The role of association in pre-schoolers solutions to 'spoon tests' of future planning

Katherine Leah Dickerson[1]
James Alexander Ainge[1]
Amanda Madeleine Seed[1*]

[1]University of St Andrews, Department of Psychology and Neuroscience, St Andrews, Fife KY16 9AZ, United Kingdom

Correspondence to Lead Contact: Amanda Seed ams18@st-andrews.ac.uk

Keywords

Episodic future thinking, associative learning, foresight, planning, mental time travel

Highlights and eTOC Blurb

Summary

Imagining the future is a powerful tool for making plans and solving problems. It is thought to rely on the episodic system which also underpins remembering a specific past event [1-3]. However, the emergence of episodic future thinking over development and evolution is debated [4-9]. One key source of positive evidence in pre-schoolers and animals is the 'Spoon Test' or Item Choice Test [4, 10], in which participants encounter a problem in one context, and then a choice of items in another context, one of which is the solution to the problem. The majority of studies report that most children choose the right item by the age of 4 [10-15, cf 16]. Apes and corvids have also been shown to pass versions of the test [17-19]. However, it has been suggested that a simpler mechanism could be driving choice: the participant simply chooses the item that has been assigned salience or value, without necessarily imagining the future event [16, 20-23]. We developed a new test in which two of the items offered to children were associated with positive outcomes, but only one was still useful. We found that older children (5-, 6- and 7-year-olds) chose the correct item at above chance levels, but younger children (3- and 4-year-olds) did not. In further tests 4-year-olds showed an intact memory for the encoding event. We

conclude that positive association substantially impacts performance on Item Choice Tests in 4-year-olds, and that future planning may have a more protracted developmental trajectory than episodic memory.

Results and Discussion

Experiment 1

We developed a new Item Choice Test (ICT) designed to rule out explanations for success based on one of the items acquiring higher value through past experience. We tested eleven 3-year-olds (*M* = 41 months, range 36 to 47 months); forty-four 4-year-olds (*M* = 55 months, range 48 to 59 months); sixty-one 5-year-olds (*M* = 66 months, range 60 to 71 months); forty-nine 6-year-olds (*M* = 77 months, range 72 to 83 months); and forty-seven 7-year-olds (*M* = 89 months, range 85 to 95 months) (Table 1). Each child was taught how to use two visually distinct boxes, which dispensed stickers when the participant placed the correct token into the machine. After learning this, children were told that one box would remain in place and they would return to it, and while the other box was being put away, they were told that it would no longer be available (the order in which these two actions was performed was counterbalanced across participants). They went to another room to complete a vocabulary test and they were then offered 3 tokens to choose between. These included the token from the box which was accessible to them (*correct*), the token which operated the unavailable box (*associate distractor*), and a third token that they had never seen before (*novel distractor*). To pass, the child had to use their memory of the encoding event (which box was left on the table) to plan for the future. If children simply chose objects that had gained incentive value by previously being paired with a reward, we would expect them to choose at random between previously useful tokens. This design was based on the majority of the other ICTs conducted with children, with the delay between encoding and choice. It should be noted that in previous ICTs with apes and corvids subject first choose and then face a delay and a need to transport the selected item. We chose the former design because we wanted to isolate the impact of including the additional distractor and draw close comparisons with the previous work with children.

Performance in the *item choice* phase improved across age categories (Table 1, Figure 1), with younger children not choosing significantly better than chance according to a binomial test, where chance likelihood to take the correct token is 1/3 (3-year olds, observed =1/11, p=0.988; 4-year olds, observed=14/44, p=0.639) and older children choosing the correct token significantly above chance levels (5-year-olds, observed =39/61, p<0.001, 6-year-olds observed = 34/49 p<0.001, and 7-year-olds, observed = 36/47, p = <0.001). Three-year-olds showed a high initial dropout rate and after 11 were tested we stopped recruiting them. We did not analyse their data further owing to the small sample size.

To evaluate episodic memory for the encoding event, we asked the children a *memory question* after they had chosen their token, namely whether or not they could remember the colour of the box on the table. Performance improved with age (Table 1). As this was an open-ended question it is feasible that children could have responded with any colour, or that they did not know. However, given that children only had experience of two box colours performance was assessed relative to a chance level of 50%: 5-, 6- and 7-year-olds (all ps <0.001) but not 4-year-olds (p= 0.639) responded significantly above chance according to a binomial test (Table 1). Atance &

Sommerville [12] found that memory for the encoding event was an important factor in determining whether or not children succeeded on a test battery of ICTs, such that when they controlled for memory performance, there was no longer an effect of age on item choice. This could suggest a link between memory and planning in development, though Atance & Sommerville stressed that a positive association is difficult to interpret as memory of the specifics of the to-be-planned for event is a prerequisite for success, regardless of the underlying cognitive mechanisms. Performance on the memory question for 5-7-year-olds was near perfect (only 5 out of 157 children were incorrect) and so examining a relationship between item choice and memory performance was not meaningful, though it is notable that in spite of this good memory performance, 48 children chose the wrong token. The performance of 4-year-olds was much more variable, with 12 of the 25 children who got the memory question correct choosing the right item, compared to only 2 of the 19 children that got the memory question wrong. This association between the two measures was significant (Fisher's Exact Test $p=0.0103$). However, considering only the 25 4-year-olds that got the memory question right, four-year-olds still performed at chance level on the choice phase (binomial test $p=0.092$). Whether or not there is an association between memory performance and item choice in 4-year-olds remains uncertain from these results, not least because the memory question is asked after token choice, which might be biasing responses. This issue is examined in more detail in Experiment 3.

We included a further performance measure inspired by the comparative literature in which subjects are evaluated on their propensity to spontaneously transport a necessary tool to the point of use [17]. Children needed to be encouraged to come back to the first context after choosing their token. We therefore did not evaluate whether or not they would spontaneously transport the token next door as in the case of the animal work, but rather we gave children a 30s period to use their chosen token on the box before prompting them to do so: when the child entered the room with the box the experimenter made eye contact with the child and gave an encouraging nod before busying herself with papers. Interestingly, levels of *spontaneous use* were lower than item choice, and increased with age (Table 1). There was a significant relationship between token choice and spontaneous use for 4, 5, 6, and 7-year-olds (FET, all $p<.01$), with those that chose the correct token being more likely to use it spontaneously (Figure 2). This could indicate that spontaneous use is a fruitful measure for future work on planning in children, though these preliminary results should be interpreted with caution. The children could see the box when this measure was taken, which could lead to higher levels of use by children with the right token. A further caveat is that negative results on this measure could have several causes, including a need for more explicit permission to approach the box.

Finally, children were given a *knowledge probe*: the 3 tokens from the choice phase were placed in front of the box for them to choose between to get a final sticker. Children performed well above chance levels on this knowledge probe from the age of 4, indicating that they remembered the details of the training (Table 1).

Experiment 2

The chance-level item choice performance of 4-year-olds in Experiment 1, in contrast to their good performance on previous ICTs, could reflect a reliance on assigning associative value to useful objects and a failure to imagine the specific configuration that they could expect in the future. However, their choices were not split between the two previously useful tokens, but rather they chose at chance between the three options, making this null result difficult to interpret. It would be instructive to know how they would perform in this paradigm if there was only one option with any previous utility, making it more similar to previous ICTs. In Experiment 2 we therefore gave 4-year-olds the same training as in Experiment 1 (namely to operate the 2 different boxes with 2 different tokens), but critically, at the time of choice they were presented with only one of these previously useful tokens (the *correct* token), alongside a token from the bank of 7 distractors used during training that did not operate either box (the *familiar* distractor), and a *novel* token. We tested a further 20 4-year-olds (Table 1). If associative strength influences choice, approximately two thirds of children should choose the correct token in Experiment 2, corresponding to the proportion that chose one of the two previously useful tokens in Experiment 1. However, if the chance performance of four-year-olds was due to the complexity of the training phase leading them to become confused or to forget the critical information needed to plan, they should continue to choose at chance.

We found that children performed significantly above chance level, in contrast to their performance in Experiment 1 (Table 1, Figure 1, binomial, test prop = .33, observed =14/20, $p<.001$). This success rate is comparable to that previously seen in the literature of ICTs in 4-year-olds [10-15, cf 16]. This difference in performance depending only on the inclusion (Experiment 1) or not (Experiment 2) of the associate distractor at the time of choice indicates that, as hypothesised, associative memory for previously assigned value or salience of an object is a critical factor for success in this kind of task. When associative memory was sufficient for success, four-year-olds passed, when it was not, they did not.

Interestingly, children answered the memory question more accurately in this variation of the experiment. All 20 four-year-olds remembered the colour of the box that would be accessible in the final part of the test, whereas only 57% of four-year-olds remembered the colour of the box in Experiment 1. As in the Atance & Sommerville test battery [12], in Experiment 1 there was a trend for memory performance to predict item choice in four-year-olds, which could be interpreted as supporting the notion that memory and planning develop in parallel in childhood: if they remember the encoding event well, they can plan for the next event, but when they do not, they can't. But why would children have remembered the encoding event better in Experiment 2 than in Experiment 1, when everything about the training situation was the same? As in the Atance & Sommerville test battery children were always asked about their memory for the identity of the task *after* they had chosen their item. The contrasting results from Experiments 1 and 2 could indicate that children's responses to memory questions are being influenced by their item choice. In Experiment 1, choosing the associate distractor could have cued the alternative box, and indeed most of the children who chose the associate distractor answered the memory question incorrectly (Figure 3). In Experiment 2 when there was no associate distractor, children's performance on the memory question was very good, though by extension of the above argument, their performance could have been assisted by choosing the token associated not only with a positive outcome but also the correct box. It remains unclear whether or not children would be able to remember the colour of the box on the table next door if they had no cue from the tokens. The status of 4-year-olds ability to remember the encoding event is important in interpreting our results, because if, when asked about it first, 4-year-olds cannot remember which box is accessible following the encoding event, it would be rational for them to split their choices between the 2 previously useful tokens, whereas if they can remember the identity of the box, the reason for them not choosing the right item would be more complicated. In Experiment 3 we therefore conducted a third variation in which the memory question about box colour was asked *prior* to item choice.

Experiment 3

A further 20 four-year-olds were tested in the same way as Experiment 1, with the exception that after completing the vocabulary test they were first asked the *memory question* about which box was on the table. Our chief focus in this experiment was to examine memory performance, but we nevertheless gave the children the choice between the correct token, the associate distractor and a novel distractor, to see how their choice would be affected by first answering the memory question. However, it should be noted that because children were first prompted to remember the identity of box on the table before item choice, they did not need to imagine what they would likely encounter next, so in this case the Item Choice measure would not be considered a test of future thinking.

We found that children performed significantly above chance level on the memory question when it was asked prior to item choice: 95% of them knew which box was available to them (binomial, observed = 19/20, $p<0.001$). Interestingly, despite having just provided this information, only 50% of children went on to pass the item choice measure (Table 1, Figure 1), suggesting that an intact memory is not sufficient for successful choice. A binomial test revealed that this performance was not statistically different from chance, though there was a trend for them to perform well, (binomial, test prop = .33, observed = 10/20, $p = 0.092$). Performance was slightly higher than in Experiment 1, but not as high as in Experiment 2, where we were able to detect above chance levels of responding with the same sample size. Closer examination of performance of 4-year-olds in Experiment 1 (memory question after choice) and Experiment 3 (memory question before choice) reveals that the principle difference between these groups concerns children that chose the wrong token (Figure 3). In Experiment 1, the majority of the children that chose the associate distractor responded incorrectly to the memory question with the box that was associated with that token. However, in Experiment 3 almost all of children (including those that went on to choose the wrong token) answered the memory question correctly. This result is consistent with other findings that suggest episodic memory at age four is fragile and can be disrupted by intervening semantic or associative information if it is in conflict with the past reality [24, 25]. Our suggestion is therefore that children could in principle have remembered the identity of the box in Experiment 1 before they chose the token – however, whether they in actual fact remembered correctly and then nevertheless chose at chance (which in turn corrupted their answer to the memory question), or simply failed to try to remember the identity of the box at all, is a question for future work.

Our results suggest that positive association can support performance on ICTs, which could lead to false positives if the test is being used as a measure of planning. When only one item had associative value, a significant number of four-year-olds chose that correct item, but when two of the items had associative value but only one had future utility, four-year-olds chose randomly. By five-years of age, children’s performance was above chance on this more stringent test. Future work on planning should ensure that explanations based on associative strength or cuing are carefully controlled for. Our results have clear implications for work on the evolution of future planning, because tasks that have been conducted to date in animals [17-19] do not fully control for success by positive association: subjects could have succeeded without imagining the future, but instead by selecting the object with associative value, or something similar to it [7, 8, 21, 26]. Our test, which does not involve verbal framing, or tool-use, would be suitable for adoption with a wide range of animal species. Nevertheless it should be noted that the comparative versions of the ICT impose challenges that the developmental versions do not [27], such as the need to retain and transport the selected tool, which our preliminary results from the spontaneous use measure suggest may be challenging for young children. In previous work with children, some studies have attempted to make success by association less likely by training children with a tool that has a different shape to the one they will need to solve the next problem (e.g. a square tool at training, and a triangle needed in the future [13, 14]), or by only presenting the problem during encoding, without describing the solution [12]. However, it remains possible that at test, children recognise the value of the target based on their past exposure to the problem and select it on that basis (for example, if they identified the object that they would need during encoding, and then recognised it at test as something they wanted), rather than by imagining the future event in which it will be useful. The current results substantiate the plausibility of lean alternatives over the rich interpretations, and so highlight the need for cautious analyses. From a wider theoretical perspective, our findings could have implications for theories that see episodic memory and episodic future thinking as being part of a single, recently evolved system [6, 28]. Our results are in line with other recent findings suggesting that future planning may emerge later than episodic memory over human development [29-32]. Previous evidence for episodic future planning in four-year-olds has been mixed, with much of the positive evidence coming from ICTs. In other future-oriented tasks, as in this study, children do not show competence until the age of five or later [25, 29-32]. While this difference in the age of emergence does not preclude a common cognitive mechanism underpinning both episodic future thinking episodic memory, it may indicate that there are significant unique components to planning, such as imagination, independent goal-setting, temporal representation, and self-control [33].

Author Contributions

All authors jointly conceived of the study, KLD designed the study materials and conducted the research, all authors contributed to analysis and manuscript preparation.

Acknowledgements

The authors would like to thank Francesca Sorrell and Paige Onouye for assistance with data collection, Keith Haynes in the St Andrews workshop for building the apparatus. KLD was funded by a Bobby Jones scholarship and by the University of St Andrews. We are grateful to Dundee Science Centre and RZSS Edinburgh Zoo for the opportunity to recruit and test children.

Declaration of Interests

The authors declare no competing interests

References

1. Szpunar, K.K., Watson, J.M., and McDermott, K.B. (2007). Neural substrates of envisioning the future. Proceedings of the National Academy of Sciences of the United States of America *104*, 642-647.
2. Atance, C.M., and O'Neill, D.K. (2001). Episodic future thinking. Volume 5. pp. 533-539.
3. Addis, D.R., Wong, A.T., and Schacter, D.L. (2008). Age-related changes in the episodic simulation of future events. Psychological science *19*, 33-41.

4. Tulving, E. (2005). Episodic Memory and Autonoesis: Uniquely Human? In The Missing Link in Cognition: Origins of Self-Reflective Consciousness, H.S. Terrace and J. Metcalfe, eds. (Oxford: Oxford University Press), pp. 3-56.
5. Suddendorf, T., and Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. Genetic Social and General Psychology Monographs *123*, 133--167.
6. Suddendorf, T., and Corballis, M.C. (2007). The Evolution of Foresight: What Is Mental Time Travel, and Is It Unique to Humans? Behavioral and Brain Sciences *30*, 299-313.
7. Suddendorf, T., Corballis, M.C., and Collier-Baker, E. (2009). How great is great ape foresight? Animal Cognition *12*, 751-754.
8. Shettleworth, S.J. (2010). Clever animals and killjoy explanations in comparative psychology. Trends in Cognitive Sciences *14*, 477--481.
9. Clayton, N.S., Bussey, T.J., and Dickinson, A. (2003). Can Animals Recall the Past and Plan for the Future? Nature Reviews Neuroscience *4*, 685-691.
10. Suddendorf, T., and Busby, J. (2005). Making Decisions With the Future in Mind: Developmental and Comparative Identification of Mental Time Travel. Learning and Motivation *36*, 110-125.
11. Scarf, D., Smith, C., and Stuart, M. (2014). A spoon full of studies helps the comparison go down: A comparative analysis of Tulving's spoon test. In Frontiers in Psychology, Volume 5. p. 893.
12. Atance, C.M., and Sommerville, J.A. (2014). Assessing the role of memory in preschoolers' performance on episodic foresight tasks. Memory (Hove, England) *22*, 118--128.
13. Suddendorf, T., and Nielsen, M.a. (2011). Children's capacity to remember a novel problem and to secure its future solution. Developmental Science *14*, 26--33.
14. Redshaw, J., and Suddendorf, T. (2013). Foresight beyond the very next event: four-year-olds can link past and deferred future episodes. Frontiers in psychology *4*, 404.
15. Scarf, D., Gross, J., Colombo, M., and Hayne, H. (2013). To have and to hold: Episodic memory in 3- and 4-year-old children. Developmental Psychobiology *55*, 125--132.
16. Russell, J., Alexis, D., and Clayton, N. (2010). Episodic future thinking in 3- to 5-year-old children: The ability to think of what will be needed from a different point of view. Cognition *114*, 56--71.
17. Mulcahy, N.J., and Call, J. (2006). Apes Save Tools for Future Use. Science *312*, 1038-1040.
18. Kabadayi, C., and Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. Science *357*, 202-204.
19. Osvath, M., and Osvath, H. (2008). Chimpanzee (*Pan troglodytes*) and Orangutan (*Pongo abelii*) Forethought: Self-Control and Pre-Experience in the Face of Future Tool Use. Animal Cognition *11*, 661-674.
20. Martin-Ordas, G. (2017). “First, I will get the marbles.” Children’s foresight abilities in a modified spoon task. Cognitive Development.
21. Redshaw, J., Taylor, A.H., and Suddendorf, T. (2017). Flexible Planning in Ravens? Trends in Cognitive Sciences *21*, 821-822.
22. Martin-Ordas, G., Atance, C.M., and Louw, A. (2012). The role of episodic and semantic memory in episodic foresight. Learning and Motivation *43*, 209--219.

23. Atance, C.M., Louw, A., and Clayton, N.S. (2015). Thinking ahead about where something is needed: New insights about episodic foresight in preschoolers. Journal of Experimental Child Psychology *129*, 98-109.
24. Atance, C.M., Metcalf, J.L., Martin-Ordas, G., and Walker, C.L. (2014). Young children's causal explanations are biased by post-action associative information. Developmental Psychology *50*, 2675-2685.
25. Atance, C.M., and Meltzoff, A.N. (2005). My future self: Young children's ability to anticipate and explain future states. Cognitive Development *20*, 341-361.
26. Suddendorf, T., and Corballis, M.C. (2010). Behavioural evidence for mental time travel in nonhuman animals. Behavioural Brain Research *215*, 292-298.
27. Scarf, D., Smith, C., and Stuart, M. (2014). A spoon full of studies helps the comparison go down: A comparative analysis of Tulving's spoon test. Volume 5. p. 893.
28. Schacter, D.L., and Addis, D.R. (2007). On the Constructive Episodic Simulation of Past and Future Events. Behavioral and Brain Sciences *30*, 331-332.
29. Brinums, M., Imuta, K., and Suddendorf, T. Practicing for the Future: Deliberate Practice in Early Childhood. Child Development, n/a-n/a.
30. Cuevas, K., Rajan, V., Morasch, K.C., and Bell, M.A. (2015). Episodic memory and future thinking during early childhood: Linking the past and future. Developmental Psychobiology.
31. Coughlin, C., Lyons, K.E., and Ghetti, S. (2014). Remembering the past to envision the future in middle childhood: Developmental linkages between prospection and episodic memory. Cognitive Development *30*, 96--110.
32. McColgan, K.L., and McCormack, T. (2008). Searching and planning: Young children's reasoning about past and future event sequences. Child Development *79*, 1477--1497.
33. McCormack, T., and Atance, C.M. (2011). Planning in young children: A review and synthesis. Developmental Review *31*, 1--31.

Figure Legends

Figure 1: Item choice distribution in Experiment 1 across age categories, and in Experiments 2 and 3. Black = correct; Grey = associate; Striped = familiar; White = novel. Asterisks indicate % correct is higher than chance (binomial test, $p < .05$).

Figure 2: Performance on the memory, knowledge probe and spontaneous use measures of Experiment 1 for children that chose the correct or an incorrect token. White = 4-year-olds; Light grey = 5-year-olds; Dark grey = 6-year-olds; Black = 7-year-olds.

Figure 3: Performance on the memory question depending on whether it was asked before or after item choice. Percentage of individuals for each choice category who performed correctly on the memory question (correctly reported the colour of the box on the table) when the memory question was asked after (Experiment 1, white bars) or prior to (Experiment 3, black bars) item choice.

Table 1: The percentage of children in each age category who responded correctly in the *item choice*, *memory question*, *knowledge probe*, and *spontaneous use* phases of Experiments 1, 2 and 3. In Experiments 1 and 3 there was a choice between the correct, associate and novel tokens, in Experiment 2 the choice was between correct, familiar and novel tokens. In Experiment 3 the memory question was asked before item choice, in Experiments 1 and 2 item choice came first. Asterisks indicate performance higher than chance (binomial test, $p < .05$).

	Age Group	n (males)	Mean age in months (StD)	Item Choice (% correct)	Memory Question (% correct)	Knowledge Probe (% correct)	Spontaneous Use (%)
Experiment 1	3	11 (8)	41 (4.5)	9	18	55	9
	4	44 (22)	55 (3.6)	32	57	80*	20
	5	61 (30)	66 (3.6)	64*	95*	95*	31
	6	49 (23)	77 (3.7)	69*	100*	94*	37
	7	47 (23)	89 (3.3)	77*	96*	96*	70
Experiment 2 No Associate	4	20 (11)	53 (3.7)	70*	100*	95*	15

Experiment 3							
Memory Question 1st	4	20 (9)	54 (3.4)	50	95*	90*	25

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

"Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Amanda Seed ams18@st-andrews.ac.uk

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The experimental group for Experiment 1 consisted of 220 children. Eight children did not reach criterion (performing six correct activations in a row on the boxes), and were therefore excluded from the experimental analysis (ages in years, months: 3,1 | 3,2 | 3,7 | 3,7 | 3.9 | 4,7 | 6,1 | 6,3), leaving 212 children in the analysis (see Table 1 for age and gender information). While 3-year-olds were initially included in the experimental group, most members of this age group could not complete the training phase so we stopped recruiting them. Children were recruited at Edinburgh Zoo. In Experiments 2 and 3 we tested a total of 40 4-year-olds (see Table 1 for age and gender information). No children were dropped from the study. Children were recruited from the Edinburgh Zoo and Dundee Science Centre. Visitors were approached and informed about the study prior to being asked to join and written consent was required from parents prior to participation. The study had ethical approval from the University of St Andrews ethics committee.

METHOD DETAILS

Materials and Apparatus

Participants were invited into a sectioned and covered area within the visitor attraction containing a table and two chairs. This area was placed inside of a walled tent so that it was visually isolated from another set of table and chairs, where the vocabulary test and item choice took place (see Figure S1). The puzzle boxes, measuring approximately 40cm x 25cm x25cm, contained a revolving dispenser that could be operated discretely by the experimenter by remote control. These rectangular boxes were visually distinct, in both colour (red vs. blue) and shape (sharp vs. rounded edges, respectively) (Figure S2a). Tokens measured approximately two inches in diameter and were distinct in both shape and colour (Figure S2b). There were ten possible tokens, two of which were assigned as the operational token for one of the boxes and one of which was excluded from the training phase so that it was novel at the time of item choice. Their roles were fully counterbalanced across participants.

Procedure

Children were invited to sit at the table, at which point the experimenter placed one of two puzzle boxes in front of the child. The experimenter first inserted a token is into the opening of the puzzle box and dispensed a sticker reward. An envelope was provided to each child to so that they could gather their rewards. Participants were then given the same token and allowed to copy the experimenter in order to obtain another sticker. After this, three tokens were placed in front of the children, one of which was the correct token to operate the box. The functional tokens were fully counterbalanced across participants. The other tokens were chosen at random from a stock of seven (excluding two functional tokens and one kept back as the final novel token during the item choice phase). The experimenter informed the child to choose one and "try for another sticker". Once the participant chose, the remaining tokens were removed from the table as the child attempted to activate the box. This phase ended once children successfully activated the box a total of three times (did not have to be consecutive). The same procedure was then repeated on the other box. After a total of three successful activations on this box (again non-consecutive), the next phase of training began. At this point the participant had to consecutively choose the right token three times on the first box. After three consecutive activations the boxes were switched over. The same rule applied to the next box. This continued until children had activated each box three times without a mistake, for a total of six correct choices in a row. We switched between the boxes up to six times before ending the training phase regardless of whether or not the participant had reached criterion. At this point, if criterion had been reached, it was determined that the child knew which token was required for each box. Choosing six correct tokens in a row was considered reaching criterion. 4 3-year-olds, 2 4-year-olds, 1 5-year-old and 1 6-year-old did not reach criterion.

Children were then instructed that they should leave their stickers on the table, because they would return to get them later. The experimenter then either drew the attention of the child to the box that was being removed from use: stating that they "can't play with that one anymore", or to the other: which, they were told, would "stay here on the table and you can play with it before you leave". At that point they would show them the other box, meaning that attention was either drawn last to the box that they could not play with or the one remaining on the table. The colour of the box left on the table was counterbalanced across participants, in addition to the box that they looked at last. This was done to ensure that participants did not always last look at the box that they could return to. We chose to do this in order to avoid either primacy or recency effects, whereby the first or last thing seen is the first to be recalled.

The experimenter then escorted the child into another area, where they performed the BPVS-III. In keeping with the standards of this diagnostic test, this portion of the experiment took approximately seven minutes. The table with the boxes was not visually accessible to the participants at this time. At the end of the test, children were offered the choice between one of three tokens. One token was the useful token for the box that would still be accessible to them. This was considered to be the *correct* choice. Another token was the one that could be used to activate the box that they could no longer use (the *associate* distractor). The final token was a novel token that they had never seen before or used in the training phase (the *novel* distractor). Children were told that they were "going to go back to get their stickers" and that they could "pick one of these to take with them." They were all then asked the memory question – 'Do you remember which colour box was on the table in the other room?' After the child had responded, they were taken directly back to the box and the experimenter waited to see if they would try to use the token. The experimenter acted preoccupied for approximately 30 seconds before making eye contact with the child and giving a firm nod. If the child used the token immediately or asked for permission to do so, we considered this to be evidence of successful transport and use, if not, they were prompted to transport and use the token. If they made an incorrect choice, we noted whether or not they tried to use this token on the box. All children who made an incorrect choice, or made a correct choice but did not use the token, were then given the opportunity to choose, as in the training phase, between one of the three tokens to use on the box. This constituted the knowledge probe. Regardless of their choices, all children were allowed to operate the box once more before the end of the experiment.

QUANTIFICATION AND STATISTICAL ANALYSIS

Our analyses were conducted online using vassarstats.net. We performed one-tailed exact binomial tests to compare the number of children that performed correctly to chance on the Item Choice measure (Table 1, Figure 1) the memory question (Table 1), and the knowledge probe (Table 1). Additionally, we performed Fisher's Exact Tests to relate performance on the memory question and spontaneous use measure with whether or not participants picked the correct token (in a 2x2 contingency table with alpha set at 0.05). For the sake of this analysis we grouped children who picked either the associate or the novel token into the same category.

DATA AVAILABILITY

The data set containing individual data from experiments 1-3 is available at figshare.com.
DOI: 10.6084/m9.figshare.6236645