, Model $\chi^2(2) = 5.02$, $p = .08$. The timing of the three judgments questions explained a significant amount of variance for the probability that preschool children who received a prime selected the biggest doll, Model $\chi^2(2) = 11.82$, $p = .003$. This group of children was more likely to select the biggest doll as the heaviest when asked to select a paperweight than when asked directly about the weight of the dolls by E1, $z = 3.41$, $p = .001$.

However, they were equally likely to select the biggest doll as the heaviest whether E2 asked them directly or asked them to select a paperweight, *z* = 1.86, *p* = .06.

Thus, children’s judgements about the weight of the doll were generally stable. When asked directly about the weight of the dolls by a second experimenter they had never met before, most of the preschool and elementary school children provided answers that were similar to those they had given to E1 after they had had an opportunity to explore the dolls. Indeed, children provided equivalent answers whether E2 asked them directly or indirectly (i.e., by asking them to select a heavy paperweight). Only one group deviated from this pattern; preschool children who had received a prime to explore the dolls were more likely to select the largest doll as a suitable paperweight.

Conclusion: Given the relative stability of children’s judgements across the three time points in both the confirming and counter-intuitive testimony conditions, we added together the three judgements children made following the opportunity to explore the dolls (i.e., their judgement with E1, their judgement with E2, and the judgement they made as part of the paperweight task they completed with E2

Table S5b. Percentage of preschool and elementary school children in the counter-intuitive condition who endorsed the biggest doll as the heaviest (1) immediately after the opportunity to explore the dolls when questioned by E1; (ii) when explicitly asked by E2; and (iii) when invited by E2 to select the heaviest paperweight.

	Following Opportunity to Explore with E1	**Initial Judgement with E2**	**Paperweight Task**
No Prime			
Preschool (n = 21)	28%	38%	43%
Elementary (n = 26)	46%	42%	54%
Total (n=47)	43%	40%	49%
Prime			
Preschool (n = 17)	29%	35%	47%
Elementary (n = 27)	52%	63%	81%
Total (n=44)	43%	52%	68%

SOM-S6

Replicating analyses of children's post-exploration weight judgments as a function of children's exploration.

In the manuscript, we investigated whether children's exploration impacted their subsequent weight judgments via a 2 x 2 x 2 ANOVA with Age Group (2: Preschool, Elementary), Priming (2: Prime, No Prime), and Exploration (2: Explored, Did Not Explore) as between-subject factors, restricting our analysis to children who had received counter-intuitive testimony. In the manuscript, we operationalized exploration as children's decision to pick up the biggest and the smallest doll during the experimenters' absence. Here, we replicate these results using a less stringent measures of exploration, i.e., whether children picked up any two dolls during the experimenter's absence. In fact, we replicate the results of the manuscript exactly because the same exact children are counted as having explored using picking up the smallest and the heaviest dolls or picking up any two dolls during the experimenter's absence as the measure of exploration.

Thus, our analysis revealed only a significant main effect of Exploration, $F(1,83) = 21.79$, $p < .001$, $\eta^2{}_p = .21$: Children who had explored judged the biggest doll to be the heaviest much more often than children who had not explored. We display this main effect in Figure 5. Children who did not explore the dolls, endorsed the smallest doll as the heaviest significantly above chance, $t(33) = 4.42$, $p < 0.001$, $d = 1.54$. In contrast, children who did explore the dolls, endorsed the biggest doll as the heaviest significantly above chance, $t(56) = 2.83$, $p = 0.006$, $d = .75$. Thus, when children gathered empirical evidence, it undermined the earlier impact of E1's counter-intuitive testimony on their judgments. The absence of any interaction between age and exploration, *F*(1,83) = .30, *p* > .25, indicates that when preschool or elementary school children explored, it impacted their judgments to the same extent, as shown in Figure 3.

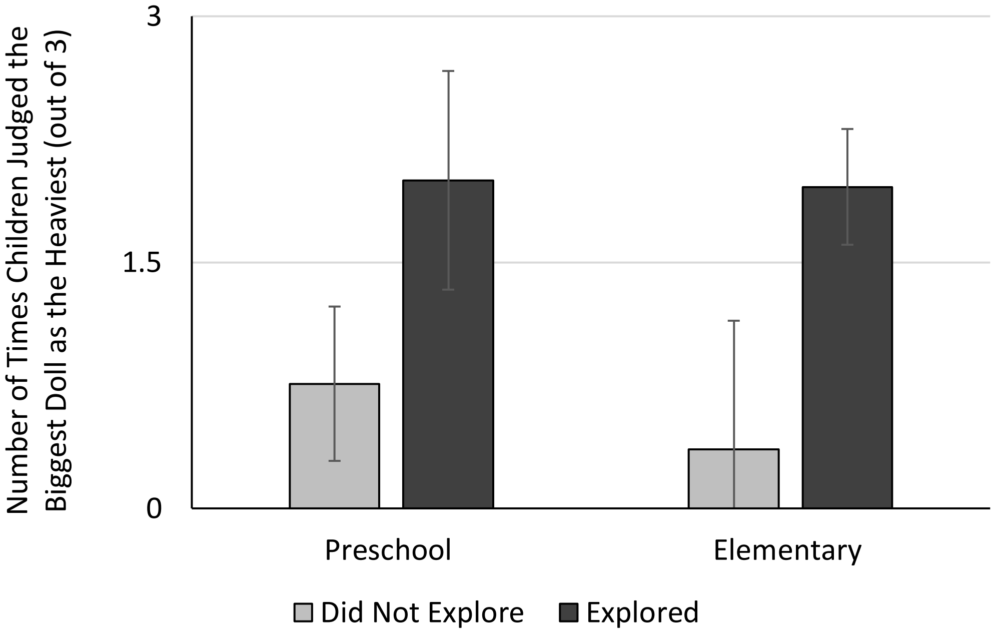


Figure S6. Average number of times preschool and elementary children judged that the biggest doll was the heaviest as a function of whether they picked up any two dolls during the experimenter's absence (Explored: 44 elementary school children, 13 preschool children; Did Not Explore: 9 elementary school children, 25 preschool children). Error bars represent 95% confidence intervals.

SOM-S7

Is the age change observed between preschool and elementary school due to cognitive maturation or schooling?

We conducted post-hoc exploratory analyses to find out. We tested for a significant interaction between age (in months) and the kind of testimony children received within each school group (i.e., preschool and elementary school). If there is an effect of age (within school group), we would expect a positive interaction between age and whether children received counter-intuitive testimony, i.e., older children are more likely to explore the dolls in the counter-intuitive testimony condition than in the confirming testimony condition. Our analyses do not find such an interaction. This is consistent with the claim that instruction at the elementary school level explains developments in children's empirical stance. However, these post-hoc tests do not rule out the cognitive maturation hypothesis. More research comparing children who begin elementary school at different ages is needed to more fully test these two hypotheses.

We conducted three sets of analyses. We looked for a positive interaction between age and whether children received counter-intuitive testimony (controlling for children's receipt of a prompt to explore) within the group of children who attended preschool and within the group of those who attended elementary school on the:

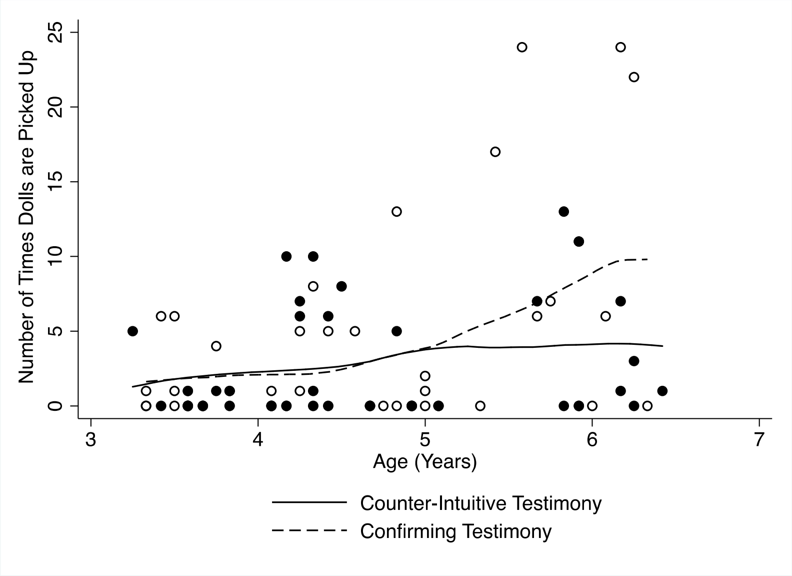
1) The total number of dolls children picked up (i.e., we summed together the number of times children picked up each of the five dolls). This is a gross measure of children's exploration.
2) The total number of times children picked up the smallest doll.

3) The total number of times children picked up the biggest doll.

We conducted separate analyses for the number of times children picked up the smallest and the biggest dolls because these are the two dolls that our prior analyses identified as having been selectively picked up by children (Manuscript, Figure 4).

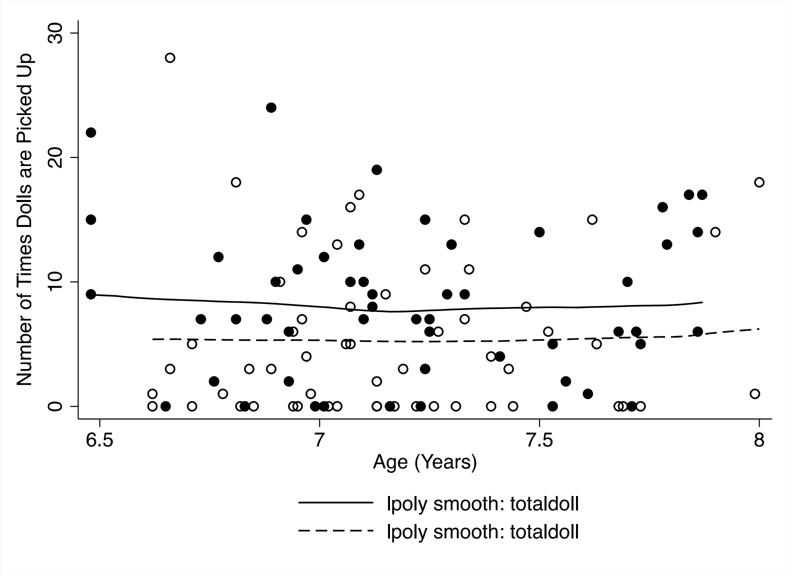
In each case ,we regressed our dependent variable on whether children received a prompt, their age (in months), whether they received counter-intuitive testimony, and the interaction between age and whether children received counter-intuitive testimony.

Preschool All Dolls



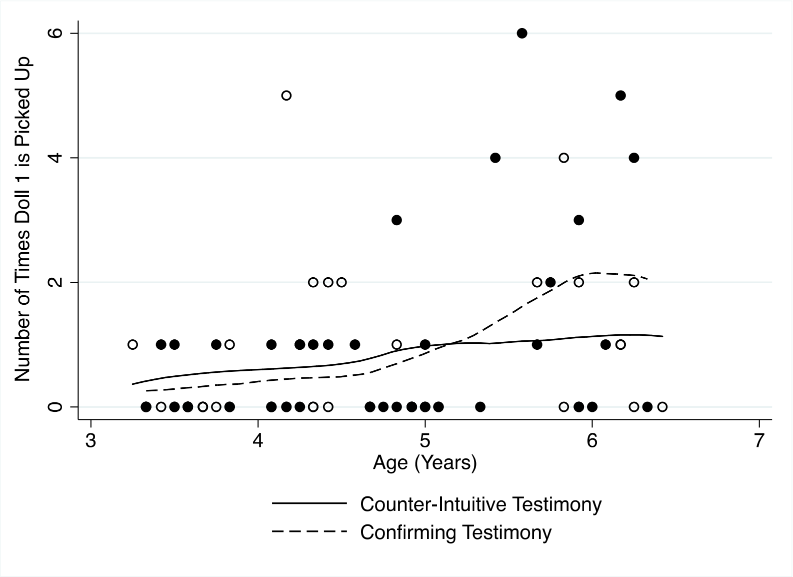
No significant interaction between Age (months) and Testimony Type, *t* = 1.81, *p* = .07

Elementary All Dolls



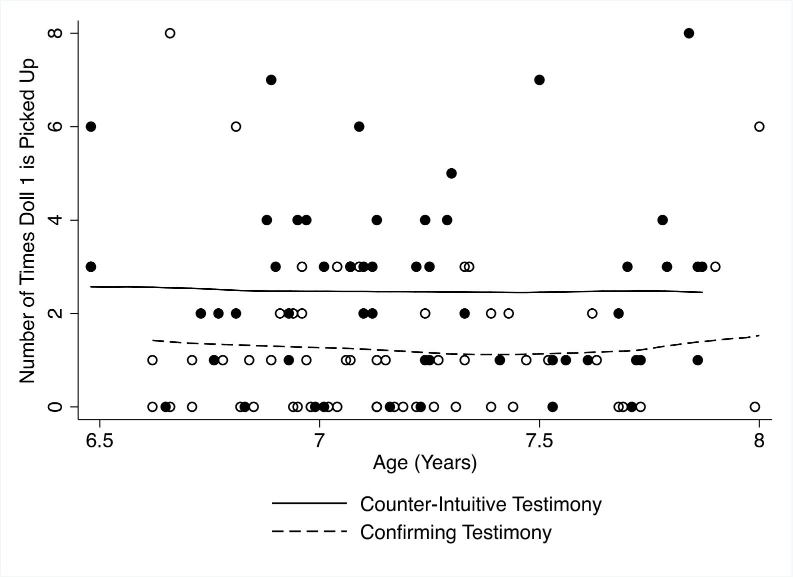
No significant interaction between Age (months) and Testimony Type, *t* = .32, *p* = .75

Preschool Smallest Doll



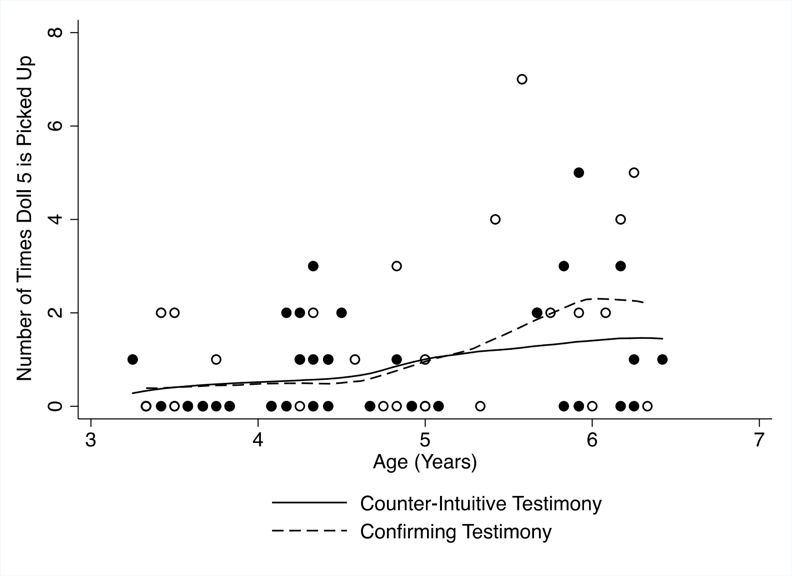
No significant interaction between Age (months) and Testimony Type, *t* = 1.45, *p* = .15

Elementary School Smallest Doll



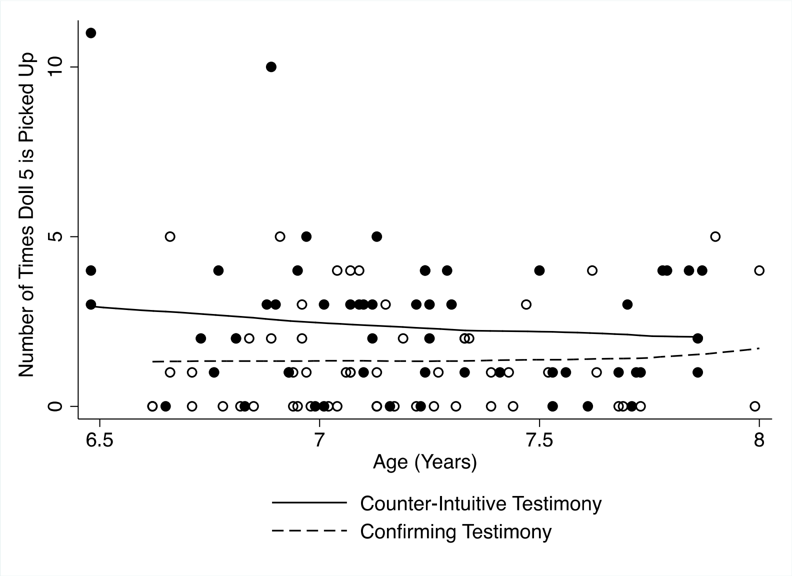
No significant interaction between Age (months) and Testimony Type, *t* = .06, *p* = .95

Preschool Biggest Doll



No significant interaction between Age (months) and Testimony Type, *t* = .87, *p* = .39

Elementary Biggest Doll



No significant interaction between Age (months) and Testimony Type, *t* = 1.73, *p* = .09