Cognitive Development 52 (2019) 100820

Contents lists available at ScienceDirect

Cognitive Development

journal homepage: www.elsevier.com/locate/cogdev

# Examining the limits of memory-guided planning in 3- and 4-year olds[⋆]

Tashauna L. Blankenship[*], Melissa M. Kibbe

*Department of Psychological and Brain Sciences and Center for Systems Neuroscience, Boston University, Boston, United States*

ARTICLE INFO

*Keywords:*
Long-term memory
Retrieval
Planning
Cognitive development

ABSTRACT

Stored memories may be drawn upon when accomplishing goals. In two experiments, we investigated limits on the ability to use episodic memories to support planning in 3- and 4-year-old children. We designed a new memory-guided planning task that required children to both retrieve memories and apply those memories to accomplish multiple, nested goals. We manipulated the difficulty of the task by varying the number of steps required to achieve the goals, and examined the impact of this manipulation on both memory retrieval and planning. We found that, overall, 4-year-olds outperformed 3-year-olds, but as task difficulty increased, all children made more errors. Analysis of these errors suggested that retrieval and planning processes might impose separate limits on memory-guided planning in early childhood, but that these limits may ease across early childhood.

Memory may be divided into multiple systems depending on its primary function. One subdivision is long-term memory, in which memories are stored over time and can be retrieved long after they are encoded (Tulving, 1972). In addition to functioning as storage for experiences and knowledge, long-term memories may be retrieved and used to support acquisition of new knowledge or planning for future events (Atance & O'Neill, 2001; Kliegel, McDaniel, & Einstein, 2000). For example, in order to plan for their upcoming birthday party, a child might remember previous parties they attended and then use those memories to make decisions about the steps they need to take and the order in which they need to take them (e.g. buy invitations, give those invitations to friends, decide how many party favors to get based who RSVPs, etc.). Thus, effectively using long-term memory for planning requires sampling from multiple relevant memories, selecting only the most relevant memories needed to accomplish a particular goal, and then executing the needed actions correctly to bring the plan into fruition (Szpunar, Spreng, & Schacter, 2014), all while inhibiting irrelevant memories or actions (e.g. Klein, Robertson, & Delton, 2011; Suddendorf, Nielsen, & Von Gehlen, 2011). Memory-guided planning therefore taps into more complex cognitions than memory retrieval alone, including working memory and inhibitory control (Blankenship, Broomell, & Bell, 2018; McCormack & Atance, 2011).

Research suggests that children as young as four years of age may engage in memory-guided planning. In one study, Suddendorf et al. (2011) examined children's ability to use episodic memories to plan for the future. They introduced 36-month-old and 48-month-old children to a puzzle box with a large triangular opening. They then showed them that inserting a triangle key activated the box. Children were later introduced to a new box with a square opening, moved to a separate room, and prompted to select a key

[⋆] This research was supported by grant F32HD094554 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) awarded to Tashauna L. Blankenship. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NICHD or the National Institutes of Health. We are grateful to the families for their participation in our research and to our research teams for their assistance with participant recruitment, data collection and coding.

[*] Corresponding author at: Department of Psychological and Brain Sciences, Boston University, Boston MA 02215, United States.
*E-mail address:* shaunalb@bu.edu (T.L. Blankenship).

https://doi.org/10.1016/j.cogdev.2019.100820
Received 19 February 2019; Received in revised form 20 September 2019; Accepted 24 September 2019
Available online 09 October 2019

from a set of differently shaped keys, one being a square. This task required children to retrieve relevant episodic memories (remembering that a triangle key opens the box with a triangular opening) and to use that information to plan for a novel future problem (picking the novel square key in anticipation to open the box with a square opening). They found that 48-month-olds selected the correct key, while 36-month-olds had more difficulty, suggesting developmental change in the ability to use episodic memory to support future-oriented planning across the fourth year of life. However, because their task tapped future-oriented thinking as well, to succeed children also needed to project themselves into the future and to generalize previously acquired knowledge to new problem scenarios, which may have posed additional challenges for younger children.

What are the sources of developmental change in memory-guided planning? One way to gain insight into this question is to explore the "capacity" of memory-guided planning. Examining the "capacity" of a memory system is common in the study of working memory (e.g., Cowan, 2016; Kibbe, 2015; Simmering, 2012), a construct that likely supports memory-guided planning since it enables concurrent maintenance and manipulation of remembered information (Marsh & Hicks, 1998). In studies of working memory capacity, the upper bounds of memory are investigated by increasing the amount of information that subjects must maintain or manipulate, and measuring the point at which subjects begin to make errors. The types of errors that subjects make can yield insights into how memory failed, and why. For example, in a serial recall task, subjects whose working memory limits are exceeded may remember the items in the list, but may fail to remember their correct order (Mathy & Varré, 2013), yielding insights into the structure of the memory system.

How is memory-guided planning limited in young children? On the one hand, memory-guided planning could be limited by the extent to which long-term memories can be stored and/or retrieved. Previous work has found age-related differences during episodic memory retrieval tasks that do not require planning across the fourth year of life (Scarf, Gross, Colombo, & Hayne, 2013), a period in which memory-guided planning processes (like episodic future thinking; Suddendorf et al., 2011) are thought to emerge. However, since the retrieval of long-term memories and the *application* of relevant memories to achieve a goal likely engage different cognitive processes, retrieval and planning may impose different limitations. Indeed, planning processes may impose greater limits, especially in early childhood, since the executive functions that support planning are also undergoing significant development during this period (e.g. Zelazo et al., 2003).

One study found limitations on planning using a task that required retrieval of semantic memories to achieve multiple-step goals. Prabhakar and Hudson (2014) compared 3- and 4-year olds performance on a two-step semantic memory-guided planning task in which children were presented with a model neighborhood and asked to choose which locations in the neighborhood they should visit to accomplish two nested goals. This task required both retrieval of relevant semantic knowledge and application of that knowledge in a particular order (i.e. purchase a birthday present, and then go to a birthday party). The demands of the task were manipulated by increasing the number of distractor locations. Results suggested that 4-year olds were able to successfully complete the task when presented with up to four distractor locations, but 3-year olds struggled when task demands were high. This study suggests that preschoolers can complete a semantic memory-guided planning task requiring two steps, and that planning using semantic memory may be limited by the extent to which children can inhibit distractors, suggesting a potential limiting role for the cognitive processes that support planning (i.e. executive functions).

In the present series of experiments, we investigated limits on memory-guided planning in 3- and 4-year-old children using a new goal-directed episodic memory task tapping into both memory retrieval and planning. Episodic memory is a form of long-term memory that involves storage and retrieval of detail-rich and context dependent memories (Tulving, 2002; Tulving, 1972). Our task was similar to Prabhakar and Hudson's (2014) semantic-guided planning task in that we presented children with nested goals (i.e., goals that must be completed in correct order) to elicit planning, and we used the same spatial context for encoding and test phases. In our task, however, children encoded episodic memories (i.e. novel item-context associations), and were then asked to apply those episodes to accomplish novel, multi-step nested goals.

We chose to focus on episodic memory for several reasons. First, the acquisition of episodic memory involves memory encoding within the experimental context, which allowed us to ensure that all children had the same experience (and the same acquired knowledge) going into our planning task. Second, the role of episodic memory in future-oriented thinking has garnered significant attention over the past several decades (Schacter, Addis, & Buckner, 2008; Atance & O'Neill, 2001; Schacter, Addis, & Buckner, 2007), yet less is known about limitations on using memory to support planning (either immediately or in the distant future) or how these interactions may change with development. Finally, research suggests that executive functions, including planning, are involved in successful episodic memory retrieval (Blankenship & Bell, 2015; Blankenship, O'Neill, Deater-Deckard, Diana, & Bell, 2016; Ruffman, Rustin, Garnham, & Parkin, 2001; Troyer, Graves, & Cullum, 1994), yet there have been no studies explicitly investigating how episodic memory may support or guide planning.

In our task, children were shown a box with four drawers embedded in it. During an episodic memory formation phase, they were taught that each of the drawers could be opened by taking one of four unique actions on objects on top of the box (e.g. that spinning a cylinder opened the drawer second from the right). Since each action/outcome association was learned within the experiment itself, required conscious understanding of specific action/outcome associations, elicited multiple unique memory traces (e.g., how to open baited drawers), and was specific to the experimental context, we could ensure that the encoded memories were episodic (and not semantic or procedural; Tulving, 1987) in nature. Children were then introduced to two animal characters, Monkey and Lion, and told that each had a favorite color, providing them with a rule to follow (i.e., Monkey = green, Lion = red). Finally, to examine memory-guided planning, we baited multiple drawers with beads in the animals' favorite colors and asked children to retrieve beads for the animals in a specific order. Thus, to succeed at the task, children needed to 1) identify which beads needed to be retrieved and in which order, 2) retrieve the relevant episodes (which action would open the baited drawers) to acquire those beads within the novel testing context, and 3) apply these episodes to take the correct actions in the correct order.

Across two experiments, children were presented with multiple trials in which they were asked to retrieve two beads (Experiment 1) or two, three, or four beads (Experiment 2) in a given order. Thus, each test trial represented a novel memory-guided planning problem: the beads could be hidden in any of the four drawers, and which animal's bead needed to be retrieved first varied across trials. To succeed on each trial, children had to construct a new plan using episodic memories acquired during the memory-formation phase of the task.

We had three primary aims. First, we wanted to investigate whether children could perform a memory-guided planning task involving more than a single step. Previous work suggested that children of this age could perform episodic future thinking tasks involving one step (Scarf et al., 2013; Suddendorf et al., 2011), and semantic memory-guided planning tasks involving two steps (Prabhakar & Hudson, 2014), yet it is currently unknown whether children can retrieve multiple relevant *episodic* memories in parallel and deploy them strategically in service of a goal. To this end, in Experiment 1, we baited two drawers and asked children to retrieve one for each animal character in a particular order (e.g. first for Lion, and then for Monkey), requiring retrieval of two relevant episodes and two-step planning. We hypothesized that both 3- and 4-year old children would be able to successfully complete goals requiring up to two actions, but that 4-year olds would be better able than 3-year olds to both retrieve the relevant episodes and execute them in the correct order, coordinating both memory retrieval and planning process (Atance & Jackson, 2009; Coughlin, Robins, & Ghetti, 2017; Prabhakar & Hudson, 2014; Scarf et al., 2013).

Second, we aimed to investigate whether there were limitations on the number of goals children could effectively plan for and execute. In Experiment 2, we parametrically increased the number of steps required to complete the memory-guided planning task, asking children to retrieve two, three, or four beads for the animals in a particular order. We predicted that, as the number of required steps increased, children's performance would decrease. However, we did not have strong predictions about age effects between 3- and 4-year olds. We reasoned that both age groups' performance was likely to decline with increasing steps, but we did not have expectations about whether performance would decline at the same rate.

Lastly, across both experiments, we wanted to investigate potential sources of limitations on memory-guided planning. Experiment 1 required children to retrieve up to two beads and Experiment 2 required children to retrieve up to four beads. In order to successfully achieve these goals, children had to both retrieve episodic information that was presented to them within the context of the task (episodic retrieval), and *use* this information in service of a novel goal (planning). We hypothesized that memory retrieval and planning may impose different limitations on memory-guided planning as the complexity of the task increased. However, since planning guided by episodic memory is dependent on retrieval (Szpunar et al., 2014), teasing apart the contributions of these processes to memory-guided planning performance is somewhat difficult. To begin to investigate this, we used a method more typically used to examine limitations in serial recall working memory tasks (see, e.g., Mathy & Varré, 2013; McCormack, Brown, Vousden, & Henson, 2000) and analyzed the types of errors children made when they were unable to perform the correct actions in the correct order. Specifically, we looked for three patterns of erroneous responses that could yield insights into which processes were contributing to children's difficulty. If children selected the correct actions but in the wrong order, this may suggest that they retrieved the relevant episodes, but were unable to effectively deploy them, potentially suggesting limitations imposed by planning, but not memory retrieval. If children selected only the first action(s) in the sequence but failed to complete the sequence or select any further actions, this may suggest that children retrieved and correctly deployed only some of the relevant episodes, potentially suggesting limitations on both memory retrieval and planning. If children engaged in other, non-systematic errors (e.g. selecting incorrect actions), this may suggest that performance breaks down considerably as task complexity increases. We also investigated whether there were developmental differences in the patterns of errors made by 3- and 4-year-olds, in order to begin to gain insights into sources of developmental change in memory-guided planning.

## 1. Experiment 1

### 1.1. Method

#### 1.1.1. Participants

Thirty-two children (16 3-year-olds, $M = 41.08$, $SD = 3.27$, 11 girls; and 16 4-year-olds, $M = 53.33$, $SD = 4.38$, 7 girls) participated (see Supplement for additional demographic details). Sample size was determined based on similar studies with this age range (Atance & Meltzoff, 2005; Hayne & Imuta, 2011; Russell, Alexis, & Clayton, 2010; Scarf et al., 2013). Post hoc power analyses suggested that we had sufficient power to detect all of our effects of interest ($> .80$; G*Power; Faul, Erdfelder, Lang, & Buchner, 2007). Three additional children participated but were excluded from analyses due to experimenter error.

#### 1.1.2. Materials

The stimuli included 1 neutral (wooden) 1 cm bead, 20 2 cm plastic beads (10 red and 10 yellow), two 16 cm stuffed animals (a monkey and a lion), two clear plastic cups, and a black foam core "magic box" (52 cm x 17.5 cm x 17.5 cm). The magic box had four 5 cm x 5 cm transparent drawers in its front and five unique "action objects" on its top, each of which could be uniquely manipulated (a small pink tube that could be lifted in and out of a larger pink tube; a yellow cylinder that could be rolled with the palm; a green accordion tube that could be bent back and forth; an orange bead that could be spun with the finger tips; and a magnetic green pompom that could be lifted an replaced; see Fig. 1).

#### 1.1.3. Procedure

The procedure used in this study was approved by [anonymized] Institutional Review Board (IRB). Children were seated at a

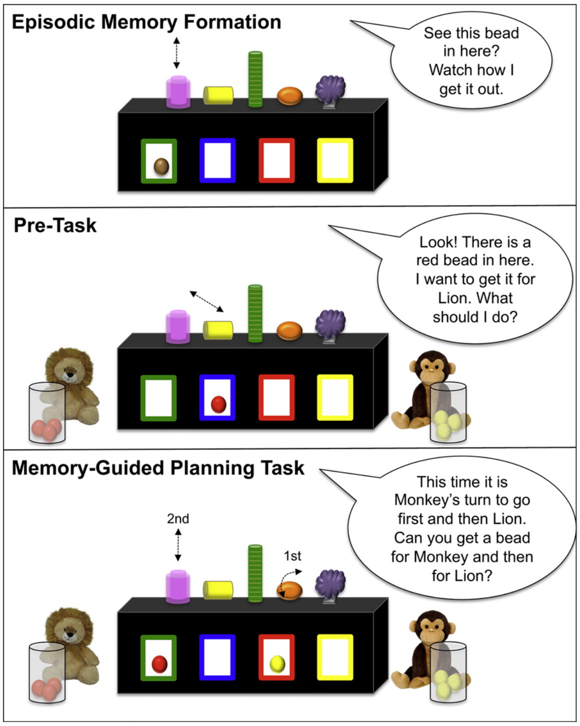


**Fig. 1.** Schematic representations of the three phases from Experiment 1. Top panel shows an example demonstration from the Episodic Memory Formation phase. Middle panel shows an example demonstration from the Pre-Task phase. Bottom panel shows an example 2- action trial from the Memory-Guided Planning task phase. Arrows represent correct actions, and numbers indicate the correct order of actions.

small table across from the experimenter, with their caregiver seated in a chair behind them. The study session was recorded and stored for coding purposes. The task was separated into three phases: Episodic Memory Formation, Task Introduction, and Memory-Guided Planning Task.

*1.1.3.1. Episodic memory formation phase.* During this phase, the experimenter showed children the magic box and told them that each drawer could be opened by manipulating an action object roughly located above the drawer. To demonstrate each action-drawer pair, the experimenter placed a plain wooden bead inside one of the drawers (see Fig. 1, top panel). She then said, "See this bead in here (pointing to one of the drawers)? Watch how I get it out." She demonstrated the relevant action (e.g. rolling the cylinder), opened the drawer, and retrieved the bead. To confirm that children had learned the action-drawer pairing, after each demonstration, the experimenter baited the same drawer with the wooden bead. She then asked children to select the action that would open the baited drawer by saying, "Look! There's a bead in here. I want to get it out. What should I do?" Children were then given the opportunity to perform the relevant action themselves. To be counted as correct, children needed to manipulate the action object as closely as possible to the way in which the experimenter manipulated it (e.g. children had to lift and return the pink cylinder; touching it was not sufficient). If children performed the wrong action, the experimenter said "That one? That one doesn't work. Watch this! See this bead in here…" and then repeated the demonstration.

All children were taught that four of the five action objects were each uniquely associated with opening a specific drawer, resulting in four action-drawer pairs[1] . The experimenter demonstrated each action-drawer pair starting from one side of the box and proceeding to the other side, and starting side was counterbalanced across children. Prior to conducting the experiment, we capped the number of repeated demonstrations per action-drawer pair at four; however, no children reached this criterion. Of the 32 participants, 20 children answered correctly when prompted the first time for each of the four action-drawer pairs, nine children needed 1 of the 4 demonstrations repeated, one child needed 2 of the 4 demonstrations repeated, one child needed 1 of the 4 demonstrations repeated twice, and one child needed 3 of the 4 demonstrations repeated.

*1.1.3.2. Pre-task phase.* This phase was designed to introduce children to the goals that they would later use during the memory-guided planning task. Children were shown the two stuffed animals, Lion and Monkey. Children were then told that each animal had a favorite color (red or yellow; see Fig. 1, bottom panel). The experimenter then placed a single colored bead in one of the drawers,

[1] The fifth action object was an irrelevant distractor object. We included this object on the apparatus in anticipation of experiment 2, in which the number of actions children were asked to take was parametrically increased, in order to make it more challenging for children to use simpler strategies such as manipulating each action object regardless of whether a given drawer was baited.

pointed to the bead, and asked children to help her get the bead for Lion or Monkey by saying, "Look there is a *yellow* (*red*) bead in here. I want to get it for Monkey (Lion) to put in her cup. Do you think you can help me? What should I do?" This phase introduced children to each animal's favorite color and the task of retrieving colored beads for each animal.

After the bead was retrieved, the experimenter placed it in a transparent cup in front of the relevant animal. These beads remained in the cups in front of each animal during later memory-guided planning trials in order to minimize memory demands (e.g. remembering whose favorite color was whose) not directly related to the memory-guided planning task itself (see Memory-Guided Planning Task, below).

Children were asked to retrieve a total of 4 beads (2 red, 2 yellow), one at a time, two beads for each animal and one for each drawer. As in the Episodic Memory Formation phase, children were given feedback as to whether they were correct, and if incorrect, the experimenter re-demonstrated the correct action, and then re-baited the drawer and invited children to try again. Thus, Task Introduction also served to reinforce the memories gained from Episodic Memory Formation phase. As in the Episodic Memory Formation phase, the number of allowed re-demonstrations was capped at four, but no children reached that criterion. Of the 32 participants, 23 children responded correctly when prompted the first time for each of the four action-drawer pairs, eight children needed 1 of the four pairs re-demonstrated, and one child needed 2 of the four pairs re-demonstrated.

*1.1.3.3. Memory-guided planning task*. The memory-guided Planning task consisted of 8 trials. On each trial, the experimenter baited two of the drawers, one with a yellow bead and one with a red bead.

The first two trials were 1-action trials. The primary purpose of these one-step trials was to give children some experience with the memory-guided planning task instructions and demands before they encountered the critical multistep trials. After the experimenter baited the drawers, the experimenter asked children to retrieve a bead *either* for Monkey *or* for Lion by saying, "Can you get Monkey's (Lion's) favorite color bead?" To succeed, children needed to identify which of the two beads belonged to the requested animal, recall which action would be required to open only that drawer (requiring retrieval of relevant memories), and then select *only* that action to retrieve that bead (requiring planning).

The remaining six trials were the critical 2-action trials. Before beginning these trials, the experimenter told children "Now, Lion and Monkey are going to take turns!" On each of the six trials, the experimenter baited two drawers, and then asked children to retrieve the beads in a specific order by saying "This time it's Lion's (Monkey's) turn to go first. Can you get a bead for Lion (Monkey) and then Monkey (Lion)?" To succeed, children had to correctly identify which drawers to open in the correct order (e.g. find the location of Monkey's bead, then find the location of Lion's bead), retrieve the relevant episodic memories required to open only those drawers, and then apply those episodes *in the correct order*, requiring strategic planning to achieve the nested two-step goal (see Fig. 1, bottom panel).

Critically, each trial presented children with a novel goal scenario that they had not previously encountered. The drawers were baited with multiple beads, and these beads could be placed in any two of the four drawers. Further, children were asked to retrieve the beads in a particular order. Together, this results in a completely novel configuration with a completely novel solution on each trial. Furthermore, the experimenter never pointed to any baited drawers or mentioned the color of the beads. Thus, success on these trials required children to develop a novel plan tailored to each trial: they had to interpret the instruction of the experimenter (i.e. Monkey first, Monkey's color is yellow; Lion next, Lion's color is red), identify which beads to target (i.e. yellow bead's location; red bead's location), retrieve the relevant memories to open the drawers to retrieve the beads (i.e. lift pink cylinder to open one drawer; spin orange bead to open other drawer), and then make a plan to deploy those episodes in the correct order (i.e. first lift pink cylinder, and then spin orange bead) and carry out those actions (see Prabhakar & Hudson, 2014, for similar logic in a semantic memory-guided planning task).

After prompting children to retrieve the beads, the experimenter kept her eyes fixed on a neutral point at the back center of the box, which allowed her to see which action(s) children selected while avoiding inadvertently influencing children's choices with her eye gaze. If children selected the correct action(s), the experimenter opened the relevant drawer(s) and retrieved the beads. If children did not select the correct action, the experimenter did not open any drawers. After each trial, regardless of whether children selected the correct actions, the experimenter said, "Let's try another one!" Children received no corrective feedback about whether they were correct or incorrect during Memory-Guided Planning Task trials.

*1.1.4. Coding*

For 1-action trials, children's responses were coded as "Correct" if they engaged in only the requested action. We also coded the types of errors children made when they did not make a correct response, in order to gain insights into the potential strategies children used in the task. Children's erroneous responses were coded as "Superset" responses when children selected the correct action, and then selected the other (unrequested) baited action; as "Incorrect Baited" responses when children selected the other (unrequested) baited action only; and as "Error" responses when children selected one or more incorrect and unbaited actions.

Of primary interest were the critical 2-action trials. For these trials, different patterns of responses can yield insights into children's ability to a) retrieve the relevant episodes, and b) to deploy the information in the correct order. Children's responses were coded as "Correct" only if they engaged in the correct actions *in the correct order,* reflecting both successful episodic memory retrieval and successful planning. We also coded three types of erroneous responses, each of which could potentially reflect different sources of limitation on memory-guided planning. Children's responses were coded as "Subset" responses if they correctly selected only the *first* requested action and then made no other selections, suggesting they retrieved and deployed only one of the two relevant episodes in the correct order. Children's responses were coded as "Swap" responses if they selected the correct actions but in the wrong order, suggesting they engaged in retrieval for both relevant episodes, but were unable to effectively use those episodes to plan. Finally,

children's responses were coded as "Error" responses if they engaged in other, non-systematic errors (e.g. selecting incorrect actions)[2]
.

A primary coder coded children's responses, and a secondary coder also coded a random 25% of the participants (8 children). Agreement between the coders was high (**κ** = .90).

### 1.2. Results

To examine whether children could effectively perform the memory-guided planning task, we first compared children's correct responses to chance for both the 1-action trials and the critical 2-action trials. For 1-action trials, chance performance was computed as the probability of selecting the correct action from among the four action-drawer pairs (1/4 = 0.25)[3] . For 2-action trials, the probability of selecting both actions in the correct order was computed as 0.25 * 0.25 = 0.06[4] .

We computed children's mean proportion Correct responses across 1- and 2-action trials (number Correct / total number of trials). To compare children's mean proportion Correct responses to the relevant chance level, we used one-sample t-tests, two-tailed. To help us interpret any statistically significant results, we also used Bayes Factor analysis (Rouder, Speckman, Sun, Morey, & Iverson, 2009), which allowed us to obtain the odds in favor of the alternative hypothesis – that children's mean proportion correct responses were different from chance – over the null hypothesis – that children's responses were not different from chance. Specifically, because our chance levels are low, we wanted to be able to examine the *strength* of our evidence that children are performing at or above chance. Bayes Factors between 1 and 3 are considered "weak evidence", between 3 and 10 are considered "moderate" evidence, between 10 and 100 are considered "strong" evidence, and greater than 100 are considered "decisive" evidence for the alternative over the null hypothesis (Jeffreys, 1961; Gallistel, 2009). These analyses were conducted using the suggested JZS prior (Rouder et al., 2009) in SPSS.

For 1-action trials, both 3- and 4-year-olds' mean proportion Correct responses were significantly above chance (chance = 0.25; 3-year-olds: $t_{15}$ = 3.31, $p = 0.005$; 4-year-olds: $t_{15}$ = 13.02, $p < 0.001$; see Fig. 2). Bayes Factor analysis yielded "strong" (JZS $BF_{10}$ = 10.12) and "decisive" (JZS $BF_{10}$ = 7,227,999) evidence for the alternative hypothesis for 3- and 4-year-olds' performance, respectively. These results suggest that children can select the relevant episodic memories and deploy them to solve a one-step memory-guided planning problem.

We then examined children's correct responses on the critical 2-action trials. Recall that, to be coded as correct, children needed to choose the correct actions in the correct order, engaging both *memory retrieval* and *planning*. Both 3- and 4-year-olds' correct responses were significantly above chance (chance = 0.06; 3-year-olds: $t_{15}$ = 6.03, $p < 0.001$; 4-year-olds: $t_{15}$ = 13.32, $p < 0.001$; see Fig. 3). Bayes Factor analysis yielded "decisive" evidence for the alternative hypothesis for both age groups (3-year-olds: JZS $BF_{10}$ = 1041.44; 4-year-olds: JZS $BF_{10}$ = 96,969,916). These results suggest that both 3- and 4-year-old children are able to engage in multi-step planning that requires the retrieval of relevant memories and the deployment of these memories in a specific order.

Next, we asked whether children engaged in more correct responses than other types of responses, and whether the frequency of each response type varied as a function of age group. We computed children's mean proportion responses for each Response Type for 1-action trials (Correct, Superset, Incorrect Baited, or Error) and 2-action trials (Correct, Subset, Swap, or Error). We then submitted these to separate 4 (Response Type) X 2 (Age Group) repeated measures ANOVAs, one for each trial type (1-action and 2-action). We followed up significant effects with Bonferroni-corrected pairwise comparisons.

For 1-action trials, we observed a main effect of Response Type ($F(3, 28) = 45.00$, $p < .001$; $\eta_p^2 = .60$): children engaged in more Correct responses than any other response type ($ps < .001$; all JZS $BF_{10} > 2115.00$). There was a Response Type X Age interaction ($F(3, 28) = 8.478$, $p < .001$; $\eta_p^2 = .220$). Four-year olds engaged in more Correct responses than 3-year olds ($p = .004$) and 3 year olds engaged in more Incorrect Baited responses than 4-year olds ($p = .03$), but there were no differences in the frequency of the other Response Types between the age groups (Fig. 2).

For the critical 2-action trials, we observed a significant main effect of Response Type ($F(3, 28) = 49.474$, $p < .001$; $\eta_p^2 = .623$): children engaged in more Correct responses than any other response type ($ps \leq .001$), and in more Error than Subset ($p < .001$) or Swap ($p = .010$) responses (Fig. 3). There was a significant Response Type X Age interaction ($F(3, 28) = 9.318$, $p \leq 0.001$; $\eta_p^2$ = .237). Four-year olds engaged in significantly more correct responses than 3-year olds ($p = .004$), but there were no differences in the frequency of the other Response Types between the age groups (Fig. 3). These results suggest that, while 4-year-olds tended to outperform 3-year-olds, rates of erroneous responses were low overall for both age groups.

[2] Prior to data collection, we planned to examine several different types of errors, informed by previous work on working memory for sequences (Mathy & Varre, 2013). These included substitutions (in which children selected one correct action and one incorrect action), insertions (in which children selected the correct actions and also selected additional, incorrect actions), and complete errors (in which children selected no correct actions). However, since children produced few error responses overall, we collapsed these into a single "Error" category.

[3] Note that there were five potential action objects on top of the box, but only four of these were paired with drawers. Thus, while children could technically engage in five total actions, we reasoned that children would be unlikely to do so, and therefore computed chance out of 4.

[4] This computation of chance allows for the possibility that children could select the same action twice in succession; that is, the probability space for selecting the second correct action is sampled with replacement. We opted for this computation of chance since there was no *a priori* reason to assume children would not do this, particularly as Sequence Length increased (see Experiment 2). Computing chance without replacement does not appreciably change the results in either Experiment (statistically significant results remain significant, non-significant results remain non-significant.)

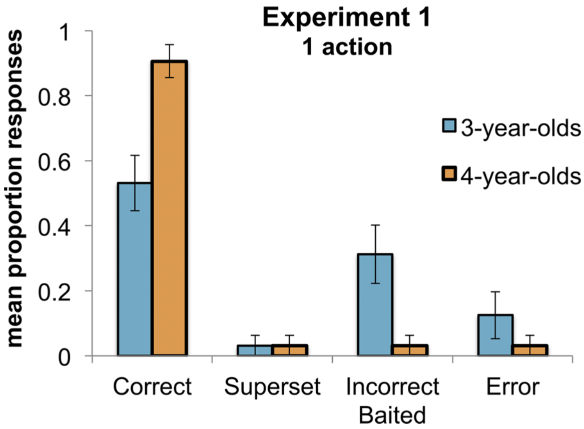


**Fig. 2.** Proportion of responses made across 1-action trials.

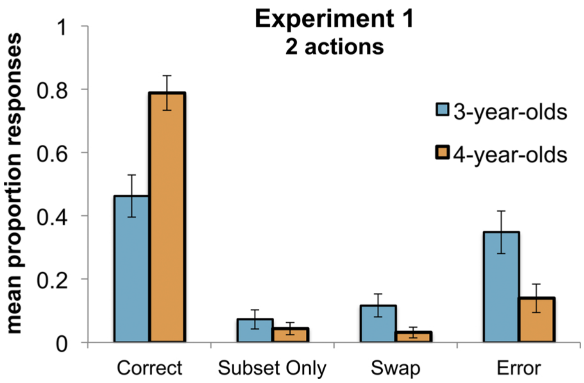


**Fig. 3.** Proportion of responses made across 2-action trials.

### *1.3. Discussion*

In Experiment 1, we examined whether 3- and 4-year old children could complete a goal-directed, multistep memory-guided planning task. Children, regardless of age, were able to do so at levels significantly above chance, suggesting that 3- and 4- year old children are able to engage in memory-guided planning. Further, children could effectively complete a memory-guided planning task requiring at least two steps, consistent with previous work examining multistep semantic memory-guided planning in this same age group (Prabhakar & Hudson, 2014). We also found that 4-year olds performed better than 3-year olds, consistent with previous work showing development of memory retrieval and planning in general during this period (Espy, 1997; Hayne, Gross, McNamee, Fitzgibbon, & Tustin, 2011; Scarf et al., 2013).

We also examined the types of errors children made in order to gain insights into two processes required for memory-guided planning: retrieval of the relevant episodic memories, and application of those memories to achieve a goal(s). We found that, overall, 3- and 4-year olds made few errors. However, when children did make errors, they more often made non-systematic errors (e.g. engaging in incorrect actions) versus subset (engaging in only the first correct action) or swap errors (engaging in both correct actions in the wrong order). Children's pattern of responses in Experiment 1 also suggests that children are not likely to be engaging in a simpler strategy of just selecting actions above baited drawers. If children were using this strategy, we would expect to observe Correct and Swap responses at roughly equivalent rates, since both responses involve selecting only the actions that open the baited drawers. Instead, our results suggest that, at least for our two-step memory-guided planning task, 3–4 year old children are able to successfully engage in memory-guided planning, but when they failed to do so, they displayed difficulty coordinating both retrieval and planning to achieve a nested two-step goal.

In Experiment 2, we sought to replicate the results of Experiment 1, while also parametrically manipulating the number of unique steps required to complete the memory-guided planning task. Specifically, we asked children to execute a memory-guided plan involving 2, 3, or 4 unique actions. We had two goals for Experiment 2. First, we wanted to examine whether we would observe a limit on multi-step memory-guided planning in 3-4-year-old children, similar to limits observed in other memory systems that involve active manipulation of information (e.g. working memory; for reviews see Kibbe, 2015; Cowan, 2016; Simmering & Perone, 2013). Second, we wanted to examine what strategies children use if these limits are exceeded, and see if these strategies may provide insights into developmental differences between memory retrieval and planning abilities. We again examined the types of errors children made as we increased the length of the required sequence of actions. We hypothesized that children would make more errors as sequence length increased, and that these errors may reflect sources of limitation on memory-guided planning, giving insights into whether these limitations were driven by retrieval processes, planning processes, both, or neither.

## 2. Experiment 2

### 2.1. Method

#### 2.1.1. Participants

Thirty-two children (16 3-year-olds, $M = 41.94$, $SD = 2.54$, 6 girls and 16 4-year-olds, $M = 53.34$, $SD = 3.37$, 9 girls; see Supplement for additional demographic information) participated. Post hoc power analyses suggested that we had sufficient power to detect all of our effects of interest (> .90; G*Power; Faul et al., 2007). An additional 7 children participated but were excluded from analyses because of experimenter error (3), task interruption (1), parental interference (1), equipment malfunction (1), or refusal to participate (1).

#### 2.1.2. Materials

The materials were the same as in Experiment 1, except 38 2 cm plastic beads (19 red and 19 green) were used.

#### 2.1.3. Procedure

The IRB protocol was identical to Experiment 1.

*2.1.3.1. Episodic memory formation.* The procedure for Episodic Memory Formation phase was the same as in Experiment 1. As in Experiment 1, the number of demonstrations per action-drawer pair was capped at 4, but no child reached that criterion. Of the 32 participants, 17 children answered correctly when prompted the first time for each of the four action-drawer pairs, 11 children needed 1 of the 4 demonstrations repeated, 1 child needed 1 of the 4 demonstrations repeated twice, 1 child needed 1 of the 4 demonstrations repeated three times, 1 child needed 3 of the 4 demonstrations repeated, and one child needed 1 of the 4 demonstrations repeated three times and 1 of the 4 demonstrations repeated once.

*2.1.3.2. Pre-task phase.* The procedure for the Pre-task phase was the same as in Experiment 1. Also as in Experiment 1, the number of re-demonstrations per action-drawer pair was capped at 4, but no child reached that criterion. Of the 32 participants, 16 children answered correctly when prompted the first time for each of the four action-drawer pairs, 10 children needed 1 of the 4 action-drawer pairs re-demonstrated, 4 children needed 2 of the 4 pairs re-demonstrated, 1 child needed 1 of the 4 pairs re-demonstrated and 1 of the 4 pairs re-demonstrated twice, 1 child needed 1 of the 4 pairs re-demonstrated three times, and 1 child needed 3 of the 4 pairs re-demonstrated.

*2.1.3.3. Memory-guided planning task.* The Memory-Guided Planning Task phase was the same as in Experiment 1, except that the number of actions required to accomplish the goal was parametrically manipulated. Children completed a total of 14 trials. As in Experiment 1, the first 2 trials were 1-action trials. Next, children completed 12 trials in which they were asked to take 2, 3, or 4 actions (4 trials per number of actions). 2-action trials proceeded as in Experiment 1. For 3-action trials, the experimenter baited two drawers with one color bead and one with the other color bead. The experimenter then said, "This time it's Monkey's (Lion's) turn to go first and then Lion (Monkey) and then Monkey (Lion) again. Can you get a bead for Monkey (Lion) and then Lion (Monkey) and then Monkey (Lion) again?" For 4-action trials, the experimenter baited four drawers; two with one color bead and two with the other color, and gave a similar instruction, this time with four steps (Fig. 4). The locations of the beads were varied across trials, resulting in a completely novel goal scenario on each trial.

Trials were blocked and presented in a fixed order: 4 2-action trials, 4 3-action trials, and 4 4-action trials. We used a block design to reduce the cognitive demands of the task by reducing potential confusion about the number of actions required due to interference between trials. We chose a fixed order (easier to harder) to help children engage with the task and to avoid children "giving up" due to task difficulty.

#### 2.1.4. Coding

The coding procedure for Experiment 2 was the same as in Experiment 1 for 1- and 2-action trials. The only difference was the addition of 3- and 4-action trials, which were coded the same as the 2-action trials (i.e., Correct, Subset, Swap, and Errors). Again, data were coded by a primary coder, and a random subset (25%; 8 participants) of the data was coded by a secondary coder. Agreement between the coders was high ($\kappa = .93$).

### 2.2. Results

#### 2.2.1. Comparing children's performance to chance

For 1-action and 2-action trials, chance performance was computed as in Experiment 1. To compute chance for 3- and 4-action trials, we took into account that these trials included multiple drawers baited with the same color bead. For 3-action trials, two drawers were baited with one color bead, and one drawer was baited with another color, and children were asked to retrieve a bead for one animal, then the other, then the first animal again. Thus, for the first of the three actions in the requested series, children

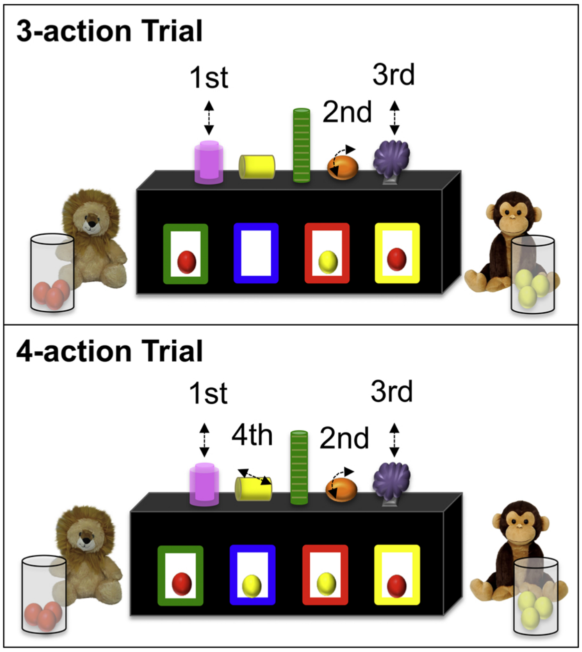


**Fig. 4.** Example 3-action (top panel) and 4-action (bottom panel) memory-guided planning task trials from Experiment 2. In both of these example trials, children were told that it was Lion's turn to go first. The arrows and numbers each represent one potential series of correct actions.

could select one of two action-drawer pairs. We therefore computed the probability of choosing the correct first action as 2/4 = .5, the probability of choosing the second correct action as 1/4 = .25, and the probability of choosing the third correct action as 1/4 = .25. The probability of engaging in this series of actions in the correct order is given by .5 * .25 * .25 = .03. For 4-action trials, in which the drawers were baited with two beads of each color, the probability of choosing the correct actions in the correct order is given by .5 * .5 * .25 * .25 = .02.

For each trial type (1-action, 2-action, 3-action, and 4-action), we compared 3- and 4-year-old children's mean proportion correct responses (number of correct responses/total number of trials, averaged across participants) to the relevant chance level using two-tailed one-sample t-tests and Bayes Factor analysis. To correct for multiple tests, we set our alpha criterion for significance at .01 for the t-tests. The results of these comparisons are presented in Table 1. Both age groups' Correct responses were significantly above chance for 1- and 2-action trials, with Bayes factor analysis yielding decisive support for the alternative hypothesis in both age groups, consistent with the results of Experiment 1. For 3-action and 4-action trials, only 4-year-olds were above chance. For 4-year-olds, Bayes factor analysis yielded strong support for the alternative hypothesis for 3-action trials, and moderate support for the alternative hypothesis for 4-action trials. For 3-year-olds, Bayes factor suggested weaker support for the alternative for 3- and 4-action trials.

### *2.2.2. Response type comparisons*

Next, as in Experiment 1, we asked whether children engaged in more Correct responses than other types of responses, and whether the frequency of each response type varied as a function of age group. We conducted separate repeated-measures ANOVAs for each Sequence Length, with Response Type as a within subjects factor and Age Group as a between subjects factor. Main effects were followed up with Bonferroni-corrected pairwise comparisons.

**Table 1**
Results of two-tailed one-sample t-tests comparing 3- and 4-year-old children's mean proportion Correct responses for each Sequence Length to chance. To correct for multiple comparisons within each age group, alpha was set at .01. P-values that reached this criterion are marked with asterisks. JZS Bayes Factors show the odds of the alternative hypothesis (that children's mean proportion Correct responses are different that would be expected by chance) over the null hypothesis. DF = 15 for each t-test.

| | | *3-year-olds* | | | *4-year-olds* | | |
|---|---|---|---|---|---|---|---|
| ***Sequence Length*** | ***Chance*** | *t* | *p* | *JZS* $BF_{10}$ | *t* | *p* | *JZS* $BF_{10}$ |
| 1-action | .25 | 5.48 | < .001* | 424.71 | 3.67 | .002* | 18.83 |
| 2-action | .06 | 5.76 | < .001* | 676.95 | 7.58 | < .001* | 11,134.83 |
| 3-action | .03 | 2.65 | .02 | 3.33 | 4.08 | .001* | 39.02 |
| 4-action | .02 | 2.70 | .02 | 3.66 | 3.19 | .006* | 8.21 |

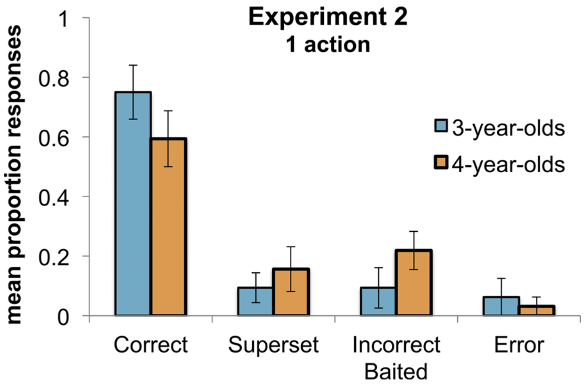


**Fig. 5.** Mean proportion of responses in 1-action trials of Experiment 2.

For 1-action trials, we observed a main effect of Response Type ($F(3, 28) = 25.033$, $p < .001$; $\eta_p^2 = .455$): children engaged in more Correct responses than any other response type (*ps* < .001; all JZS $BF_{10} > 1250.00$; see Fig. 5), consistent with Experiment 1. There was no Response Type X Age interaction ($F(3, 28) = 1.154$, $p = .326$; $\eta_p^2 = .037$), suggesting that children's pattern of responses were similar across both age groups.

Next we analyzed the critical 2-, 3, and 4-action trials. For 2-action trials, we observed a significant main effect of Response Type ($F(3, 28) = 25.240$, $p < .001$; $\eta_p^2 = .457$): children engaged in more correct responses than subset ($p < .001$), swap ($p < .001$) or error responses ($p = .009$), and in slightly more errors than subset responses ($p = .024$; see Fig. 6). There was no Response Type X Age interaction ($F(3, 28) = .966$, $p = .375$; $\eta_p^2 = .031$); this pattern held for both age groups.

For 3-action trials, we also observed a significant main effect of Response Type ($F(3, 28) = 6.924$, $p \leq .001$; $\eta_p^2 = .188$), with children engaging in more correct than subset responses ($p \leq .001$), in more swaps than subset responses ($p < .001$), and in more error than subset responses ($p < .001$; see Fig. 6). The frequencies of Correct and Swap responses were not significantly different, suggesting that, for 3-action trials, children engaged in the correct actions in the correct order and engaged in the correct actions but

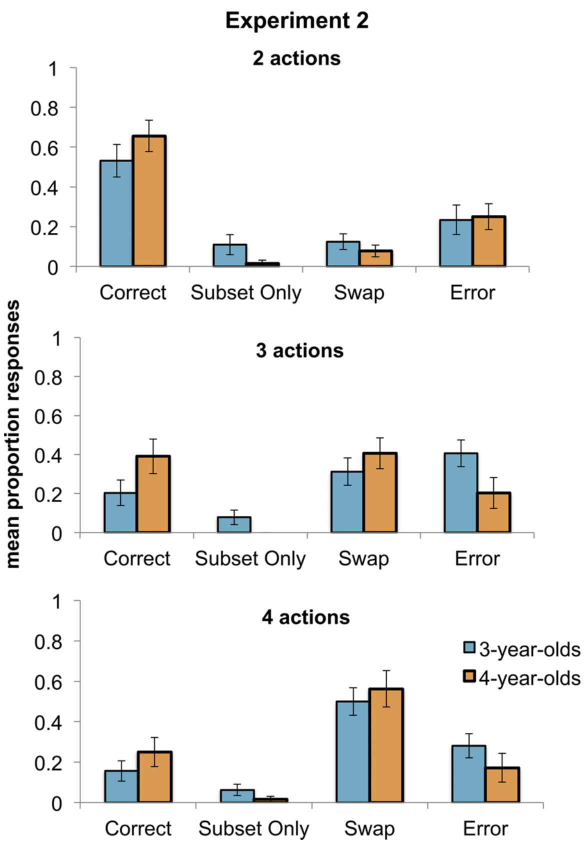


**Fig. 6.** Proportion mean responses for all response types for 2-, 3-, and 4-action trials in Experiment 2.

*not* in the correct order at roughly equivalent rates. The Response Type X Age interaction was not significant ($F(3, 28) = 2.564$, $p = .08$; $\eta_p^2 = .079$); although 3-year-olds' proportion correct responses were not different from chance while 4-year-olds' proportion correct responses were significantly above chance (Table 1), there was no significant difference in Correct responses between the age groups ($p = .10$), suggesting 4-year-olds' above-chance performance for a three-step memory-guided planning task may reflect an ability that is still emerging.

For 4-action trials, there was also a significant main effect of Response Type ($F(3, 28) = 16.748$, $p < .001$; $\eta_p^2 = .358$) driven primarily by children engaging in more Swap than Correct ($p = .004$), Subset ($p < .001$), or Error responses ($p = .016$). Children also engaged in more Correct than Subset responses ($p = .013$; See Fig. 6). Again, there was no Response Type X Age interaction ($F(3, 28) = .890$, $p = .417$; $\eta_p^2 = .029$), suggesting that the frequency of response types was similar across 3- and 4-year-olds.

### *2.2.3. Comparison of children's responses as a function of sequence length*

In the next series of analyses, we examined the impact of Sequence Length on children's responses in the memory-guided planning task, and whether 3- and 4-year-olds children's patterns of responses differed as Sequence Length increased. For these analyses, we were interested only in the critical 2-, 3-, and 4-action trials, in which sequence length was parametrically manipulated and children had the potential to engage in four types of responses: Correct, Subset, Swap, and other Error.

We predicted that, as Sequence Length increased, Correct responses would decrease, and that this would be true for both 3- and 4-year-olds. However, we did not have strong predictions about how Sequence Length would impact the other Response Types. If children are able to successfully engage in retrieval of the relevant episodes, but have difficulty planning, or executing them in the correct order, as Sequence Length increases, we may observe Swap responses increasing as Sequence Length Increases. Alternatively, as Sequence Length increases, children may use a strategy of executing only a subset of the correct actions in the correct order, akin to recalling only some of the items on a to-be-remembered list. If this is the case, we may observe Subset responses increase as Sequence Length increases. It is also possible that, as Sequence Length increases, children may make more non-systematic Errors; that is, if their capacity for memory-guided planning is exceeded, they may respond randomly, in which case Error responses would increase with Sequence Length. Finally, we also may observe age differences in these potential patterns, which would suggest that children's strategies as the memory-guided planning task becomes more challenging might undergo developmental change between 3 and 4 years of age.

To examine these possibilities, we analyzed the proportion of children's responses coded as each Response Type using four 3 (Sequence Length: 2-action, 3-action, or 4-action) X 2 (Age Group: 3-year-olds or 4-year-olds) repeated measures ANOVAs. Where relevant, significant effects were followed up with Bonferroni-corrected pairwise comparisons.

### *2.2.4. Correct responses*

We observed a significant main effect of Sequence Length ($F(2, 29) = 24.489$, $p < .001$; $\eta_p^2 = .449$); children engaged in more Correct responses on 2-action trials versus both 3- and 4-action trials ($ps < .001$), with no significant difference between 3- and 4-action trials ($p = .232$). There was no main effect of Age ($F(1, 30) = 2.876$, $p = .10$; $\eta_p^2 = .087$), and no Sequence Length X Age interaction ($F(2, 29) = .335$, $p = .716$; $\eta_p^2 = .011$), suggesting that this pattern held for both age groups. While 4-year-olds' Correct responses were above chance across Sequence Length, the frequency of their Correct responses decreased significantly as the planning task became more challenging.

### *2.2.5. Swap responses*

We observed a significant main effect of Sequence Length ($F(2, 29) = 23.160$, $p < .001$; $\eta_p^2 = .436$); children engaged in fewer Swap responses on 2-action trials than on both 3- and 4-action trials ($ps < .001$), the inverse of the pattern observed for Correct Responses. There was no main effect of Age ($F(1, 30) = .391$, $p = .54$; $\eta_p^2 = .013$), and no Sequence Length X Age interaction ($F(2, 29) = .675$, $p = .50$; $\eta_p^2 = .022$), suggesting that children's frequency of engaging in Swap responses as Sequence Length increased did not vary as a function of Age.

### *2.2.6. Subset responses*

We observed no main effect of Sequence Length ($F(2, 29) = .888$, $p = .385$; $\eta^2 = .029$), children's frequency of engaging in Subset responses did not vary as a function of Sequence Length. We did observe a small main effect of Age ($F(1, 30) = 4.401$, $p = .04$; $\eta_p^2 = .128$), but no Sequence Length X Age interaction ($F(2, 29) = .691$, $p = .458$; $\eta_p^2 = .023$). Overall, 3-year-olds engaged in more subset responses than 4-year-olds.

### *2.2.7. Error responses*

We observed no main effects of Sequence Length ($F(2, 29) = 1.04$, $p = .36$; $\eta_p^2 = .033$) or Age ($F(1, 30) = 1.83$, $p = .19$; $\eta_p^2 = .057$), and no Sequence Length X Age interaction ($F(2, 29) = 1.83$, $p = .17$; $\eta_p^2 = .058$), suggesting children's non-systematic error responses were similar regardless of sequence length or age of participant.

### *2.2.8. Comparison of Experiments 1 and 2*

The procedures of Experiments 1 and 2 were identical up to the 6$^{th}$ memory-guided planning task trial, allowing the 1- and 2-action memory-guided planning task trials of Experiment 2 to serve as a replication of Experiment 1. We therefore asked whether children's performance on Experiment 1 was replicated in Experiment 2. To do so, we conducted two repeated-measures 4 (Response Type) X 2 (Age Group) X 2 (Experiment) repeated measures ANOVAs, one for 1-action trials and one for 2-action trials. Since

Experiment 1 included six 2-action trials while Experiment 2 included only four of these trials, we analyzed only the first four 2-action trials from Experiment 1. These were followed up, where relevant, with Bonferroni-corrected pairwise comparisons.

For 1-action trials, we observed a main effect of Response Type ($F(3, 58) = 65.662, p < .001; \eta_p^2 = .523$): children engaged in more Correct responses than any other response type ($ps < .001$), consistent with the individual results of Experiments 1 and 2. There was no Response Type X Age interaction ($F(3, 58) = 1.393, p = .251; \eta_p^2 = .023$) nor Response Type X Experiment interaction ($F(3, 58) = 0.741, p = .501; \eta_p^2 = .012$). There was a Response Type X Age X Experiment interaction ($F(3, 58) = 6.907, p < .001; \eta_p^2 = .103$): 4-year olds engaged in more Correct responses in Experiment 1 than in Experiment 2 ($p = .028$)

For the critical 2-action trials, we observed a significant main effect of Response Type ($F(3, 58) = 58.155, p < .001; \eta_p^2 = .492$); children engaged in more Correct responses than any other response type ($ps < .001$), consistent with the individual results of Experiments 1 and 2. Children also engaged in more Error than Subset or Swap responses ($ps \leq .001$) on these trials. We also observed a significant Response Type X Age interaction ($F(3, 58) = 7.079, p = .002; \eta_p^2 = .186$); 4-year olds engaged in significantly more Correct responses than 3-year olds ($p = .004$). Critically, there was no Experiment X Response Type interaction ($F(3, 58) = .052, p = .941; \eta_p^2 = .001$) and there was no Response Type X Age X Experiment interaction ($F(3, 58) = 2.949, p = .06; \eta_p^2 = .047$): suggesting that children's pattern of responses did not differ across Experiments.

## 3. Discussion

In Experiment 2, we examined whether 3- and 4-year old children would be able to complete our memory-guided planning task when the task required children to plan and execute two or more steps. Our results replicated and extended our findings from Experiment 1. We found that children, regardless of age, were able to complete a two-step memory-guided planning task at above-chance levels, while performance declined as the number of required steps exceeded two. As sequence length increased, 4-year-olds were significantly above chance for sequences requiring up to four unique actions, while 3-year-olds were closer to chance levels. However, 3- and 4-year-olds' performance did not differ significantly, and all children made more errors as sequence length increased, suggesting that effectively executing multi-step memory-guided planning may be an emerging skill.

The types of errors children made yielded some insights into potential sources of limitations on memory-guided planning. As sequence length increased, rates of non-systematic errors remained consistent, potentially reflecting a breakdown of engagement in the task that is not dependent on task difficulty (for example, if children failed to attend to all or part of the experimenter's instruction on that trial, or failed to engage with the task on that trial). Three-year-olds made significantly more Subset errors than 4-year-olds, suggesting that 3-year-olds may be more limited in their ability to execute all the steps of a nested multi-step goal. Finally, both age groups made more Swap errors as Sequence Length increased, potentially reflecting limitations on planning; even when children are able to retrieve the relevant memories required to complete the goal, they may not be able to deploy them in the correct order when the number of nested steps is high.

However, it is important to consider whether the increase in Swap errors as the number of required steps increased might reflect non-memory-guided strategies. Because there was a finite number of potential actions that children could take, as the length of the sequence increased, the potential for random errors also decreased. Specifically, on 4-action trials, when faced with a memory-guided planning problem that was too challenging, children may have resorted to a simpler strategy of selecting all four of the potential actions (for example, engaging in the actions sequentially from right to left). This pattern of performance would be coded as "Swap" under our coding scheme, since children engaged in the correct actions but in the wrong order. However, while this strategy would require children to remember how to engage in each action, and to engage in only the correct number of actions (and not, e.g., 3 or 5), it would not necessarily entail that children were retrieving the information that was relevant for *that particular trial*. This is not a concern for the 3-action trials coded as Swap errors, since using this "select all" strategy during 3-action trials would be coded as an Error under our coding scheme. Thus, it is possible that Swap behaviors during 4-action trials may reflect different underlying strategies than Swap behaviors during 2- and 3-action trials.

To explore this possibility, we further examined children's Swap responses during 4-action trials. We asked what proportion of Swap responses appeared to reflect the "select all" strategy by dividing Swap responses into three categories: Select-all, in which children selected all four actions systematically from right to left or from left to right; Subset + Error, in which children engaged in the first one or two actions correctly, and then responded erroneously; and Random, in which children selected four objects in a non-systematic order. Note that we opted to code Random responses as different from Select-all responses, because it was difficult to determine children's strategies on these trials. For example, children may have been using the "select all" strategy, or they may have made a planning-based error (e.g. selecting the incorrect animal's bead first). We then compared the frequency of these response types across our two age groups (see Supplement for analysis details). We found that children engaged in select all strategies on roughly half of the trials coded as Swap errors, but they did not do so more frequently than the other strategies. This result, combined with our error analyses above, suggests that children may be engaging in a variety of strategies as the task becomes more challenging. Implications for our understanding of limitations on memory-guided planning are discussed in the General Discussion.

## 4. General discussion

In two experiments, we examined 3- and 4-year-olds children's ability to complete a novel task in which children were asked to engage in memory-guided planning in order to accomplish goals with multiple steps. Our task required children to encode episodic memories over the course of a brief experiment and then retrieve those memories in order to accomplish a novel goal requiring two or more steps.

We had three aims. Our first aim was to investigate whether children could perform a memory-guided planning task involving more than a single step. In Experiment 1, we found that 3- and 4-year-olds both could reliably engage in memory-guided planning requiring two steps, which was replicated in children's performance on 2-action trials in Experiment 2. Overall, 4-year-olds performed better than 3-year-olds, although this difference in performance was not robust across Experiments 1 and 2, suggesting variability in memory and planning skills across early childhood. These results are the first, to our knowledge, to examine young children's performance on a multiple-step memory-guided planning task.

Our second aim was to examine whether there were limitations on the number of steps children could effectively plan and execute in a memory-dependent planning task, and to investigate how these limitations may vary as a function of children's age. In Experiment 2, we found that 3- and 4-year-old children could reliably execute plans involving two steps, but performance declined as the number of required steps increased. When the task required three or four steps, 4-year-olds were significantly above chance, while 3-year-olds' performance fell just short of our strict cut-off for statistical significance, yet we found no differences in performance between 3- and 4-year-olds. Indeed, while 4-year-olds were above chance levels, they were still engaging in far fewer correct responses as the number of steps increased.

Our third aim was to investigate sources of limitations on memory-guided planning by attempting to tease apart limitations on memory retrieval and planning abilities. We did so by analyzing the kinds of errors children made to gain insights into the kinds of strategies they used as the memory-guided planning task became more challenging. These analyses yielded somewhat equivocal results. We found that children were more likely to engage in "swap" responses – taking the correct actions in the wrong order - as the number of required steps increased, but rates of subset responses and random errors remained relatively constant across sequence length. These results suggest that children may be able to retrieve the relevant memories, but may struggle to execute a plan using that information. However, further analysis of swap responses painted a more nuanced picture of children's performance. Children sometimes selected the correct actions in random order, sometimes resorted to a "select-all" strategy of choosing the actions in a fixed order (e.g. from left to right), and sometimes engaged in the first one or two actions in the correct order, while randomly selecting the remaining actions. This suggests that, as the task became more complex, children used a variety of strategies in their responses, potentially suggesting that there are individual differences that dictate how children respond when their limits are exceeded (Siegler, 2005).

An additional age effect was found for error types, with 3-year-olds engaging in more subset responses than 4-year-olds. This result may suggest that 3-year-olds are more limited in their ability to retrieve episodic memories (O'Neill & Gopnik, 1991; Scarf et al., 2013). Perhaps, requesting four episodes simultaneously overwhelms their developing episodic memory systems. Another explanation is that 3-year-olds have a harder time maintaining in working memory the required memories, requested goals, or both. For example, if 3-year-olds struggle to remember four requests (e.g., "Can you get a bead for Monkey and then Lion and then Monkey and then Lion again?") they may only retrieve a bead for Monkey and then Lion. This difference in performance may reflect differences in working memory span across this age range (e.g. Simmering, 2012), again suggesting an important limiting role for executive functions during memory-based planning. However, these interpretations should be considered with some caution, as this effect was only present when examining subset responses alone.

Overall, our results suggest that limitations on memory-guided planning may be explained by limitations on planning processes more so than limitations on retrieval of relevant memories. This does not necessarily mean that limitations on retrieval do not contribute to memory-guided planning performance, but that executive functions may contribute more to variation in performance (Blankenship & Bell, 2015; Blankenship et al., 2018; Mahy & Moses, 2011). These results are in agreement with Prabhakar and Hudson (2014), who found that increasing the number of distractors impacted multi-step semantic memory-guided planning, suggesting that executive functioning processes were a limiting factor.

Our results have implications for our understanding of the development of future-oriented thinking in young children. Previous work has suggested that engaging in successful planning through use of episodic or semantic memory requires self-projection into the future (Atance & O'Neill, 2001; Feeney & Roberts, 2012). While our task did not directly measure children's ability to simulate or project into the future, our results suggest an important divergence between retrieval of memories and application of those memories to achieve a future goal which can inform theories of how future-oriented thinking develops across early childhood.

### *4.1. Task-dependent limitations on memory-guided planning*

We observed limitations in children's ability to perform a memory-guided planning task: in our task, that limit appears to reliably be two sequential goals, with some easing of these limits between 3 and 4 years of age. However, it is an open question about whether the limits on memory-guided planning can be *quantified*, as limits on working memory often are (see, e.g., Cowan, 2016; Miller, 1956).

There are several factors of our task that likely influenced the observed memory-guided planning limits. For example, in our task, the number of actions children needed to take was apparent from the number of visible beads. This was by design, in order to reduce moment-to-moment working memory demands, but it could facilitate children using a strategy of selecting the action above the relevant bead (although they still had to remember how to execute the action). Our blocked design also meant that children did not have to switch between numbers of goals across each trial, further reducing the working memory demands of the task (although this meant children may have been more fatigued at higher sequence lengths). Finally, children could learn the simple relationship between animals and colored beads, and use that knowledge to guide planning. However, in real world settings, children are unlikely to receive this kind of scaffolding, and memory-guided planning may be even more limited when working memory demands of the task are increased. Children's patterns of errors could likewise be impacted by the working memory demands of the task. For example, without visual feedback about the number of steps required for the task, children may make more subset errors. Future work would attempt to examine more closely the role of working memory load and other sources of task demands in limiting episodic memory and memory-guided planning across development. Indeed, existing literature suggests that working memory and episodic memory are related in children (Blankenship & Bell, 2015; Blankenship, O'Neill, Ross, & Bell, 2015). However, it is important to note that working memory cannot be the only factor driving success or failure in our memory-guided planning task, since children had to engage and coordinate strategic memory retrieval and planning processes in addition to holding information online during the task.

Our task was designed to examine how episodic memories may be used to support planning. However, long-term memories, it can be argued, exist on a continuum (Szpunar et al., 2014), and complex memories can incorporate semantic, episodic, and/or procedural knowledge. The nature of the long-term memories could potentially impact the limitations imposed on memory-guided planning across development. Semantic information, for example, is not context-bound and is easier to maintain in working memory from very early in development (i.e. Dempster, 1981; Kibbe & Leslie, 2019). Prabhakar and Hudson (2014) examined semantic memory-guided planning and also found limitations on this ability, imposed by the need to suppress distractors in their task. It is possible that the retrieval and application of multiple semantic memories may impose different limitations than the retrieval and application of episodic memories. Procedural memories, that require little conscious cognitive effort, could also impose fewer limitations on planning. Indeed, it is possible that our task may have engaged more than episodic memory. For example, it is possible that experience retrieving the relevant memories during learning resulted in some engagement of procedural memory for the relevant actions. Future work would investigate how the format of long-term memories impacts memory-guided planning, perhaps by manipulating the amount of experience children receive with the relevant actions before the critical planning trials.

It is also important to note that the episodic memories that children acquired in our task were rather arbitrary; there is no clear meaningful relationship between rolling a cylinder and opening a drawer. Again, this was by design, to ensure that the task was tapping episodic rather than semantic memory and that the information was completely novel to children. However, meaningfulness eases demands on episodic memory (e.g. Konkle, Brady, Alvarez, & Oliva, 2010) and working memory (e.g. Kibbe & Feigenson, 2014, 2016; Miller, 1956; Chase & Simon, 1973; Mathy & Feldman, 2012), and meaningfulness could therefore potentially ease observed limits on memory-guided planning. Our lab is currently investigating this possibility.

Given the variety of cognitive processes involved in memory-guided planning, quantifying a capacity of this system is not likely. Rather, our results highlight the importance of systematically examining potential sources of limitation on memory-guided planning. We investigated one such source: the number of steps required to achieve a novel goal. Future work would further examine other sources of limitations across development.

### *4.2. Summary*

In summary, we developed a novel task that required children to engage two processes necessary for memory-guided planning: retrieval of relevant episodic memories and application of those episodes to accomplish a novel goal. We found that both 3- and 4-year-olds could effectively engage in memory-guided planning for goals requiring two steps, but performance declined as the number of steps required for the task increased. 4-year-olds tended to outperform 3-year-olds, suggesting development of this ability in early childhood. Analysis of children's patterns of errors suggested that limitations on multi-step memory-guided planning task may be driven by the cognitive processes that support planning more than memory retrieval.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.cogdev.2019.100820.

## References

Atance, C. M., & Jackson, L. K. (2009). The development and coherence of future-oriented behaviors during the preschool years. *Journal of Experimental Child Psychology, 102*(4), 379–391.

Atance, C. M., & Meltzoff, A. N. (2005). My future self: Young children's ability to anticipate and explain future states. *Cognitive Development, 20*(3), 341–361.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences, 5*(12), 533–539.

Blankenship, T. L., & Bell, M. A. (2015). Frontotemporal coherence and executive functions contribute to episodic memory during middle childhood. *Developmental Neuropsychology, 40*(7-8), 430–444.

Blankenship, T. L., Broomell, A. P., & Bell, M. A. (2018). Semantic future thinking and executive functions at age 4: The moderating role of frontal brain electrical activity. *Developmental Psychobiology, 60*, 608–614.

Blankenship, T. L., O'Neill, M., Deater-Deckard, K., Diana, R. A., & Bell, M. A. (2016). Frontotemporal functional connectivity and executive functions contribute to episodic memory performance. *International Journal of Psychophysiology, 107*, 72–82.

Blankenship, T. L., O'Neill, M., Ross, A., & Bell, M. A. (2015). Working memory and recollection contribute to academic achievement. *Learning and Individual Differences, 43*, 164–169.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55–81.

Coughlin, C., Robins, R. W., & Ghetti, S. (2017). Development of episodic prospection: Factors underlying improvements in middle and late childhood. *Child Development, 90*, 1109–1122.

Cowan, N. (2016). Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on Psychological Science, 11*(2), 239–264.

Dempster, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin, 89*(1), 63.

Espy, K. A. (1997). The shape School: Assessing executive function in preschool children. *Developmental Neuropsychology, 13*(4), 495–499.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Feeney, M. C., & Roberts, W. A. (2012). *Comparative mental time travel: Is there a cognitive divide between humans and animals in episodic memory and planning. The oxford handbook of comparative evolutionary psychology*236–260.
Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439.
Hayne, H., Gross, J., McNamee, S., Fitzgibbon, O., & Tustin, K. (2011). Episodic memory and episodic foresight in 3-and 5-year-old children. *Cognitive Development, 26*(4), 343–355.
Hayne, H., & Imuta, K. (2011). Episodic memory in 3-and 4-year-old children. *Developmental Psychobiology, 53*(3), 317–322.
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Clarendon Press: Oxford: Oxford University Press.
Kibbe, M. M. (2015). Varieties of visual working memory representation in infancy and beyond. *Current Directions in Psychological Science, 24*(6), 433–439.
Kibbe, M. M., & Feigenson, L. (2014). Developmental origins of recoding and decoding in memory. *Cognitive Psychology, 75*, 55–79.
Kibbe, M. M., & Feigenson, L. (2016). Infants use temporal regularities to chunk objects in memory. *Cognition, 146*, 251–263.
Kibbe, M. M., & Leslie, A. M. (2019). Conceptually rich, perceptually sparse: Object representations in 6-month-old infants' working memory. *Psychological Science, 30*(3), 362–375.
Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology General, 139*(3), 558–578.
Kliegel, M., McDaniel, M. A., & Einstein, G. O. (2000). Plan formation, retention, and execution in prospective memory: A new approach and age-related effects. *Memory & Cognition, 28*(6), 1041–1049.
Klein, S. B., Robertson, T. E., & Delton, A. W. (2011). The future-orientation of memory: Planning as a key component mediating the high levels of recall found with survival processing. *Memory, 19*(2), 121–139.
Mahy, C. E., & Moses, L. J. (2011). Executive functioning and prospective memory in young children. *Cognitive Development, 26*(3), 269–281.
Marsh, R. L., & Hicks, J. L. (1998). Event-based prospective memory and executive control of working memory. *Journal of Experimental Psychology Learning, Memory, and Cognition, 24*(2), 336.
Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition, 122*(3), 346–362.
Mathy, F., & Varré, J. S. (2013). Retention-error patterns in complex alphanumeric serial-recall tasks. *Memory, 21*(8), 945–968.
McCormack, T., & Atance, C. M. (2011). Planning in young children: A review and synthesis. *Developmental Review, 31*(1), 1–31.
McCormack, T., Brown, G. D., Vousden, J. I., & Henson, R. N. (2000). Children's serial recall errors: Implications for theories of short-term memory development. *Journal of Experimental Child Psychology, 76*(3), 222–252.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology, 27*(3), 390.
Prabhakar, J., & Hudson, J. A. (2014). The development of future thinking: Young children's ability to construct event sequences to achieve future goals. *Journal of Experimental Child Psychology, 127*, 95–109.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.
Ruffman, T., Rustin, C., Garnham, W., & Parkin, A. J. (2001). Source monitoring and false memories in children: Relation to certainty and executive functioning. *Journal of Experimental Child Psychology, 80*(2), 95–111.
Russell, J., Alexis, D., & 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*(1), 56–71.
Scarf, D., Gross, J., Colombo, M., & Hayne, H. (2013). To have and to hold: Episodic memory in 3-and 4-year-old children. *Developmental Psychobiology, 55*(2), 125–132.
Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience, 8*(9), 657.
Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences, 1124*(1), 39–60.
Siegler, R. S. (2005). Children's learning. *The American Psychologist, 60*(8), 769.
Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology, 111*(4), 695–707.
Simmering, V. R., & Perone, S. (2013). Working memory capacity as a dynamic process. *Frontiers in Psychology, 3*, 567.
Suddendorf, T., Nielsen, M., & Von Gehlen, R. (2011). Children's capacity to remember a novel problem and to secure its future solution. *Developmental Science, 14*(1), 26–33.
Szpunar, K. K., Spreng, R. N., & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences, 111*(52), 18414–18421.
Troyer, A. K., Graves, R. E., & Cullum, C. M. (1994). Executive functioning as a mediator of the relationship between age and episodic memory in healthy aging. *Aging and Cogntion, 1*(1), 45–53.
Tulving, E. (1972). Episodic and semantic memory. *Organization of memory, 1*, 381–403.
Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology, 6*(2), 67–80.
Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1), 1–25.
Zelazo, P. D., Müller, U., Frye, D., Marcovitch, S., Argitis, G., Boseovski, J., et al. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development* i-151.