Journal of Experimental Child Psychology 181 (2019) 1–16

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

Thinking about the future: Comparing children's forced-choice versus "generative" responses in the "spoon test"

Cristina M. Atance [a,*], Seyda Nur Celebi [a], Sarah Mitchinson [b], Caitlin E.V. Mahy [c]

[a] School of Psychology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
[b] Department of Psychology, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada
[c] Department of Psychology, Brock University, St. Catharines, Ontario L2S 3A1, Canada

ARTICLE INFO

Article history:
Received 8 June 2018
Revised 13 December 2018

Keywords:
Development
Cognition
Episodic future thinking
Generativity
spoon test
category fluency

ABSTRACT

One of the most popular methods to assess children's foresight is to present children with a problem (e.g., locked box with no key) in one room and then later, in another room, give them the opportunity to select the item (e.g., key) that will solve it. Whether or not children choose the correct item to bring back to the first room is the dependent measure of interest in this "spoon test." Although children as young as 3 or 4 years typically succeed on this test, whether they would pass a more stringent version in which they must verbally generate (vs. select) the correct item in the absence of any cues is unknown. This is an important point given that humans must often make decisions about the future without being explicitly "prompted" by the future-oriented option. In Experiment 1, using an adapted version of the spoon test, we show that as the "generative" requirements of the task increase, 3-, 4-, and 5-year-olds' ($N = 99$) performance significantly decreases. We replicate this effect in Experiment 2 ($N = 48$ 3-, 4-, and 5-year-olds) and also provide preliminary evidence that the capacity to verbally generate the correct item in a spoon test may draw more heavily on children's category fluency skills than does their capacity to select this item among a set of distracters. Our findings underscore the importance of examining more generative forms of future thought in young children.

* Corresponding author.
E-mail address: atance@uottawa.ca (C.M. Atance).

https://doi.org/10.1016/j.jecp.2018.12.006

Introduction

Humans' capacity to think about the future manifests itself in various ways. For example, when a person lost in thought is asked what he is thinking about, he might respond "my upcoming holiday," or when a work colleague is asked why she brought a screwdriver to work, she might refer to the loose drawer handle on her desk. Indeed, there are numerous times during the day when people think about the future and act with the future in mind. Our capacity to mentally project ourselves into the future—sometimes referred to as *episodic future thinking* (EFT; Atance & O'Neill, 2001)—is a highly adaptive ability that results in behavior that is deliberate and planful rather than immediate and reactive. Indeed, behaviors that are valued in our society, including saving and delaying gratification, would be impossible without some concept of the future and our future selves. Although early in development children's needs are largely taken care of by others, before long they too develop the foresight that characterizes the adults of our species. Much headway has been made during the past few decades to better understand children's future-oriented thought and behavior. These abilities become increasingly important as children gain independence and begin schooling, and thus understanding their emergence and development is critical.

Children's EFT develops rapidly during the preschool years (Atance & Meltzoff, 2005; Atance & Sommerville, 2014; Hudson, Mayhew, & Prabhakar, 2011; Mahy, Grass, Wagner, & Kliegel, 2014; Payne, Taylor, Hayne, & Scarf, 2014; Russell, Alexis, & Clayton, 2010; Suddendorf, Nielsen, & von Gehlen, 2011) and has important implications for everyday functioning in domains such as planning (e.g., McCormack & Atance, 2011), delaying gratification (e.g., Thompson, Barresi, & Moore, 1997), and saving (e.g., Atance, Metcalf, & Thiessen, 2017). Much of what we know about EFT development comes from laboratory studies that have used both verbal and behavioral methods to tap into children's conceptual understanding (for a review, see Atance & Mahy, 2016). In what follows, we focus on behavioral EFT tasks that are modeled after Tulving's (2005) "spoon test." The spoon test was inspired by an Estonian folktale (and by an earlier proposal by Suddendorf, 1994) in which a young girl dreams about going to a friend's party. At this party, there is delicious chocolate pudding being served, but the young girl did not bring her own spoon and so cannot eat any. The next evening, she places a spoon under her pillow to avoid reexperiencing her disappointment. As Tulving (2005) noted, the young girl's behavior illustrates her ability to reflect on her past experience (i.e., episodic memory) and attempt to secure a solution to her problem (i.e., EFT).

This example has now served as the basis for numerous studies assessing EFT/mental time travel in both human children and nonhuman animals (for a review of both populations, see Scarf, Smith, & Stuart, 2014). The earliest and most systematic adaptation of the spoon test with children was carried out by Suddendorf et al. (2011; but see also Suddendorf & Busby, 2005). The basic setup in this study and others is as follows. Children are introduced to a novel problem (e.g., locked box with stickers but no functional key) in one location (e.g., Room A), and after a delay in a different location (e.g., Room B) they are told that they will be returning to the first location. However, before returning, children are presented with a set of items that includes one (e.g., key) that can address the problem. Whether children select the correct item is the dependent variable of interest.

Using this kind of item-choice task, Suddendorf et al. (2011) compared 3- and 4-year-olds' performance in "instant" and "delay" conditions. In both conditions, children were presented with a locked box that could be opened with a triangular key to obtain a desirable object. The box was then removed and replaced by one with a square keyhole but for which no key was available. In the instant condition children were led to the other side of the room and were presented with a set of keys, including a square-shaped key, and asked which one they wanted to take back with them, whereas in the delay condition children waited for 15 min in another room before selecting. Both 3- and 4-year-olds selected the correct key at above-chance levels in the instant condition, but only 4-year-olds did so in the delay condition. Children also received an alternate version of this task in which they were introduced to a puppet, "Ellie the Elephant," who likes to eat bananas but were told that there were no bananas in the room to feed her. Similar to the results of the just-described box task, 3- and 4-year-olds in the instant condition performed significantly above chance, whereas only the 4-year-olds did so in the delay condition. These findings suggest that by age 4 children are able to remember a past problem and show the necessary foresight to later solve it (but see Payne et al., 2014, and Scarf, Gross, Colombo, & Hayne, 2013, for evidence that even 3-year-olds can solve certain kinds of item-choice tasks).

More recently, using clever variants of the spoon test, researchers have shown that children can also plan for a more remote future event. For example, Redshaw and Suddendorf (2013) used the same general approach as Suddendorf et al. (2011) but instituted a 5-min delay between the time when children chose the item and the time when they were given the opportunity to address the problem. Despite the more ‘‘deferred" nature of the task, 4-year-olds still selected the correct item at above-chance levels.

Using a different approach, Martin-Ordas (2018) developed a two-step task in which 3-, 4-, and 5-year-olds went into a room, where they were shown a locked box with marbles and a separate bag with crayons. Then, in a second room, they were shown a marble run game. After visiting both rooms, children were told that they would be returning to them 15 min later. Although children were assigned to three different conditions, the most relevant (and comparable to previous versions of the spoon test) was one in which, after the 15-min delay, children were presented with a key, a coloring sheet, and a block. In this case, the correct choice is the key because it would allow children to first open the box that contained the marbles and then use the marbles in the marble game. (Note that only children who stated that they preferred to play with marble runs, rather than with paper and crayons, were included in this experiment.) Children were also asked in which order they would visit the rooms; in this case, the required sequence is to go to the first room that contained the locked box and then to the second room that contained the marble game. Only the 5-year-olds were significantly above chance in selecting the correct item (i.e., key) to bring back with them to the two rooms, and only children in this age group were significantly above chance in stating that they would visit the rooms in the correct sequence. These results suggest that this task taps into a more challenging aspect of future thinking than the one-step versions of the spoon test and also accurately reflects the kinds of future-oriented problems with which humans are often faced (e.g., first I must get the car keys from the house, and then I can check whether I forgot my mittens in the car).

In sum, a number of developmental studies have been inspired by Tulving's (2005) spoon test and have yielded important insights about children's future-oriented reasoning along with important methodological advances in its study. First and foremost, researchers have converged on a method that is easily administrable and that consistently shows that by 3 or 4 years of age (depending on the particular version of the task given; see also Atance & Mahy, 2016), children are able to select an item that can address a future problem. Second, at least one variant—a two-step spoon test (Martin-Ordas, 2018)—has been implemented in which children need to consider temporal order (i.e., which room they will visit first) when making their item selections. In this case, it is only by 5 years of age that children succeed at above-chance levels.

Based on this body of work, it is tempting to conclude that by 5 years of age (if not younger) children have acquired the kind of foresight so elegantly described in Tulving's (2005) spoon test. Yet, at least one critical aspect of Tulving's spoon test has yet to be adequately translated to developmental tasks. An important difference between Tulving's scenario and existing spoon tests is that in the former the young girl, spontaneously or of her own volition, mentally generates the item (i.e., spoon) that is needed to address the future problem. Indeed, no one lays out for her a set of choices (e.g., bowl, spoon, and glass) from which to choose. In contrast, existing item-choice tasks typically involve a predetermined set of choices—including the correct choice and distracters—from which children must choose. By using such forced-choice options, it is possible that children are simply *recognizing* the item (‘‘A-ha, the spoon is required to eat the pudding") rather than first *recalling* the event in question (as presumably the young girl in Tulving's scenario did) and then mentally generating the required item (i.e., ‘‘spoon").

The notion of ‘‘generating" is important given that Tulving (2005) stated that a central requirement of the spoon test is that the participant ‘‘must *deliberately* engage in behavior" (p. 44, italics added). Yet, arguably, when items are physically presented to children in a forced-choice format, their behavior (e.g., selecting the correct one) is not deliberate in the way that Tulving described (i.e., children are *not* required to seek out or verbally generate the item in question). Equally problematic is that, at this point, the task may be solved by processes not necessarily related to foresight but rather related to associative learning (e.g., Atance, Louw, & Clayton, 2015; Dickerson, Ainge, & Seed, 2018; Russell et al., 2010). Although there has been considerable effort to reduce such associative mapping (e.g., using multiple keys as distracters to reduce the possibility that children merely link "key" with "lock"; e.g., Suddendorf et al., 2011), it is possible that merely seeing the item of need cues children to its connection with the earlier problem.

Similar arguments were recently made by Moffett, Moll, and FitzGibbon (2018), who noted that in existing forced-choice spoon tests "children neither have to seek out nor create the means to an end, but can simply identify the solution among a set of items that are placed right under their nose" (p. 867). To overcome this limitation, these researchers designed a task in which children needed to "create" the means to address a future problem. Their task also involved two rooms, with 4- and 5-year-olds starting in the Little Room, where they drew pictures that they subsequently placed in a backpack. They then traveled to the Big Room, where they played a picture game. However, in this room they learned that a particular picture (e.g., banana) was missing, and thus the game could not be completed. Children then returned to the Little Room, where they had the opportunity to draw the missing picture that would later allow them to complete the game in the Big Room. Children were given three trials that followed this general structure, and the dependent measure of interest was whether children would, in the Little Room, draw the required picture to then transport to the Big Room. Thus, an important difference in this task as compared with previous ones is that children needed to freely recall and construct a needed item (e.g., banana) rather than select (or, as the authors argued, "perceptually recognize") it from a set. Children's success on this task was analyzed by comparing performance in the above-described Future Need condition with that in a No Future Need condition in which the target item (e.g., banana) was present. Although 4-year-olds' mean score for the three trials in the Future Need condition was less than 1 (i.e., ~17% correct) and thus relatively low, it was nonetheless significantly higher than 4-year-olds' performance in the No Future Need condition, suggesting that this age group was showing some sensitivity to future needs. The 5-year-olds' level of task success in the Future Need condition was significantly higher than that of the 4-year-olds (~50% correct) and also higher than the 5-year-olds' performance in the No Future Need condition. As with Martin-Ordas's (2018) task, it is clear that this is a more challenging version of the spoon test and one that better assesses children's ability to more spontaneously generate a future-oriented solution than previous versions.

The current study

In a similar vein, the goal of the current study was to assess 3-, 4-, and 5-year-olds' performance on a novel version of the now-classic spoon test in which they were asked to verbally generate a response rather than select one from a set of predetermined choices. The particular version of the spoon test we used was an adaptation of Suddendorf et al.'s (2011) food task in which children are introduced to an animal character who loves a food that is currently unavailable. We randomly assigned children to one of three conditions. In the first one, the *forced-choice* condition, children learned about an animal (e.g., hippo) who loves a food (e.g., apples) that is unavailable. As in most previous versions of the spoon test, children were presented with a set of items in a second room and needed to select one to bring back to the original room. In contrast, our second condition required that children verbally generate an item from a specific category (e.g., fruit) that would address the problem they had encountered in the first room. The important difference between this condition, the *generate-category* condition, and the forced-choice condition is that we did not present children with three possible options (e.g., apple, lemon, and orange) but rather simply asked them, "What would be a good fruit to bring back with you?" Finally, in our third condition, the *generate-object* condition, children were asked the more open-ended unconstrained question, "What would be a good *thing* to bring back with you?" Thus, implementing these three conditions allowed us to directly compare children's responses between contexts in which they were "cued" to differing extents.

We predicted that children's performance on these tasks would be significantly affected by age, with older children outperforming younger children. This prediction was based on the results of previous studies in which there tends to be an improvement in performance on item-choice tasks between 3 and 5 years of age. We also predicted a linear effect of condition, such that children would perform best in the forced-choice condition and worst in the generate-object condition, with performance in the generate-category condition in between. Here, our logic was that the more "generative" children needed to be (vs. being cued), the worse their performance would be.

Experiment 1

Method

Participants

Our sample consisted of 99 typically developing English-speaking children: 30 3-year-olds (13 boys; M_{age} = 42.10 months, range = 35–47), 36 4-year-olds (11 boys; M_{age} = 55.00 months, range = 48–59), and 33 5-year-olds (18 boys; M_{age} = 65.73 months, range = 60–72). An additional 6 children did not complete the testing session due to uncooperativeness and, thus, were excluded from the sample. Most children were White (73%) and from middle- to upper-middle-class backgrounds (84% of the families reported an income of $75,000 CDN/year or higher).

Item-choice tasks

All children completed two item-choice tasks modeled after Suddendorf et al.'s (2011) food task: a fruit task and a vegetable task. Children were randomly assigned to one of three task conditions: (a) forced-choice (n = 31 [10 3-year-olds, 11 4-year-olds, and 10 5-year-olds], M_{age} = 54.10 months), (b) generate-category (n = 35 [10 3-year-olds, 15 4-year-olds, and 10 5-year-olds], M_{age} = 54.71 months), and (c) generate-object (n = 33 [10 3-year-olds, 10 4-year-olds, and 13 5-year-olds], M_{age} = 55.15 months). The two tasks and each of their three conditions are described below.

Fruit task. In the main testing room (Room A), children were given a plastic toy lemon and introduced to "Terri the Tiger" who loves to eat lemons. The experimenter (E) asked children to feed Terri the lemon. After doing so, children were told that Terri was very full, did not want any more lemons, and was going to take a nap. E then put Terri away/out of sight. After that, children were introduced to "Heather the Hippo" who loves to eat apples but did not have any apples to eat (i.e., "Heather loves to eat apples, but there are no apples here for Heather the Hippo to eat. Too bad!"). E patted the hippo's head while saying "Aww, too bad!" Children were then invited to take a 5-min break in the greeting room (Room B), where they had time for free play. At the end of the break, children were told that it was time to go back to the other room (Room A).

In the *forced-choice* condition, E randomly laid out three choices/plastic toy fruits (i.e., apple, lemon, and orange) in the area outside both rooms and told children, "But first, here are three fruits! Can you tell me what would be a good *fruit* to bring back to the other room?" If children picked the apple, they went into Room A and fed the hippo. If they picked one of the other two fruits (i.e., orange or lemon), they brought it to feed the hippo, and E stated that this was the hippo's second favorite food.

In the *generate-category* condition, E did not provide children with any choices but rather, after telling children that it was time to go back to the other room, stated the following: "But first, I want you to close your eyes and think hard about something. Can you tell me what would be a good *fruit* to bring back to the other room?"

In the *generate-object* condition, again without any choices, children were told, "But first, I want you to close your eyes and think hard about something. Can you tell me what would be a good *thing* to bring back to the other room?"

If children correctly stated "apple" in these generative conditions, E brought out an apple from a hidden drawer in the hallway and they went back into Room A to feed the hippo. If children were unable to verbally generate the correct response, they went back to Room A, where E emphasized that the hippo was still in the room and simply said "Hi, Heather!" without making any reference to the apple. In all three conditions, E also asked children to justify their choice (i.e., "Why would you bring a(n) [type of fruit] to the other room?").

Children received two scores for their performance in the fruit task: one for their item-choice response (0 or 1) and one for their explanation (0 or 1). In the forced-choice condition children received 1 point if they chose the apple, and in the two generative conditions they received 1 point if they stated the word "apple." Any other response received a score of zero (0). For their ensuing explanation, children received 1 point if they provided a future-oriented response (e.g., referred to the hippo and/or Room A) and 0 points if they gave reasons unrelated to the future (e.g., "because I like it").

Vegetable task. The vegetable task was identical to the fruit task except for the puppets involved and what they liked to eat. In this task, children were first introduced to "Bonnie the Bear" who loves to eat cucumbers; children fed her a cucumber, and E said that Bonnie was full. Children were then introduced to "Danny the Dog" who loves to eat carrots but did not have any carrots to eat. In the *forced-choice* condition, E randomly laid out three vegetables (i.e., pepper, cucumber, and carrot) and asked, "What would be a good *vegetable* to bring back to the other room?" In the *generate-category* condition, without giving any choices, children were asked to close their eyes and think about "What would be a good *vegetable* to bring back to the other room?" In the *generate-object* condition, again without any choices, children were asked to close their eyes and think about "What would be a good *thing* to bring back to the other room?" Children received 1 point if they picked the carrot in the forced-choice condition or stated "carrot" in the generative conditions. As in the fruit task, children received two scores: one reflecting the item they chose (0 or 1) and one reflecting their ensuing explanation (0 or 1).

Procedure

Children were tested at an on-campus laboratory ($n = 88$) and a local museum ($n = 11$) (task scores did not differ significantly as a function of testing location), and all sessions were audio- and video-recorded. Children always completed the fruit task first and the vegetable task second. These tasks were separated by approximately 5 min, during which children completed other cognitive tasks that are not reported here. At the end of the session, children received a small prize for their participation. The procedure received ethical approval from the University of Ottawa's Office of Research Ethics and Integrity.

Results and discussion

Table 1 shows means and standard deviations for children's item-choice scores as a function of age and condition. There was no significant effect of task type (i.e., fruit vs. vegetable) on either children's item-choice scores (McNemar test, $p = .804$) or their explanation scores (McNemar test, $p = 1.00$). Thus, we collapsed scores across both tasks into one variable (range = 0–2) for children's item-choice and explanation scores separately. There were also no significant gender differences on either children's item-choice scores, $F(1, 97) = 2.36$, $p = .127$, or their explanation scores, $F(1, 97) = 0.62$, $p = .435$, and so we collapsed across this factor as well.

Effects of age and condition on children's item choices

To determine whether children's ability to select the correct item varied as a function of age or condition, we ran a 3 (Age: 3, 4, or 5 years) × 3 (Condition: forced-choice, generate-category, or generate-object) analysis of variance (ANOVA) with item-choice scores as the dependent variable. This analysis revealed a significant main effect of condition, $F(2, 90) = 19.62$, $p < .001$, $\eta^2_p = .30$, but no significant effect of age, $F(2, 90) = 1.79$, $p = .172$, $\eta^2_p = .04$. The Age × Condition interaction was marginally significant, $F(4, 90) = 2.29$, $p = .066$, $\eta^2_p = .09$ (see Fig. 1). Because we predicted a linear increase in children's performance, with the generate-object condition being the lowest and the forced-choice condition being the highest, we ran a polynomial trend analysis. This analysis yielded a significant linear trend ($p < .001$), such that mean performance was lowest in the generate-object condition ($M = 0.55$, $SD = 0.87$), higher in the generate-category condition ($M = 1.29$, $SD = 0.83$), and highest in the forced-choice condition ($M = 1.71$, $SD = 0.59$). Thus, as predicted, children experienced more difficulty when they were asked to verbally generate (as opposed to select) an item that would be useful in the future.

Table 1
Means and standard deviations for item-choice tasks as a function of age in Experiment 1.

Condition	3-Year-olds	4-Year-olds	5-Year-olds
Forced-choice (range = 0–2)	1.70 (0.67)	1.73 (0.47)	1.70 (0.67)
Generate-category (range = 0–2)	1.10 (0.74)	1.07 (0.96)	1.80 (0.42)
Generate-object (range = 0–2)	1.00 (0.94)	0.20 (0.63)	0.46 (0.88)

Note. Standard deviations are in parentheses.

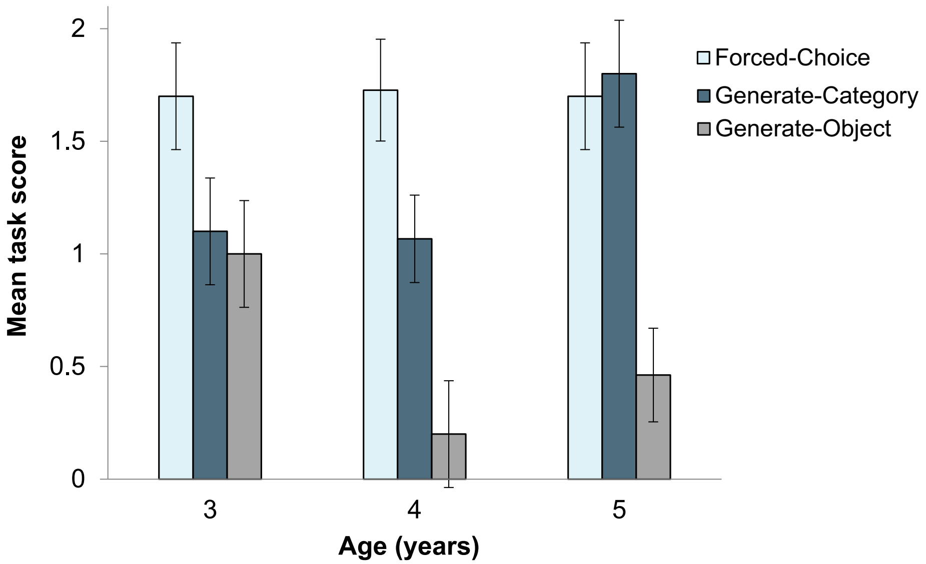

Fig. 1. Mean task scores as a function of age and condition in Experiment 1. Standard errors are represented in the figure by the error bars attached to each column.

Given that our interaction was marginally significant, we ran two series of simple main effects examining age differences at each level of condition (Bonferroni-adjusted alpha level of $p < .016$) and condition differences at each level of age (Bonferroni-adjusted alpha level of $p < .016$). None of the age differences at each level of condition was significant after applying the Bonferroni correction: forced-choice, $F(2, 90) = 0.01$, $p = .993$; generate-category, $F(2, 90) = 3.00$, $p = .042$; generate-object, $F(2, 90) = 2.98$, $p = .056$. Condition differences at the different levels of age were as follows: 3-year-olds, $F(2, 90) = 2.55$, $p = .084$; 4-year-olds, $F(2, 90) = 10.89$, $p < .001$; 5-year-olds, $F(2, 90) = 11.67$, $p < .001$. Pairwise comparisons (again using a Bonferroni-adjusted alpha level of $p < .016$) for the 4-year-olds indicated that they performed significantly better in the forced-choice condition than in the generate-object condition ($p < .001$) and performed better in the generate-category condition than in the generate-object condition ($p = .006$). Although 4-year-olds also performed better in the forced-choice condition than in the generate-category condition ($p = .029$), this was not significant after applying the Bonferroni correction. Pairwise comparisons for the 5-year-olds (also using an adjusted alpha level of $p < .016$) indicated that they performed better in the forced-choice condition than in the generate-object condition ($p < .001$) and performed better in the generate-category condition than in the generate-object condition ($p < .001$). No significant difference was obtained between the forced-choice and generate-category conditions ($p = .766$).

Children's item-choice errors. It is important to address the item-choice errors that children made in each of the three conditions. Children made very few errors (15%) in the forced-choice condition, and of these errors, unsurprisingly, children simply chose one of the two other (incorrect) options presented. In the generate-category condition, children erred 36% of the time, and all of these errors entailed the generation of an incorrect fruit/vegetable. However, the majority of errors (9 of 13) in the vegetable task (which always came after the fruit task) were fruit labels (e.g., apple, lemon, banana), suggesting some interference from the previous fruit task. This was true for all three age groups. For the fruit task (which always came first), children made just as many errors (12 total), but these were mostly the verbal generation of an incorrect fruit, with this being true for each age group. This pattern suggests that it was not solely interference that accounted for children's errors in the generate-category condition.[1] Finally, in the generate-object condition, children erred 73% of the time. Although there was some variability in the kinds of responses that children provided after being asked "What would be a good *thing* to bring back to the other room?" the most common response (accounting for 71% of errors) was to name a toy (e.g., "ball," "marbles," "my toys," "a teddy," "puzzles," "some cars"). This was true 50% of the time for 3-year-olds, 78% of the time for 4-year-olds, and 80% of the time for 5-year-olds. Interestingly, in the vegetable task in this condition, children rarely erred by stating a fruit (2 of 24 total errors), suggesting that interference from the fruit task occurred infrequently.

Verbal explanations. Children's explanation data were relatively straightforward. That is, most children who selected or verbally generated the correct item also referenced the animal's need for that item (e.g., "because the hippo wants one"). This was true 91% of the time in the forced-choice condition, 89% of the time in the generate-category condition, and 100% of the time in the generate-object condition. These data are consistent with results from comparable tasks in the literature showing that children's correct item choices are accompanied by explanations that reference the previously encountered problem (e.g., Suddendorf et al., 2011). However, whether in the forced-choice condition (at least) these explanations genuinely reflect "foresight," as opposed to more of an "ad hoc" or associative explanation, is an issue to which we return in the General Discussion.

In sum, our results suggest that, irrespective of age, children experienced increasing difficulty when asked to verbally generate an item versus select an item (e.g., apple) that would be useful in the future (e.g., feeding a hippo). Moreover, whereas all three age groups were successful in the forced-choice condition, 3- and 4-year-olds (at least) had more difficulty than 5-year-olds in generating a response even when provided with the category (i.e., fruit or vegetable) that corresponded to the item in question (although the age difference in the generate-category condition was not statistically significant after applying the Bonferroni correction). Finally, the pattern of data in our generate-object condition was somewhat unexpected. That is, whereas 5-year-olds had the highest mean performance in the generate-category condition, this advantage was not apparent in the generate-object condition and, in fact, in this condition 3-year-olds had the highest mean score (but again this difference was not statistically significant). Our hunch is that the generate-object condition was simply too open-ended for children, and the fact that they could verbally generate any object whatsoever (as opposed to the more constrained category "fruit" in the generate-category condition) seemed to lead them to identify objects they desired (e.g., toys) rather than focus on the past problem and the need to address it. As such, this particular condition might not be optimal for testing children's foresight. Note that although 5-year-olds performed better ($\sim$50% correct) in Moffett et al.'s (2018) more open-ended drawing task than they did in our (open-ended) generate-object condition ($\sim$25% correct), in the former they were also constrained by what they could draw. Accordingly, this methodology may have better focused them on the task at hand (e.g., needing to draw a banana) and reduced the tendency to generate any item they wished.

Experiment 2

In Experiment 2, we build on the results of Experiment 1 in two ways. First, for the reasons just described, we excluded the generate-object condition from our design and directly compared

[1] We also reran the 3 (Age: 3, 4, or 5 years) × 3 (Condition: forced-choice, generate-category, or generate-object) ANOVA with item-choice score on the fruit task only (which always came before the vegetable task) as the dependent variable, and the results we obtained were virtually identical to the analysis using scores on both tasks. There was a significant main effect of condition, $F(2, 90) = 12.51$, $p < .001$, $\eta_p^2 = .22$, no significant effect of age, $F(2, 90) = 1.38$, $p = .257$, $\eta_p^2 = .03$, and no Age × Condition interaction, $F(4, 90) = 1.49$, $p = .211$, $\eta_p^2 = .062$.

children's performance in the forced-choice and generate-category conditions. Accordingly, we ran a within-participants design in which children were first asked the open-ended generate-category question (i.e., "Can you tell me what would be a good fruit to bring back to the other room?"), immediately followed by the forced-choice question. Here, as in Experiment 1, three items were laid out and children needed to select the one to bring with them to the other room.

The second way in which we build on Experiment 1 is by exploring whether children's performance in the generate-category condition is related to their performance on category fluency and generativity tasks. Category fluency tasks require individuals to list as many instances as they can of a given category (e.g., "foods") in a limited amount of time (e.g., 60 s) and are argued to draw on both semantic memory and executive functioning ability (e.g., Pastor-Cerezuela, Fernández-Andrés, Feo-Álvarez, & González-Sala, 2016). More important, they require the individual to generate a novel response and, in this sense, overlap with what is required of children in our generate-category condition.

Generativity tasks are similar to category fluency tasks but are also argued to rely on divergent thinking ability—often considered to be an aspect of creativity (e.g., Suddendorf & Fletcher-Flinn, 1997, 1999; Ward, 1968). For example, in a generativity task, children are asked to state all the things they can think of that are "red." Similar to the logic of category fluency tasks, those children who can generate a response to what is arguably a less well-defined category (e.g., "all the things that are red" vs. "foods") may be at an advantage in our generate-category condition. Moreover, previous research has found a link between EFT and creativity in both younger and older adults (Addis, Pan, Musicaro, & Schacter, 2016) and in children (Suddendorf, 2010). More specifically, Suddendorf (2010) reported that 3- to 5-year-olds' scores on the kind of generativity task just described were associated with the number of responses that they generated when asked to report events that would happen "tomorrow." In light of this, we were curious about whether those children who performed better in the generate-category condition were also those who had higher scores on the generativity task.

To explore these possible relations, we included two category fluency tasks (Adlam, Patterson, Bozeat, & Hodges, 2010) and one generativity task (Suddendorf & Fletcher-Flinn, 1997). Common to all three tasks is that they assess children's ability to search their memory and generate items that fall into a specific category. As such, we predicted a link between children's performance on these tasks and their performance in the generate-category condition. This is because in the latter children must also draw on their memory of the past and more spontaneously generate an item that will be useful in the future.

Method

Participants

Our sample consisted of 48 typically developing English-speaking children: 16 3-year-olds (6 boys; M_{age} = 41.69 months, range = 36–47), 16 4-year-olds (8 boys; M_{age} = 53.56 months, range = 48–59), and 16 5-year-olds (9 boys; M_{age} = 66.38 months, range = 60–71). An additional 4 children did not complete the testing session due to uncooperativeness and were excluded from the sample. Children were mostly White (77%) and from middle- to upper-middle-class backgrounds (79% of the families reported an income of $75,000 CDN/year or higher). Children were recruited from two participant databases at the University of Ottawa and at Brock University, where the testing was conducted.

Procedure

Children were tested in a laboratory setting, and tasks were administered in the following fixed order: fruit task, category fluency tasks (animal and food), vegetable task, and generativity task. As in Experiment 1, the intervening time between the fruit and vegetable tasks was approximately 5 min. All sessions were audio- and video-recorded, and parents watched their children on a computer screen from an adjoining room. At the end of the session, children received a small gift for their participation. This procedure received ethical approval from the University of Ottawa's Office of Research Ethics and Integrity and Brock University's Research Ethics Board.

Measures

Item-choice tasks. Children completed the same fruit and vegetable item-choice tasks as in Experiment 1 with the same animals and fruits/vegetables. Thus, children were introduced to an animal character (i.e., hippo or dog) whose favorite food (i.e., apple or carrot) was unavailable. After children learned this, they went to Room B for a 5-min break. However, whereas in Experiment 1 children were assigned to one of three conditions, each child in Experiment 2 completed both the generate-category and forced-choice conditions in this order within the same task (described below).

Fruit task. After the 5-min break in Room B, the experimenter first asked children the generate-category question. Similar to Experiment 1, children were told that it was time to go back to the other room, to close their eyes, and to think about what would be a good fruit to bring back with them. As in Experiment 1, children were also asked to explain their choice. Regardless of whether or not children generated the correct fruit (i.e., "apple"), the experimenter randomly laid out three fruits (i.e., apple, lemon, and orange) on the floor (thereby providing children with the available response options, including the correct choice). Children were then told that they could take one of the fruits back with them and were asked which one they wanted to take. Children were again asked to explain their choice. If by the end children had verbally generated/selected the apple, they went into Room A and fed the hippo. If they had not generated or subsequently selected the apple, they still fed the hippo and the experimenter told them that their chosen option (i.e., lemon or orange) was the hippo's second favorite food.

In each of the two parts of this task (i.e., generative and forced-choice), children received a score for their item generation/selection (0 or 1) and a score for their explanation (0 or 1) for a total of four scores. For both the generative and forced-choice questions, children received a score of 1 if they generated/chose the apple and a score of 0 if they generated/chose any other option. For their ensuing explanation, children received a score of 1 if they provided a future-oriented response (e.g., referred to the hippo and/or Room A) and a score of 0 for reasons unrelated to the future (e.g., "because it smells good").

Vegetable task. After the 5-min break in Room B, the experimenter told children that it was time to go back to the other room and that they should close their eyes and think about what would be a good vegetable to bring back with them. Children were then asked to explain their choice. After children provided a reason for their choice, the experimenter randomly laid out three vegetables (i.e., pepper, cucumber, and carrot) on the floor and asked the forced-choice question (even if they had correctly responded to the open-ended question). Thus, children were asked to select the vegetable that they wanted to take back to the other room with them. Children were again asked to explain their choice. If by the end children had generated or subsequently selected the carrot, they went into Room A and fed the dog. If they selected one of the other two vegetables (i.e., cucumber or pepper), the experimenter stated that this option was the dog's second favorite food. Children's responses were scored in the same way as for the fruit task.

Category fluency tasks. Children completed two category fluency tasks that were adapted from Adlam et al. (2010): an animal task and a food task. In the animal task children were asked to generate as many animal names as they could in 60 s, and in the food task they were asked to generate as many food names as they could in 60 s. The experimenter prompted children with "Can you think of another?" or "Anything else?" after each response or when children paused. Children received 2 points for each different animal (e.g., lion, dog, shark) and food (e.g., lettuce, strawberry, apple juice) they generated. However, in the animal task, children received only 1 point for variations of an animal that they had earlier stated (e.g., "lioness" after stating "lion," "puppy" after stating "dog"). And in the food task children received only 1 point if they generated the broader category label (e.g., meat, fruit) rather than a specific food (e.g., beans, strawberries). Children's points were summed for two total scores reflecting how many animals and how many foods they had generated.

Generativity/creativity task. Children were asked to name all the things they could think of that are *red.* This task was also timed for 60 s. This task was adapted from one given to children by Suddendorf and Fletcher-Flinn (1997). Children were given 1 point for each object that is almost always red (e.g., fire, apple, lips). Children were not, however, given any points for naming objects that could be red (e.g., jacket) but for which "red" is not a defining feature. Children's points were summed for a total score. Although this task is quite similar to a category fluency task, we kept them separate because the generativity task requires children to generate options for a less well-defined category than the category fluency tasks and is also considered an early marker of creativity.

Interrater reliability was assessed for the three tasks using the video-recordings of 35% of the participants. Agreement between the primary and secondary coders ranged from .94 to .99; all disagreements were resolved in ensuing discussions between the coders.

Results and discussion

Table 2 presents means and standard deviations for children's item-choice scores for the generative and forced-choice questions (see below for their explanation data). Children's item choices on the generative question did not differ as a function of task type (i.e., fruit or vegetable; McNemar test, $p = 1.00$), nor did their scores differ on the cued questions (McNemar test, $p = .18$). For this reason, we collapsed item-choice scores across these two tasks. The same was true of children's explanations (McNemar test, $p = .289$, and McNemar test, $p = .581$, for explanations following their generative item choices and forced-choice item choices, respectively). Finally, children's responses on the generate-category questions and the forced-choice questions did not differ as a function of gender, $F(1, 46) = 0.28$, $p = .601$, and $F(1, 46) = 0.01$, $p = .928$, respectively. This was also true for explanations following the generate-category and forced-choice questions, $F(1, 45) = 2.29$, $p = .138$, and $F(1, 46) = 0.69$, $p = .412$, respectively.

Effects of age and condition on children's item choices

We ran a repeated-measures ANOVA with age (3, 4, or 5 years) as a between-participants factor and condition (generate-category or forced-choice) as a within-participants factor. This analysis revealed a significant effect of age, $F(2, 45) = 3.91$, $p = .027$, $\eta_p^2 = .15$, and condition, $F(1, 45) = 35.35$, $p < .001$, $\eta_p^2 = .44$, but no significant Age × Condition interaction, $F(2, 45) = 1.72$, $p = .191$, $\eta_p^2 = .07$. Our condition effect was due to children performing better in the forced-choice condition ($M = 1.69$, $SD = 0.59$) than in the generate-category condition ($M = 0.94$, $SD = 0.78$) (see Table 2). To further explore our age effect, we ran a series of Tukey's post hoc tests showing that, overall, 5-year-olds ($M = 1.50$, $SD = 0.68$) outperformed 3-year-olds ($M = 1.03$, $SD = 0.68$), $p = .030$, but there were no significant differences between the 3- and 4-year-olds ($M = 1.41$, $SD = 0.61$), $p = .099$, or between the 4- and 5-year-olds, $p = .858$.

Children's item-choice errors. As in Experiment 1, children made very few errors (16%) in the forced-choice condition, and of these errors they simply chose one of the two other (incorrect) options presented. In the generate-category condition, children erred 53% of the time ($n = 51$), and the majority of these errors ($n = 38$) entailed the generation of an incorrect fruit/vegetable. Similar to Experiment 1, a common error (13 of 25 errors) in the vegetable task (which always came second) was to provide a fruit label (e.g., apple, orange, watermelon), which again suggests some interference from the previous fruit task. This was true for all three age groups. Of the remaining 12 errors, children generated either an incorrect vegetable ($n = 5$), another food ($n = 1$), or an unrelated item/no response ($n = 6$). As in Experiment 1, this pattern suggests that it was not solely interference that was hampering children's performance.[2] For the fruit task (which always came first), children made just as many errors (26 in total), but these were mostly the verbal generation of an incorrect fruit ($n = 18$), another food ($n = 3$), or an unrelated item/no response ($n = 5$).

[2] We also reran the repeated-measures ANOVA with age (3, 4, or 5 years) as a between-participants factor and condition (generate-category or forced-choice) as a within-participants factor on children's item choices for the fruit task only (which always came before the vegetable task). Although we did not obtain a significant effect of age (likely due to loss of power), the effect nevertheless trended in this direction, $F(2, 45) = 2.36$, $p = .106$, $\eta_p{}^2 = .10$. Importantly, our effect of condition remained highly significant, $F(1, 45) = 16.70$, $p < .001$, $\eta_p{}^2 = .27$, and there was no significant Age × Condition interaction, $F(2, 45) = 0.46$, $p = .636$, $\eta_p{}^2 = .02$.

Table 2
Mean scores and standard deviations for all tasks as a function of age in Experiment 2.

Task	3-year-olds	4-year-olds	5-year-olds
Item-choice tasks			
Generate-category	0.50 (0.63)	1.06 (0.77)	1.25 (0.77)
Forced-choice	1.56 (0.73)	1.75 (0.45)	1.75 (0.58)
Cognitive tasks			
Category fluency animal	8.19 (4.90)	10.81 (4.34)	18.25 (6.09)
Category fluency food	7.19 (4.18)	9.56 (3.39)	16.44 (5.24)
Generativity	1.38 (0.96)	2.06 (1.12)	2.94 (1.69)

Note. Standard deviations are in parentheses.

Verbal explanations. We first examined children's explanations for their correct item choices in the generate-category condition. Here, 67% of children's explanations referenced the animal's need for that item (e.g., "for the doggy"). In contrast, children's incorrect explanations mostly referred to a current need (e.g., "I love apples") or lack of knowledge (e.g., "I don't know"). In the forced-choice condition, 58% of children's explanations were correct (i.e., referenced the previous need/problem), and again incorrect explanations tended to reference a current need or lack of knowledge.

Relations between item-choice tasks and category fluency and generativity tasks

Children's scores on the two category fluency tasks and one generativity task are shown in Table 2. One-way ANOVAs indicated that there was a significant effect of age on each of the three tasks: category fluency animal, $F(2, 45) = 16.35$, $p < .001$, $\eta^2 = .42$; category fluency food, $F(2, 45) = 19.63$, $p < .001$, $\eta^2 = .47$; and generativity, $F(2, 45) = 5.84$, $p < .01$, $\eta^2 = 21$. Follow-up Tukey's tests showed that 5-year-olds performed significantly better than 3- and 4-year-olds in both category fluency tasks (all ps < .01) and that 5-year-olds outperformed 3-year-olds in the generativity task ($p = .004$). No other differences were significant.

We first ran Pearson's product-moment correlations between children's scores on the category fluency tasks, the generativity task, and item-choice scores in the generate-category condition (range = 0–2) and forced-choice condition (range = 0–2). As can be seen in Table 3, children's scores in the generate-category condition were significantly correlated with their performance on both of the category fluency tasks but not on the generativity task. In contrast, children's performance in the forced-choice condition was not significantly correlated with any of these three tasks. We next ran second-order correlations controlling for children's age in months. The one significant correlation that remained was between the generate-category condition and the category fluency food task.[3] These findings provide partial support for the idea that those children who are more proficient at generating instances that fit a given category are also those who perform better when asked to think about an item (from a particular category) that will be useful in the future. We further discuss the implications of this relation in the next section.

General discussion

A growing body of work during the past decade has begun to shed light on young children's capacity to think about the future. Much of this work has focused on whether children can select the means to address a future problem and has been heavily inspired by Tulving's (2005) spoon test. The first attempts to craft developmental versions of the spoon test supported Tulving's prediction that only children aged 4 years and older would succeed (e.g., Suddendorf & Busby, 2005; Suddendorf et al.,

[3] Responses to the two category fluency tasks may have been affected by children having received the fruit task beforehand. In the fruit task, children were exposed to two animals (hippo and tiger) and three fruits (apple, orange, and lemon). Thus, we reran all relevant analyses removing mention of these items from children's total scores on the category fluency tasks. The same age-related patterns were still obtained (i.e., age significantly affected children's performance on both category fluency tasks) and, importantly, the category fluency food task remained significantly correlated with the generate-category scores after controlling for age in months.

Table 3
Correlations between item-choice tasks and category fluency and generativity tasks in Experiment 2.

Variable	2	3	4	5	6
1. Age in months	.362*	.198	.661**	.694**	.414**
2. Generate-category		.187 (.127)	.296* (.081)	.478** (.338*)	.179 (.034)
3. Forced-choice			.072 (−.080)	.105 (−.045)	.098 (.018)
4. Category fluency animal				.786** (.606**)	.544** (.395**)
5. Category fluency food					.519** (.354*)
6. Generativity					

Note. Partial correlations controlling for age in months are in parentheses. $N = 48$.
* $p < .05$.
** $p < .01$.

2011). Several subsequent studies then showed that even 3-year-olds succeed when, for example, the encoding event is made particularly salient (or motivating) for them (e.g., Payne et al., 2014). Yet, even more recently, clever variations of the spoon test have shown that children as old as 5 years experience difficulty (although typically being above chance) when they must either plan a two-step sequence (Martin-Ordas, 2018) or create the means to address a previously encountered problem (Moffett et al., 2018).

Our data both replicate and extend these previous findings. First, we also show that 3-year-olds succeed on certain versions of the spoon test, in our case the forced-choice condition. Although this may have been because children experienced a salient/memorable event (e.g., some children expressed sadness that the animal did not have his or her favorite food available), we suspect that it is more likely due to the fact that the delay we imposed between children encountering the problem and having the opportunity to solve it was 5 min. In contrast, Suddendorf et al.'s (2011) original version of the fruit task involved a 15-min delay. It is possible that this difference in delay time is what led our 3-year-olds to succeed (78–85% correct across both studies) and led their 3-year-olds to fail (~30% correct). This is plausible given that when both Suddendorf et al. and Scarf et al. (2013) ran "no-delay" versions of the item-choice task, 3-year-olds were successful in 70% and 75% of the instances, respectively. Moreover, in Scarf et al.'s study, this percentage dropped to 33% after a 30-min delay. This led these researchers to conclude that it is not the encoding of the event that hampers children's performance on item-choice tasks but rather the retention of this information over a delay (see also Atance & Sommerville, 2014). Indeed, as the amount of delay in item-choice tasks increases, success decreases. Our goal was not to manipulate delay but instead to choose one length of delay and keep it constant across conditions. Yet, our data and those of others clearly point to length of delay—and hence episodic memory—as being a critical factor in children's performance on item-choice tasks.

More important, however, the data from both of our experiments clearly show that when children must verbally generate a response without being "cued" by the correct item being physically present, their performance is negatively affected. Even though 5-year-olds tended to show more proficiency than 3- and 4-year-olds in the generate-category condition (especially in Experiment 1), their performance was not at ceiling—this, despite being constrained by the category label (i.e., fruit or vegetable) label with which they were provided.

Why was it difficult for children to generate the correct item in the generate-category condition? As we argued in the Introduction, presenting children with the items themselves (of which the correct option is embedded) may in fact remove the requirement of mentally traveling back in time to reexperience the problem (e.g., "hippo needs an apple"), and instead through a process of association, connection with the earlier problem, or combination of both the correct item is selected. However, when children must *generate* the correct item, it seems less likely that an associative process is instantiated. Nevertheless, we recognize that the higher-level category label (e.g., fruit) that we provided children in our generate-category condition also served as a cue to the correct response. Indeed, the reason we implemented the generate-object condition is because we deemed it an even more stringent test of children needing to fully generate—in the absence of any category-level cue—the item of need. However, as discussed in Experiment 1, this condition was problematic. If anything, 5-year-olds (the oldest children in our sample) may have been "overly generative" and identified items that they desired (e.g., different kinds of toys) but that were unrelated to the past problem. An intriguing possibility is that 3-year-olds, who had the highest mean score on this task, were in fact less generative and more constrained by the current context than 5-year-olds.

Another aspect of our results that sheds light on why children had difficulty with the generate-category condition is the significant correlation we found between it and performance on one of our category fluency tasks. Category fluency tasks require that individuals search their memory to generate appropriate exemplars of a given category. This would also seem to be a basic requirement of our generate-category condition, without which children could not succeed. Because children are not cued by the correct item, they are required to generate this item from a potentially large set of exemplars—even when they are constrained by a given category (e.g., fruit), as in our generate-category condition. Yet, being able to generate these exemplars is, in our view, necessary but not sufficient to solve our task. Children must then go the extra step of "blocking" incorrect category exemplars (e.g., orange) to generate the correct one (e.g., apple). To do so is what arguably requires episodically reexperiencing the past problem (i.e., that the hippo needs an apple) and generating the correct option to address it. Interestingly, our one generativity task that is argued to be an early marker of divergent thinking/creativity was not correlated with the generate-category condition. This makes sense insofar as the generate-category condition does not require children to think "divergently" but rather requires them to "converge" on the correct option. Admittedly, our interpretation of these correlational findings awaits further systematic research, especially given that only the category fluency food task was correlated with performance in the generate-category condition after controlling for age. Regardless, an important future direction is to delve more deeply into the various mechanisms that underlie different forms of future-oriented thought. In doing so, it would also be desirable to control for both receptive and especially expressive language abilities to ensure that any relation between future-oriented thought and variables such as category fluency is not solely a function of language development.

We set out to emulate as closely as possible—within a controlled lab setting—Tulving's (2005) spoon scenario and, specifically, the need for children to verbally generate a future-oriented response. Nonetheless, we acknowledge that we still fell short of this goal. Arguably, in the laboratory environment, it is challenging to set up a scenario in which children encounter a problem and are then left to spontaneously solve it. However, one promising route may be to use the laboratory as the context in which to introduce the initial problem and then, with varying degrees of prompts/scaffolding, determine whether children can generate a future-oriented solution. For example, one could imagine giving children a small backpack that they take home with them after visiting the lab and then enlisting the parents' help to ask children before they return to the lab the next day whether they would like to pack their bag. The dependent measure of interest would be whether children spontaneously seek out the needed item from their own home. Our prediction is that very few 5-year-olds would do so, but perhaps with a series of more targeted prompts success rates would increase. An advantage of this methodology is that it would also reduce the verbal demands of the task (i.e., children would "seek out" rather than "verbally generate" the item of need); indeed, one of the strengths of Tulving's spoon test (see also Suddendorf, 1994) is that it does not require a verbal response on behalf of children. Although we do not think that children fared worse in our generative conditions as compared with the forced-choice condition because of added language requirements (e.g., children of all ages were very familiar with the items we used), developing a "nonverbal" generative condition, as well as administering both receptive and expressive language measures, would help to further support this claim.

It is important to note that although our generative conditions were not "spontaneous" in the way that Tulving (2005) described, we believe that our task modification (along with the modifications of Martin-Ordas (2018) and Moffett et al. (2018)) is an important bridge between the more forced-choice versions of Tulving's spoon test and yet-to-be determined fully spontaneous versions. However, an important issue to keep in mind when comparing different versions of the spoon test (as we did in the current study) is that chance levels will necessarily vary as a function of the number of items used in the forced-choice version. We used three items in our forced-choice version, which set the chance rate at 33%; however, had we opted for more items, the chance rate would have been lower (and perhaps children's performance with it), thereby potentially affecting comparisons between forced-choice and more generative task versions.

Our findings fit nicely within the broader research on children's future-oriented thinking. In particular, another way in which researchers have studied young children's concept of the future has been to ask them questions about what they will be doing "tomorrow" (e.g., Busby & Suddendorf, 2005) or at specific points in time (e.g., Hayne, Gross, McNamee, Fitzgibbon, & Tustin, 2011; Quon & Atance, 2010). When asked the more open-ended question "Can you tell me something you are going to do tomorrow?" 3-year-olds perform quite poorly (~30% correct), whereas 4- and 5-year-olds perform significantly better yet still not at ceiling (~70% correct). However, when children are asked about more specific events (e.g., "What are you going to do the next time you go to the park?"; Quon & Atance, 2010) or events are generated for them (e.g., "Tomorrow your mum said that you're going to the farmer's market. What can you tell me about this?"; Hayne et al., 2011), 3-year-olds show increased proficiency. In this respect, there is an important parallel between these results from "verbal" methodologies and ours (and others) using a behavioral approach. In both cases, younger children need more contextual support or "prompting" to respond in a future-oriented manner.

As with any cognitive ability, future thinking does not develop in an "all or none" fashion but rather its different components build on one another. With respect to spoon-like tasks in particular, data (including ours) show that children as young as 3 years can pass when the event in question is particularly salient and/or the delay between problem presentation and item selection is short. By 4 or 5 years of age, children become quite proficient at item-choice tasks that follow the general setup used by Suddendorf and colleagues (e.g., Suddendorf et al., 2011). However, there are at least three task modifications of which we are now aware that increase difficulty: (a) being required to plan more than one step in the future (Martin-Ordas, 2018), (b) needing to create the means to solve the problem (Moffett et al., 2018), and (c) needing to verbally generate the response option in the absence of being presented with the items (current data). In light of this, a possible future direction is to implement a longitudinal study that tracks the extent to which children progress through a predictable sequence of task variations during the preschool and early school years. This kind of study would continue to shed light on one of our most adaptive cognitive abilities and how it develops.

Acknowledgments

This research was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada to the first and fourth authors.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2018.12.006.

References

Addis, D. R., Pan, L., Musicaro, R., & Schacter, D. L. (2016). Divergent thinking and constructing episodic simulations. *Memory, 24*, 89–97.

Adlam, A.-L. R., Patterson, K., Bozeat, S., & Hodges, J. R. (2010). The Cambridge semantic memory test battery: Detection of semantic deficits in semantic dementia and Alzheimer's disease. *Neurocase, 16*, 193–207.

Atance, C. M., Louw, A., & 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.

Atance, C. M., & Mahy, C. E. V. (2016). Episodic future thinking in children: Methodological and theoretical approaches. In K. Michaelian, S. B. Klein, & K. K. Szpunar (Eds.), *Seeing the future: Theoretical perspectives on future-oriented mental time travel* (pp. 367–385). New York: Oxford University Press.

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

Atance, C. M., Metcalf, J. L., & Thiessen, A. J. (2017). How can we help children save? Tell them they can (if they want to). *Cognitive Development, 43*, 67–79.

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

Atance, C. M., & Sommerville, J. A. (2014). Assessing the role of memory in preschoolers' performance on episodic foresight tasks. *Memory, 22*, 118–128.

Busby, J., & Suddendorf, T. (2005). Recalling yesterday and predicting tomorrow. *Cognitive Development, 20*, 362–372.
Dickerson, K. L., Ainge, J. A., & Seed, A. M. (2018). The role of association in pre-schoolers' solutions to "spoon tests" of future planning. *Current Biology, 28*, 2309–2313.e2.
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*, 343–355.
Hudson, J. A., Mayhew, E. M. Y., & Prabhakar, J. (2011). The development of episodic foresight: Emerging concepts and methods. *Advances in Child Development and Behavior, 40*, 95–137.
Mahy, C. E. V., Grass, J., Wagner, S., & Kliegel, M. (2014). These pretzels are going to make me thirsty tomorrow: Differential development of hot and cool episodic foresight in early childhood? *British Journal of Developmental Psychology, 32*, 65–77.
Martin-Ordas, G. (2018). "First, I will get the marbles": Children's foresight abilities in a modified spoon task. *Cognitive Development, 45*, 152–161.
McCormack, T., & Atance, C. M. (2011). Planning in young children: A review and synthesis. *Developmental Review, 31*, 1–31.
Moffett, L., Moll, H., & FitzGibbon, L. (2018). Future planning in preschool children. *Developmental Psychology, 54*, 866–874.
Pastor-Cerezuela, G., Fernández-Andrés, M.-I., Feo-Álvarez, M., & González-Sala, F. (2016). Semantic verbal fluency in children with and without autism spectrum disorder: Relationship with chronological age and IQ. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00921.
Payne, G., Taylor, R., Hayne, H., & Scarf, D. (2014). Mental time travel for self and other in three- and four-year-old children. *Memory, 5*, 675–682.
Quon, E., & Atance, C. M. (2010). A comparison of preschoolers' memory, knowledge, and anticipation of events. *Journal of Cognition and Development, 11*, 37–60.
Redshaw, J., & Suddendorf, T. (2013). Foresight beyond the very next event: Four-year-olds can link past and deferred future episodes. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00404.
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*, 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*, 125–132.
Scarf, D., Smith, C., & Stuart, M. (2014). A spoon full of studies helps the comparison go down: A comparative analysis of Tulving's spoon test. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00893.
Suddendorf, T. (1994). *Discovery of the fourth dimension: Mental time travel and human evolution* Master's thesis. Hamilton, New Zealand: University of Waikato http://cogprints.org/729/.
Suddendorf, T. (2010). Linking yesterday and tomorrow: Preschoolers' ability to report temporally displaced events. *British Journal of Developmental Psychology, 28*, 491–498.
Suddendorf, T., & Busby, J. (2005). Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learning and Motivation, 36*, 110–125.
Suddendorf, T., & Fletcher-Flinn, C. M. (1997). Theory of mind and the origin of divergent thinking. *Journal of Creative Behavior, 31*, 169–179.
Suddendorf, T., & Fletcher-Flinn, C. M. (1999). Children's divergent thinking improves when they understand false beliefs. *Creativity Research Journal, 12*, 115–128.
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*, 26–33.
Thompson, C., Barresi, J., & Moore, C. (1997). The development of future-oriented prudence and altruism in preschoolers. *Cognitive Development, 12*, 199–212.
Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). New York: Oxford University Press.
Ward, W. C. (1968). Creativity in young children. *Child Development, 39*, 737–754.