Received: 13 March 2018 | Revised: 14 April 2020 | Accepted: 20 May 2020

DOI: 10.1002/dev.22016

RESEARCH ARTICLE

Developmental Psychobiology WILEY

Applying computational modeling to assess age-, sex-, and strategy-related differences in *Spin the Pots*, a working memory task for 2- to 4-year-olds

Laura Zimmermann[1] | Hannah E. Frank[2] | Francys Subiaul[3,4,5] | Rachel Barr[1]

[1]Department of Psychology, Georgetown University, Washington, DC, USA

[2]Department of Psychology, Temple University, Philadelphia, PA, USA

[3]Department of Speech, Language, and Hearing Sciences, The George Washington University, Washington, DC, USA

[4]Department of Anthropology, Center for the Advanced Study of Human Paleobiology, The George Washington University, Washington, DC, USA

[5]Institute for Neuroscience and Mind-Brain Institute, The George Washington University, Washington, DC, USA

Correspondence
Laura Zimmermann, Department of Psychology, Georgetown University, 306 A White-Gravenor Hall, 3700 O Street N.W., Washington, DC 20057, USA.
Email: ljz7@georgetown.edu

Funding information
NSF; Rachel Barr and Peter Gerhardstein, Grant/Award Number: BCS-1023772; Francys Subiaul, Grant/Award Number: BCS-0748717

Abstract

Working memory (WM) develops rapidly during early childhood. In the present study, visual WM (VSM) was measured using the well-established *Spin the Pots* task (Hughes & Ensor, 2005), a complex non-verbal eight-location object occlusion task. A self-ordered hiding procedure was adopted to allow for an examination of children's strategy use during a VWM task. Participants ($N = 640$) between the ages of 2 and 4 years were tested under semi-naturalistic conditions, in the home or in a museum. Computational modeling was used to estimate an expected value for the total trials to complete *Spin the Pots* via a random search and child performance was compared to expected values. Based on this approach, we determined that children who found six stickers retrieved them in significantly fewer trials than the expected value, excluding chance performance and implicating VWM. Results also showed age-related and sex-related changes in VWM. Between 2 and 4 years of age, 4-year-olds performed significantly better than younger children and girls out-performed the boys. Spontaneous use of a color matching hiding strategy was associated with a higher success rate on the task. Implications of these findings for early development of VWM are discussed.

KEYWORDS

computational modeling, early childhood, executive functioning, working memory, sex differences

1 | INTRODUCTION

Working memory (WM) is a key component of a larger executive functioning (EF) system that also includes inhibitory control and cognitive flexibility (Miyake, Friedman, Witzki, Howerter, & Wager, 2000). WM is defined as the ability to mentally retain, update, and/or manipulate information for a short time, on the scale of seconds (Atkinson & Shiffrin, 1971). Impairments in WM result in difficulties selecting, maintaining, and updating information (Shimamura, 2000). Most researchers consider the maintenance of information in memory different from the manipulation and updating of information in memory. The former is generally referred to as short-term memory. The latter is generally considered to be WM (Garon, Bryson, & Smith, 2008; Gathercole, 1999). While WM and inhibitory control are dissociable (Best & Miller, 2010; Garon et al., 2008), complex working memory tasks that involve maintaining, manipulating, and updating multiple items co-activate WM and inhibitory control components of the EF system (Garon et al., 2008). Inhibitory control failure during WM tasks is indexed via perseveration, which is the repeated search to a previously rewarded location (Diamond, 1990). Based on theoretical models (e.g., Baddeley, 2012) researchers have generally measured WM using both verbal and non-verbal tasks.

WM improves dramatically during the preschool years (Garon et al., 2008) and development continues well into adolescence

(Best & Miller, 2010; Mandolesi, Petrosini, Menghini, Addona, & Vicari, 2009; Overman, Pate, Moore, & Peuster, 1996). Item recall increases between 3 and 5 years on verbal measures of WM using *digit* and *word span tasks* (Espy & Bull, 2005) and the non-verbal *radial arm maze* (Mandolesi et al., 2009). Performance on WM measures at age 5 have been shown to be a better predictor of academic success in reading and math than IQ (Alloway & Alloway, 2010; Bull & Scerif, 2001). Between the ages of 6 and 16, better WM continues to predict future achievements in mathematics and reading (Dumontheil & Klingbert, 2012; Gathercole, Pickering, Knight, & Stegmann, 2004). Beyond academic success, WM and inhibition have also been associated with social and emotional development (Bernier, Carlson, & Whipple, 2010).

WM tasks used during early childhood have a critical limitation: an overreliance on verbal responses. These verbal WM tasks are confounded by the complexity of the verbal reports required (e.g., *NIH Toolbox List Sorting Working Memory Test; LSWM*, Bauer & Zelazo, 2013), often resulting in floor effects with children under 5. Specifically, the LSWM test of the NIH Toolbox requires complex verbal recall, and performance is notably poor in 3- to 5-year-olds (Riggins, 2013). Other non-verbal tasks suffer from technical limitations and task complexity (e.g., Cambridge Neuropsychological Test Automated Battery (*CANTAB*) *Spatial Working Memory Task*, Luciana & Nelson, 1998). Although performance improved on items of the *CANTAB spatial WM* task in children between 4 and 8 years, 4-year-olds performed poorly (Luciana & Nelson, 1998). Given these limitations, we currently lack a cohesive and coherent understanding of the development of WM in young children.

One way to reduce verbal demands is to examine non-verbal types of working memory (WM). WM is comprised of visual memory (VWM), our capacity to remember what we see: shapes, colors, or features of stimuli, and spatial memory (SWM), our capacity to remember information about locations and movement (Baddeley & Hitch, 1974; Logie, 1995). Non-verbal WM tasks are more developmentally appropriate for young children who are not yet proficient in verbal language. Typically, non-verbal WM tasks involve finding a rewarding occluded object, measured via reaching. In order to successfully complete non-verbal WM tasks, participants must attend to, manipulate, and update a number of perceptual cues, including color and location.

VWM tasks require attention to individual item cues, such as color or size. In the *scrambled boxes* task (Diamond, Prevor, Callender, & Druin, 1997), children view an array of different boxes and have to search for a reward within each location per trial; however, the location of the boxes changes after each trial. In the *scrambled boxes* task, the child must attend to specific individual item features of the boxes such as the color, rather than the spatial location. The scrambled boxes task also involves inhibitory control to a previously rewarded item. The scrambled boxes task assesses VWM, specifically, as spatial location is randomized from trial to trial.

Age-related changes have been identified in the *scrambled boxes task*. Diamond and colleagues (1997) directly compared performance on the *scrambled boxes* tasks and reported age-related changes. In the 15- to 30-month group (three boxes task), older toddlers needed fewer reaches to open all boxes and fewer reaches to the same location than younger toddlers. In the 3.5- to 7-year-old group (six boxes), performance improved with age in the scrambled conditions (Diamond et al., 1997). The task *Spin the Pots* (Hughes, 1998; Hughes & Ensor, 2005), which is the focus of the current research, is a variant of the scrambled boxes task.

As well as age-related changes, studies suggest that there may be sex-related differences in WM. Diamond et al. (1997) showed that 3.5- to 7-year-old girls performed better than boys on two tasks that involved integration of color cues with spatial cues. Specifically, girls performed better on the three pegs color tasks, and the six-item *scrambled boxes* tasks. The findings suggest that color cues may provide an important cue for preschoolers during WM tasks.

The present study will test 2- to 4-year-old children on the *Spin the Pots* task to better understand developmental trajectories in non-verbal VWM, as well as any sex differences underlying performance early in development. The *Spin the Pots* task (Hughes, 1998) is a multi-location visual search task that was derived from methodology used with non-human primates (Petrides & Milner, 1982). It was an eight-box variant adapted from Diamond's and colleagues (1997) 'six boxes scrambled' task. It was developed to assess non-verbal WM and inhibitory control during early childhood. Hughes (1998) gave 3- and 4-year-olds 15 trials to find eight hidden rewards. Hughes and Ensor (2005) later gave 2-year-old children 16 trials in which to find six hidden stickers under eight different pots; the location of the pots was rotated 180 degrees (scrambled) after each trial. Similar to the scrambled boxes task (Diamond et al., 1997; Petrides & Milner, 1982), this non-verbal, internally ordered, and scrambled task requires children to retrieve six hidden objects from eight distinctly colored cups. In *Spin the Pots*, the color of the cup is a useful cue for children to encode and update across trials as location changes per trial. To succeed on the task, children must use VWM to maintain and update their memory for cups where stickers have been retrieved. We hypothesized that with each additional year, children would become more efficient at solving the task (better WM score and fewer perseverations). Given some prior sex-related differences in the WM field, we decided to test for sex-related differences as well, but we did not have a directional hypothesis.

Age-related and sex-related individual differences in performance on the *Spin the Pots* task have been examined. Hughes (1998) reported no age or sex-related differences in 3- and 4-year-olds but their WM measure only included whether or not the rewards were found and did not consider error rates. In a follow-up study, Hughes and Ensor (2007) conducted a longitudinal study of 2-, 3-, and 4-year-olds and found an age-related increase in performance from age 2 to age 3 and also from 3 to 4 years where performance almost reached ceiling (see also Blakey & Carroll, 2018). It is interesting to note that on a composite score of executive functioning which included *Spin the Pots*, girls performed significantly better than boys at age 3 (Hughes & Ensor, 2007). There were no sex-related differences in 2- and 3-year-olds when the experimenter hid the stickers (Blakey & Carroll, 2018; Blakey, Visser, & Carroll, 2016).

Children in the present study participated in a self-ordered child-directed hiding phase prior to testing (children hid the stickers themselves) rather than an experimenter-directed or -assisted hiding phase. We based our decision on the prior work on self-ordered WM tasks (see Beck, Schaefer, Pang, & Carlson, 2011; Hughes, 1998; Pinto-Hamuy & Linck, 1965; Petrides & Milner, 1982, Diamond et al., 1997). The original Hughes (1998) task specified that there was a self-ordered hiding phase but the Hughes and Ensor (2005) protocol did not specify whether it was a self-ordered or an experimenter-ordered hiding phase. Hughes and Ensor (2007) reported that children were asked to help the experimenter hide the stickers (see also Huber, Yeates, Meyer, Fleckhammer, & Kaufman, 2018). In prior studies, some researchers adopted a self-ordered protocol (Beck et al., 2011; Müller, Liebermann-Finestone, Carpendale, Hammond, & Bibok, 2012; Roman, Ensor, & Hughes, 2016), while others adopted an experimenter-ordered protocol (Blakey & Carroll, 2018; Blakey et al., 2016; Choi, Kirkorian, & Pempek, 2017; Johansson, Marciszko, Brocki, & Bohlin, 2015). In the present study, we opted for a self-ordered version of the task.

Unfortunately, past studies of *Spin the Pots* did not assess estimates of random guessing making it difficult to evaluate prior findings age- and sex-related differences in performance. Prior studies did not assess random guessing because it is difficult to statistically assess whether children found the stickers at a rate that would be different from the probability of a random search. A primary aim of the present study was to utilize computational modeling via a Monte Carlo simulation to estimate the expected value at which all six stickers would be found after a random search.

In sum, the present study extends prior work on the *Spin the Pots* task in the following ways. It will add to a growing body of literature examining WM in young children, examining a large sample of children across a wide age range on a well-established non-verbal WM task, which does not rely on complicated verbal instructions. A limitation of prior studies using the *Spin the Pots* measure, however, is that performance was not compared to random search. Rather than comparing the number of correct responses between children of different ages, Monte Carlo simulations which have been used frequently to assess psychometric properties and developmental processes (e.g., Bolenz, Reiter, & Eppinger, 2017; van den Bos, Bruckner, Nassar, Mata, & Eppinger, 2018), were conducted to calculate a statistically expected value of the number of trials to retrieve the stickers under random search conditions. Third, the use of a self-ordered procedure allowed for a closer look at children's strategy use (color or linear) during the hiding phase and whether it was related to their search behaviors.

2 | METHOD

2.1 | Participants

Participants included 640 typically developing children (306 boys). Data were collected for 24- to 36-month-old children in their homes in Washington, DC, and 30- to 52-month-olds in the Smithsonian National Zoo and the Smithsonian National Museum of Natural History. This study is an analysis of a large dataset which includes prior studies that were collected in the home or in the museum over the course of 3 years and, thus, the frequencies of children at each age are unequal. Portions of data collected on other tasks with these participants have been published elsewhere, but did not include the *Spin the Pots* data (Barr et al., 2016; Moser, Zimmermann, Dickerson, Grenell, Barr, & Gerhardstein, 2015; Subiaul, Zimmermann, Renner, Schilder, & Barr, 2016; Zimmermann, Moser, Gerhardstein, & Barr, 2015; Zimmermann, Moser, Lee, Gerhardstein, & Barr, 2017). There was roughly an equal number of males and females at each age group (see Table 1). Independent groups of children were tested at 2 years (n = 334, M age = 27 months 21 days, SD = 3.53 months), 3 years (n = 205, M age = 39 months 4 days, SD = 3.81 months), and 4 years (n = 101, M age = 52 months 12 days, SD = 5.58 months). We include age as a categorical variable because data were drawn from studies with narrow age ranges and we did not have a normally distributed continuous age range. Although the age groups are not equal, the sample size is large and well-powered. Participants were primarily Caucasian (67.2%) and from college-educated families (M years of education = 17.09, SD = 3). The remaining sample included the following races: mixed (10.3%), African-American (4.8%), Asian (4.2%), Native American (0.2%), other (0.8%), and the remaining participants (n = 12.5%) did not report the race of the child. With regard to ethnicity, 11.6% of the sample were Hispanic. Approximately one third of children lived in homes where more than one language was routinely spoken (32%; bilingual homes). Fifty children (7.8%) were excluded from the analysis: 12 due to experimenter error, 3 for technical error with the video, 11 for failure to interact with the experimental stimuli, 4 due to parental or sibling interference, and 20 for knocking over the cups during testing.

TABLE 1 The number of males and females that found 2 to 6 stickers on *Spin the Pots* as a function of age group.

		Age			
		2 years	3 years	4 years	Total (n)
Males	3 stickers	1	0	0	1
	4 stickers	13	6	0	19
	5 stickers	42	22	6	70
	6 stickers	118	65	44	227
		174	93	50	317
Females	2 stickers	0	1	0	1
	3 stickers	2	0	0	2
	4 stickers	7	3	1	11
	5 stickers	38	24	9	71
	6 stickers	113	84	41	238
		160	112	51	323
Total		334	205	101	640

2.2 | Apparatus and stimuli

The *Spin the Pots* (Bernier et al., 2010; Hughes & Ensor, 2005) apparatus is comprised of eight distinctly colored opaque cups, six stickers that match the color of the cups, and a lazy Susan. A lazy Susan is a turntable (rotating tray) that was attached to the bottom of the stimulus that allowed it to rotate 360 degrees. All eight cups fit inside the lazy Susan in a circle with equal spacing between them. An opaque cover was used to cover the cups in between trials and had a handle on top of the cover in order to easily cover and uncover the lazy Susan, see Figure 1. It was 15 cm tall, 35 cm in diameter, and 110 cm in circumference.

2.3 | Design and procedure

After obtaining informed consent, primary caregivers were asked to complete a general questionnaire (including contact information, education, career, parental education, child language exposure, and media exposure).

Hiding phase. For the *Spin the Pots* task (Hughes, 1998; Hughes & Ensor, 2005), the experimenter encouraged the child to hide the six colored stickers themselves under six of the eight brightly colored cups, leaving two cups empty. After all stickers were hidden, the experimenter showed the child the two cups that did not have a sticker and said, "Look, no stickers under these cups!" The opaque cover was placed over all the cups on the lazy Susan and the entire tray was spun 180 degrees.

Search Phase. The experimenter uncovered the cups and instructed the child to find one of the stickers. If the child found a sticker, the experimenter praised the child, the sticker was set aside or given to the child's caregiver, and the lid was replaced and the tray was spun 180 degrees again. After each trial, the tray was spun 180 degrees to counterbalance the position of the cups. If the child did not find a sticker, the experimenter gave appropriate feedback (e.g., "no sticker there, let's try again"), the lid was replaced, and the tray was spun 180 degrees again. This task required the child to hold the color of the cups that did not have stickers in mind and to update this memory after each trial. The task ended when the child found all six stickers or reached sixteen trials. A subset of children did not find all six stickers within sixteen trials. In other studies, researchers (Hughes, 1998; Hughes & Ensor, 2005) gave children up to 16 trials to find all the stickers, but in the present study children were given additional trials if the child engaged with the apparatus within 1 min of the start of the trial, and had not yet retrieved all 6 stickers. Testing continued for up to 35 trials, otherwise the task ended ($M_{2years} = 15.16$, $SD_{2years} = 5.18$, $M_{3years} = 14.50$, $SD_{3years} = 4.89$, $M_{4years} = 12.23$, $SD_{4years} = 4.13$; $M_{boys} = 15.09$, $SD_{boys} = 5.14$, $M_{girls} = 13.90$, $SD_{girls} = 4.85$). The number of trials and success rate variables were normally distributed.

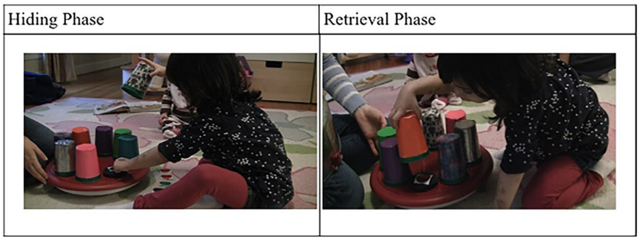

FIGURE 1 During the hiding phase of Spin the Pots, the child hides 6 stickers under distinctly colored cups. During the retrieval phase, children try to retrieve one sticker per trial

3 | RESULTS

3.1 | Coding

Task performance was videotaped for subsequent coding. For the *Spin the Pots* task, each child was given the following scores: color strategy, linear strategy, success rate, perseveration, alternate perseveration, and the location of the cup searched.

3.1.1 | Hiding phase

Color Strategy

This measure quantifies hiding behavior of the child prior to the test phase. If the child matched at least four of the six stickers to their correct cup (i.e., green smile sticker under green cup), they received a point. Matching three or fewer stickers was not defined as color strategy use.

Linear strategy

This measure quantifies hiding behavior of the child prior to the test phase. If the child hid at least four of the six stickers in a linear fashion (without skipping cups) around the circumference of the apparatus, they received a point. Linear hiding of three or fewer stickers was not defined as linear strategy use.

3.1.2 | Test phase

Success rate

Given that the number of trials could vary across children, the success rate was calculated by dividing the number of stickers retrieved by the total number of trials.

Perseveration rate

This is the number of times the child chose a cup that was selected on the previous trial (whether it did or did not have a sticker on the first search). This allows us to quantify errors based on the feature of the cup. For example, selecting purple, then purple again across two trials would equate to one point on this measure. The number of perseverations was divided by total trials completed.

3.1.3 | Alternate perseveration rate

This is the number of times the child chose a cup that was selected two trials ago (whether it did or did not have a sticker on the first search). This allows us to quantify errors based on the location of the cup as the apparatus rotated 180 degrees each trial. For example, selecting purple, then orange, and then purple across three trials would equate to one point on this measure. The number of alternative perseverations was divided by total trials completed.

3.1.4 | Location of cup

On each test trial, the four cups closest to the child were considered "near" and the four cups further from them were called "far." For each trial it was determined whether the child reached for a cup in the four nearest to them or four on the other half of the apparatus furthest from them. The total number of times a child searched in the near and far regions was computed. Only one selection, near or far, was possible per trial.

3.1.5 | Reliability

An independent coder scored 14% of the videos to determine reliability of the strategy use (location or color) codes. Inter-rater reliability on strategy use was above the acceptable level of 0.70 (Landis & Koch, 1977), with *kappa* = 0.74. A second independent coder scored 14% of the videos and recorded the color of the cup selected per trial and whether or not a sticker was retrieved. All other calculations could be derived from these two measures. Inter-rater reliability on these codes was also above the acceptable level, *kappa* = 0.92.

3.2 | Data analysis plan

First, preliminary analyses were conducted to assess whether demographic variables (parental education and bilingual status) or test location (home, museum) were associated with primary *Spin the Pots* scores (success rate, perseveration rate, and alternate perseveration rate).

3.2.1 | Monte Carlo simulations

Next, two Monte Carlo simulations were performed to determine the null hypothesis distribution to estimate the expected value of a random search. It is necessary to generate a null distribution because the probability for completion of the *Spin the Pots* task is not direct (e.g., flipping a fair coin has a "direct" probability of 0.5 for heads and tails). *Spin the Pots* searches on preceding trials alter the probability of future searches, making them conditional. For example, on trial 1, the probability of finding a sticker is 6/8. If you find a sticker, then the next trial's probability is 5/8. If you fail to find a sticker, the probability stays at 6/8 until X trials when you do find a sticker.

To be mathematically specific, generating the estimate for random search is complicated because of two factors:

1. The experiment has "path-dependency" which causes the experimental space to change with each success.
2. Stopping rules ensure that the experiment is truncated when a child successfully finds six stickers. For example, the truncation value was set at 200 trials which far exceeds the actual end point in this study. The model also has to allow for a child completing the task in as few as six trials. A separate Monte Carlo simulation had to be run to determine the null distribution for finding five stickers because the stopping rule differed (five stickers rather than six). That is, because of path dependency, random search for five stickers by definition will yield a different expected value than random search for six stickers.

For a sufficiently large *N*, a Monte Carlo simulation approximates the probability distribution of the experiment. At $N = 6$ million, the resulting probability distribution for the experiment was stable and assumed to converge. Additional simulations could have been run, but would not have yielded greater precision.

To simulate a random search on this task, we calculated the expected value for the probability distribution and ran computational models for finding five and six stickers. Only those children who performed below the expected value were likely completing the task using working memory rather than random search.

After running the Monte Carlo simulations, we tested whether performance differed from the expected value for each model. Specifically, children's performance was compared to the expected values using one-way *t*-tests to determine whether or not the groups significantly differed from the expected number of trials. If it was less than the expected number of trials, we inferred that children were using working memory to retrieve the stickers rather than random search. Once we identified the subgroup who completed the task at or below the expected number of trials, we examined factors associated with task performance. Then, we assessed age- and sex-related differences in task performance by conducting a 3 (age) x 2 (sex) multivariate analysis of covariance (MANCOVA) for the following variables: success rate, perseveration rate, alternate perseveration rate with color strategy, and time per trial as covariates.

3.3 | Preliminary analyses

Preliminary analyses including bilingual status and parental education were entered into the model on success rate, perseveration rate, and alternate perseveration rate. No significant ($p < .05$) main effects or interactions involving bilingual status or parental education emerged in the preliminary analyses, and data were collapsed across these variables in further analyses. Furthermore, an equal number of 3-year-olds were run in the home and the museum and did not differ demographically. Preliminary analyses indicated that there were no significant ($p < .05$) main effects involving being tested in the home or museum location in 3-year-olds. Data were collapsed across locations in further analyses.

TABLE 2 Mean time to complete *Spin the Pots*, total trials, time per trial, first error, success rate perseveration rate, alternate perseveration rate, proportion of children who used color strategy and linear strategy, and # of trials near location selected (SDs), as a function of age and sex of the child for those who retrieved six stickers

		Time			Performance				Strategy use		
Sex	*N*	Time to complete (s) (*SD*)	Total trials (*SD*)	Time per trial (s) (*SD*)	First error (*SD*)	Success rate (*SD*)	Perseveration rate (*SD*)	Alternate perseveration rate (*SD*)	Color strategy proportion (*SD*)	Linear strategy proportion (*SD*)	Location trials near (*SD*)
2-year-olds											
Male	118	161.79	14.84	10.82	2.62	0.44	0.74	2.75	0.19	0.22	11.92
		(68.52)	(4.75)	(2.69)	(1.33)	(0.14)	(1.19)	(2.69)	(0.39)	(0.42)	(3.84)
Female	113	146.91	13.32	11.12	2.76	0.50	0.62	2.28	0.13	0.26	10.82
		(65.74)	(4.67)	(3.44)	(1.51)	(0.16)	(0.82)	(2.18)	(0.33)	(0.44)	(4.60)
3-year-olds											
Male	65	129.74	13.89	9.34	2.47	0.49	0.72	2.63	0.30	0.36	10.32
		(58.31)	(4.92)	(2.02)	(1.24)	(0.19)	(1.04)	(2.35)	(0.46)	(0.48)	(4.25)
Female	84	123.10	13.04	9.71	3.01	0.51	0.68	1.98	0.29	0.39	9.69
		(44.99)	(4.29)	2.74	(1.26)	(0.17)	(0.96)	(1.86)	(0.46)	(0.49)	(3.45)
4-year-olds											
Male	44	103.59	12.14	8.53	3.23	0.54	0.71	1.23	0.20	0.50	8.89
		(35.92)	(3.50)	1.47	(1.48)	(0.16)	(0.88)	(1.18)	(0.41)	(0.51)	(3.58)
Female	41	91.68	10.63	8.73	3.38	0.63	0.24	1.46	0.27	0.32	8.02
		(97.85)	(11.41)	8.63	(3.30)	(0.20)	(0.54)	(2.38)	(0.45)	(0.47)	(4.41)

Table 1 shows the range of performance of children across age and sex. The number of stickers retrieved ranged from two to six.

3.4 | Expected value

As previously noted, we used Monte Carlo simulations to identify the expected number of trials that it would take to retrieve six stickers hidden in eight locations if children were searching randomly. The expected value to retrieve all six stickers is 18.35 trials (*SD* = 6.69) and to retrieve five stickers is 11.6 trials (*SD* = 4.44).

To test whether performance differed by age, sex, and number of stickers retrieved, when the expected value was taken into account, a centered mean was computed for total trials by subtracting total trials from 18.35 for those who retrieved six stickers and subtracting total trials from 11.6 for those who only retrieved five stickers. A negative score indicates performance below the expected value. The centered mean allowed for a direct comparison between those who found five or six stickers, allowing for different rates of expected values.

An omnibus 3 (age: 2, 3, 4 years) x 2 (sex: male, female) × 2 (stickers: 5, 6) ANOVA across the number of trials using the centered mean revealed a main effect of age, $F(2,575) = 4.79$, $p = .009$, $\eta_p^2 = 0.016$, no main effect of sex, $F(2,575) = 3.01$, $p = .08$, $\eta_p^2 = 0.005$, and a main effect for number of stickers, $F(2,575) = 424.95$, $p < .001$, $\eta_p^2 = 0.43$. There were no significant two-way or three-way interactions (all *F*s < 1). It was not clear from the ANOVA which ages and which sticker numbers were significantly different from the expected value of trials from the Monte Carlo simulation. Therefore, follow-up *t*-tests were needed to determine which group's success rate was better or worse than the expected value.

Under the null hypothesis, we would expect that each age group would perform at the expected value. However, we were able to reject the null hypothesis ($p < .001$) for all age groups who found six stickers using the expected value of 18.35. Children ($n = 465$) who retrieved six stickers performed significantly better than the expected value at age 2: $t(230) = -13.58$, $p < .0001$, Cohen's $d = 0.89$, age 3: $t(148) = -13.17$, $p < .0001$, Cohen's $d = 1.08$, and age 4: $t(85) = 16.55$, $p < .0001$, Cohen's $d = 1.80$. In other words, they retrieved six stickers in fewer trials than the expected value of 18.35 trials. We interpret this result to mean that children were not simply searching randomly, but attempting to keep track of and update information across trials. Furthermore, *t*-tests were also calculated separately for males and females at each age. The pattern of results was the same at each age and for each sex of the child.

However, this was not true for children who only retrieved five stickers. Across age groups and both sexes, children who retrieved five stickers did not complete the task in significantly fewer trials than the expected value of 11.36 trials (all *t's* not significantly better than expected). Given that children who only found five stickers ($N_{\text{2-year-olds}} = 80$, 24% of all 2-year-olds tested; $N_{\text{3-year-olds}} = 46$, 22.4% of all 3-year-olds tested; and $N_{\text{4-year-olds}} = 15$, 15% of all 4-year-olds tested) did not perform better than the expected value, their data were not analyzed further. A chi-square calculating the frequency with which children at each age found five or six stickers was not significant, $\chi^2(2) = 2.39$, $p = .30$, indicating that there were no age-related differences in the frequency with which five or six stickers were retrieved. Taken together, therefore, the findings indicate that those who retrieved six stickers were unlikely to have been searching at random, whereas the performance of those who only retrieved five stickers did not differ from random search.

Results from the simulations were also used to determine the expected number of trials until the first error. The expected number of trials until the first error is 1.33 (*SD*=0.67). All age groups with children who found all six stickers made their first error significantly later than the expected first error, 2-year-olds: $t(230) = 14.50$, $p < .0001$, Cohen's $d = 0.96$, 3-year-olds: $t(143) = 13.61$, $p<.0001$, Cohen's $d = 1.13$, and 4-year-olds: $t(79) = 13.11$, $p<.001$, Cohen's $d = 1.47$. It is important to note that initially those who retrieved only five stickers also made their first error significantly later than expected by chance, 2-year-olds: $t(79) = 8.55$, $p < .001$, Cohen's $d = 0.96$, 3-year-olds: $t(45) = 8.83$, $p < .001$, Cohen's $d = 1.30$, and 4-year-olds: $t(14) = 4.77$, $p < .001$, Cohen's $d = 1.23$. Again, given that the first error occurred significantly later than expected, we interpret this finding to indicate that children were not simply guessing at random and that initially those who retrieved five and six stickers started the search in a similar way.

3.5 | Examining WM performance in those who retrieved six stickers

In our subsequent analyses, we focused on children who retrieved six stickers. This decision was made because this group completed the task in significantly fewer trials than the expected value. They also made their first error significantly later than the expected value, allowing us to infer that they were solving the task using WM. We tested the number of trials and found that there were three outliers who had trials lengths over 30 trials. We removed these outliers from further analyses. The mean scores and standard deviations for the Spin the Pots task as a function of age in years and sex of the child are reported in Table 2 for those children who retrieved six stickers ($n = 463$).

3.6 | Strategy

3.6.1 | Location

In order to assess whether the choice of cups that were closest to them (near) was related to task performance and individual differences, a series of first-order correlations among rate of choosing cups nearest to them, perseveration errors, sex, and age of the child were conducted with those who retrieved six stickers (see Table 3).

There was a significant positive correlation between choosing cups located in the bottom half of the apparatus (closest to the child) on the test trials and higher rates of perseveration, $r(461) = 0.24$, $p < .001$, $r^2 = 0.06$, or 6% of the variance and alternate perseveration, $r(461) = 0.73$, $p < .001$, $r^2 = 0.53$ or 53% of the variance. Although the effect was small, older children ($r(461) = −0.26$, $p < .001$, $r^2 = 0.07$) and girls ($r(461) = −0.11$, $p < .02$, $r^2 = 0.01$) were significantly less likely to select cups nearest to them.

3.6.2 | Linear strategy

A total of 139 children (32.1%) used a linear hiding strategy with four or more stickers. Girls and boys were equally likely to use a linear strategy ($n_{girls} = 71$; $n_{boys} = 68$) and 3- and 4-year-old children had more frequent linear strategy use ($n_{2\text{-year-olds}} = 55$ (16.5%), $n_{3\text{-year-olds}} = 54$ (26.9%) $n_{4\text{-year-olds}} = 30$ (34.9%)). A chi-square test of independence confirmed that younger children were less likely to use the linear strategy ($\chi^2(2, N = 456) = 12.02$, $p = .002$, $\phi = 0.16$, small effect size). Linear strategy was not correlated with success rate, perseveration rate, or alternate perseveration rate. Due to the fact that a linear strategy was not associated with our main performance measures on this task, we did not consider this measure further.

3.6.3 | Color strategy

Three- and 4-year-old children had more frequent color strategy use than 2-year-olds ($n_2 = 35$ (10.5%), $n_3 = 41$ (20.4%), $n_4 = 15$ (17.4%)). A chi-square test of independence confirmed that younger children were less likely to use this color strategy, $\chi^2(2, N = 452) = 10.05$, $p = .007$, $\phi = 0.15$, small effect size. There a significant positive correlation between children's color matching stickers to the cups and success rate, $r(452) = 0.13$, $p < .01$. There was no significant correlation between perseveration and children's color matching ($r(452) = −0.07$, $p = .16$) or between alternate perseveration and color matching ($r(452) = −0.04$, $p = .42$). An equal number of males and females used the color strategy.

3.7 | Multivariate analysis of covariance

Based on our finding that the six-sticker group performed better than the expected value estimate calculated from the Monte Carlo simulations, we focused our analysis on children who retrieved six stickers. We hypothesized both age and sex differences on *Spin the Pots* and, therefore, the model included age (years), sex of child, and an age x sex interaction term. The three *Spin the Pots* measures were success rate, perseveration rate, and alternate perseveration rate. These variables were moderately correlated. The first-order correlational analysis indicated the color strategy but not linear strategy was correlated with *Spin the Pots* measures and, therefore, we added color strategy as a covariate. As shown in Table 2, time per trial was longer for younger children than older children. We were concerned that time per trial may increase cognitive load for younger children and result in poorer VWM. For that reason, we also entered time per trial as a covariate into the model. Due to unequal variance, a MANCOVA was conducted and Wilks Lamda corrections are reported throughout.

A MANCOVA was conducted to examine the association between age and sex of the child with *Spin the Pots* measures (success rate, perseveration rate, and alternate perseveration rate) using time per trial and color strategy as covariates. The overall model for age of the child was significant, $F(6,880) = 5.30$, $p = .01$, $\eta_p^2 = 0.04$, and sex of the child, $F(3, 440) = 3.67$, $p = .01$, $\eta_p^2 = 0.02$, was significant, but there was no interaction between age and sex of the child, $F(6, 880) = 1.99$, $p = .07$, $\eta_p^2 = 0.01$. Color strategy was a significant covariate, $F(3, 440) = 3.21$, $p = .02$, $\eta_p^2 = 0.02$, but time per trial was not. Table 4 shows the between-subjects effects. All effect sizes are small except for the significant main effect of age on success rate, which is a medium effect size. The results indicated that children who had a color strategy had a higher success rate. Likewise, those who had shorter time per trial also had a higher success rate. Girls had a higher success rate than boys and older children had higher success rates overall. Older children also had a lower alternate perseveration rate than younger children. A post-hoc Student Newman-Keuls (SNK, $p < .05$) analysis across all groups on both success rate and alternate perseveration rate indicated that the 4-year-olds significantly exceeded the performance of the 2- and 3-year-olds.

TABLE 3 Correlations among success rate, perseveration, alternate perseveration, color strategy, linear strategy, time per trial, and location on *Spin the Pots* for children who retrieved 6 stickers

Measure	1	2	3	4	5	6
1. Success rate	–					
2. Perseveration rate	−0.42**	–				
3. Alternate perseveration rate	−0.57**	0.07	–			
4. Color strategy	0.13**	−0.06	−0.06	–		
5. Linear strategy	−0.02	0.003	0.009	−0.27**	–	
6. Time per trial	0.02	−0.001	0.002	−0.10*	−0.11*	–
7. Location	−0.77**	0.24**	0.73**	−0.11*	−0.04	0.03

* $p<.05$, ** $p<.01$.

TABLE 4 Results of the MANCOVA showing the effects of strategy use during the hiding phase of *Spin the Pots*, time per trial, age, sex, and age x sex interaction on success scores for children who retrieved 6 stickers

		d.f.	*F*	Sig.	Partial Eta Squared
Color strategy	Success rate	1, 442	8.53	0.004	0.02
	Perseveration rate	1, 442	0.30	0.58	0.00
	Alternate perseveration rate	1, 442	0.13	0.72	0.00
Time per trial	Success rate	1, 442	3.97	0.05	0.01
	Perseveration rate	1, 442	0.00	0.99	0.00
	Alternate perseveration rate	1, 442	0.94	0.33	0.00
Age (years)	Success rate	2, 442	13.94	<0.001	0.06
	Perseveration rate	2, 442	0.92	0.40	0.00
	Alternate perseveration Rate	2, 442	7.54	0.001	0.03
Sex of child	Success rate	1, 442	10.26	0.001	0.02
	Perseveration rate	1, 442	2.97	0.09	0.01
	Alternate perseveration rate	1, 442	1.30	0.26	0.00
Age * Sex	Success rate	2, 442	1.54	0.22	0.01
	Perseveration rate	2, 442	3.40	0.03	0.02
	Alternate perseveration rate	2, 442	0.73	0.48	0.00
$R^2 = 0.021$ (Adjusted $R^2 = 0.006$).					

There was no difference between 2- and 3-year-olds, (see Tables 2 and 4). For perseveration rate, there was no main effect of age or sex, but there was an age x sex interaction (see Tables 2 and 4). To test where the interaction occurred, we conducted two follow-up one-way ANOVAs: one for boys and one for girls. We found that for boys there was no significant main effect of age on perseveration rate, $F(2, 223) < 1$, and for girls there was a significant main effect of age on perseveration rate, $F(2, 234) = 3.94$, $p = .02$, $\eta_p^2 = 0.03$. A post-hoc SNK test ($p < .05$) indicates that the 4-year-olds girls had significantly lower perseveration rates than 2- and 3-year-old girls.

4 | GENERAL DISCUSSION

The present study examined age- and sex-related changes in VWM development across an extended age range. Children were permitted additional trials to complete the task, and children's performance was compared to estimated expected values to increase the strength of our interpretation of the findings. Use of the Monte Carlo simulations to calculate expected values overcomes a significant limitation of previous studies which did not provide an estimate of random search.

Consistent with our hypotheses, we found small but significant age-related and sex-related differences in VWM. The study established that by age 4, children are performing significantly better than 2- and 3-year-olds on a self-directed hiding version of the *Spin the Pots* task. Furthermore, results indicated that across different measures, there were small but significant differences between boys and girls, where girls consistently outperformed boys on this task. Prior studies using the *Spin the Pots* task reported *Spin the Pots* as part of an executive functioning composite score (see Bernier, Carlson, Deschênes, & Matte-Gagné, 2012; Hughes & Ensor, 2005). Taken together, our novel approach of determining the expected value and comparing it to a large sample of children across a wide age range lends support for the continuation of the task as a useful multi-trial non-verbal measure of working memory to be included in cognitive batteries and longitudinal studies.

Previous research documenting multiple object search working memory tasks (Diamond et al., 1997) has identified sex differences in inhibitory control and, in particular, lower perseveration rates in girls than boys. A meta-analysis of children ages 3 months to 13 years has also documented a moderate sex difference for inhibitory control, suggesting that girls have a better ability to control responses or behaviors than boys (Else-Quest, Hyde, Goldsmith, & Van Hulle, 2006). The present findings are consistent with the meta-analysis, showing a lower rate of perseveration in 4-year-old females, even though the rate of perseveration was low overall. Specifically, our results indicated that girls completed the task in significantly fewer trials, took significantly less time, had significantly fewer perseverations, and had fewer other types of error as well. This finding is consistent with those of Blakey and Carroll (2018) who examined the intersection of cognitive flexibility, inhibitory control, and working memory. They found that better inhibitory control was associated with better performance on the

Spin the Pots task in 2- and 3-year-olds. Specifically, 2- and 3-year-olds who were better at ignoring distracting stimuli had higher scores on *Spin the Pots*. These results are entirely consistent with task demands in *Spin the Pots*, which requires privileging item cues, while inhibiting attention to location cues. It is also feasible that the sex-related difference was due to linguistic differences between boys and girls which allowed girls to better identify or label the colors of the cups and stickers. Additional research exploring this sex difference is needed. This research is warranted given prior studies documenting an advantage for males on spatial reasoning tasks (Joh, 2016; Mandolesi et al., 2009), a female advantage on linguistic tasks, as well as research on the intersection of cognitive flexibility and working memory (Blakey & Carroll, 2018).

Our pattern of age-related and sex-related changes differs, however, from previous *Spin the Pots* studies in this age range. In a number of studies, there were no reported sex-related differences (e.g., Blakey & Carroll, 2018; Blakey et al., 2016) but these studies included smaller sample sizes. Findings on age-related differences have been mixed, with some researchers reporting age-related differences (Hughes & Ensor, 2007) and others reporting no differences in 2- to 4-year-olds (e.g., Blakey et al., 2016). Children in the present study participated in a self-ordered child-directed hiding phase prior to testing (children hid the stickers themselves) rather than an experimenter-directed or -assisted hiding phase. We based our decision on the prior work on self-ordered WM tasks (see Pinto-Hamuy & Linck, 1965; Diamond et al., 1997; Petrides & Milner, 1982). While this self-ordered hiding might make retrieval easier for some children, additional working memory demands may have been present for those children who took longer to hide their stickers or who did not try to remember the color of the cups. These findings suggest that 2-year-olds may be highly sensitive to additional demands of the self-ordered hiding phase. Prior manipulations of the number of locations and/or the number of hidden objects to retrieve have shown that children are sensitive to cognitive load on the *Spin the Pots* task (Batchelor, Inglis, & Gilmore, 2015; Hammond, Müller, Carpendale, Bibok, & Liebermann-Finestone, 2012; Hostinar, Stellem, Schaefer, Carlson, & Gunnar, 2012; see also Barr et al., 2016 for analogous findings on a different measure). Additional investigation directly comparing performance on self-ordered versus experimenter-ordered hiding phases prior to the test phase is warranted. Examining these parameters along with some of the previously tested cognitive load manipulations mentioned above could refine the *Spin the Pots* task to allow for better assessment of VWM at different ages and on multiple occasions. Findings from these investigations would also be relevant to early educators as potential ways to improve performance by children who are having difficulty keeping track of the information.

One advantage of the child-directed self-ordered protocol was that we could examine whether strategy use during the hiding phase predicted VWM. We found that many of the children did not take a strategic approach to hiding the stickers and randomly placed the stickers under the pots. Some children, however, behaved in more deliberate ways during this phase, some placing the stickers in a linear fashion and others matching the color of the sticker to the color of the pots. Given that the hiding event occurred before the search phase, children were not aware of the search task parameters in advance. Although ~ 50% of children across age used a strategy (linear or color), only the color strategy could actually help them on this scrambled version of the task. Had we tested them using a stationary pots task, the linear strategy may have been effective. While only a small subset of the entire sample (~20%) used the color strategy, it significantly predicted a better success rate on the task, but the linear strategy adopted by more children was not a significant predictor of performance. The effectiveness of the color strategy is consistent with findings from other spatial reasoning tasks (Diamond et al., 1997; Joh, 2016). The color strategy was used equally by boys and girls. Future research using a training study method is needed to test whether teaching children a color-hiding strategy or demonstrating a color-hiding strategy increases task performance (see also Joh, 2016).

Spin the Pots is a complex task. While *Spin the Pots* was initially designed and tested with 2-year-olds, our results suggest that performance is poor at this age. This may be due to the complex nature of the task. One reason is that the spatial location changes each trial as the apparatus is rotated 180 degrees (similar to *scrambled boxes*). A second reason is that two cups remain empty, meaning that children have to keep track of and update information about the two empty cups in addition to the new empty cups from which they retrieve stickers. In addition, there are documented age-related changes in children's ability to both use landmarks to track objects, and to track objects through spatial rotation to infer locations of hidden objects (Okamoto-Barth & Call, 2008). Specifically, in a two-location search task, 3-year-olds performed well on visible displacements and invisible rotation if a marker was on top of the target cup, but poorly with invisible displacement in the absence of landmark cues. Inferring rotations was achieved later in development, with 5-year-olds tracking 180 degree rotations independent of landmark cues. This partially accounts for why tracking eight locations, even with highly distinctive cups in the *Spin the Pots* task, is challenging for 2- to 4-year-olds. Our version of *Spin the Pots* increases the task complexity further as children are tracking both their own retrieval of stickers based on their hiding locations, as well as subsequent successful and unsuccessful search behaviors across trials. A limitation of the study is that there is inconsistency in the research in the hiding phase and the results of the present study are limited to a self-hiding protocol. As mentioned above future research should systematically compare how performance differs as a function of self-ordered versus experimenter-directed hiding.

5 | IMPLICATIONS

The present findings add to a small but growing body of literature on developmental and sex-related changes in WM during early childhood. The development of visual WM, in particular, is crucial for acquiring skills in complex tasks such as mathematics and problem solving that require substantial information tracking and updating, and ultimately impact academic success. Both VWM and spatial skills are critical for early education and are precursors to success in Science, Technology, Engineering, and Math (STEM) disciplines. There are significant differences in the home environment in activities that enhance spatial transformation skills. For example, parents of 2- to 4- year-olds are both more engaged and use more spatial language with boys than they do with girls, and expose boys to more difficult puzzles than girls (see Levine, Ratliff, Huttenlocher, & Cannon, 2012). More frequent puzzle play was associated with better spatial transformation for boys, and higher puzzle quality was associated with better spatial transformation for girls. It is important to integrate findings from spatial tasks that typically show male advantages for representing spatial information (e.g., Levine, Huttenlocher, Taylor, & Langrock, 1999) with others that show a female advantage in integration of individual item cues during visual search (e.g., Diamond, 1997).

The reported age differences also have important implications for early education. Specifically, educators should consider different strategies to enhance learning for younger children who have greater difficulty on these tasks. Perhaps the addition of color cues during tasks that have spatial elements such as number lines, sorting objects, pattern detection, mental rotation, puzzles, and the translation of geometric shapes may facilitate performance in boys and girls by taking advantage of their ability to use color cues successfully at an early age.

ACKNOWLEDGMENTS

This data collection was made possible by a grant to Rachel Barr and Peter Gerhardstein (BCS-1023772) and Francys Subiaul (BCS-0748717) both from the National Science Foundation. The authors thank Amanda Grenell and Ruth Speidel for their assistance with data collection. Courtney Nugent, Christina Sharkey, Sydney Simmons, and Ruthie Speidel helped in coding this large dataset. Special thanks to Joshuah Touyz for his consultation on the Monte Carlo approach. The authors declare no conflicts or financial interests or holdings. We thank the museums for allowing us to use their facilities and families for their participation in the study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Laura Zimmermann https://orcid.org/0000-0002-3215-5248
Hannah Frank https://orcid.org/0000-0003-2396-4585
Francys Subiaul https://orcid.org/0000-0002-5873-9524
Rachel Barr https://orcid.org/0000-0002-5855-9718

REFERENCES

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Atkinson, R. C., & Shiffrin, R. M. (1971). The control processes of short-term memory. Institute for Mathematical Studies in the Social. Sciences, Stanford University.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Baddeley, A. D., & Hitch, G. J. (1974). In G. A. Bower (Ed.), Working memory. In *The psychology of learning and motivation* (vol. *8*, pp. 47–89). New York, NY: Academic Press.

Barr, R., Moser, A., Rusnak, S., Zimmermann, L., Dickerson, K., Lee, H., & Gerhardstein, P. (2016). The impact of memory load and perceptual cues on puzzle learning by 2-year-olds. *Developmental Psychobiology*, *58*, 817–828. https://doi.org/10.1002/dev.21450

Batchelor, S., Inglis, M., & Gilmore, C. (2015). Spontaneous focusing on numerosity and the arithmetic advantage. *Learning and Instruction*, *40*, 79–88. https://doi.org/10.1016/j.learninstruc.2015.09.005

Bauer, P. J., & Zelazo, P. D. (2013). NIH toolbox cognitive battery (CB): Summary, conclusions, and implications for cognitive development. *Monographs of the Society for Research in Child Development*; *78*(4):133–146. Chapter IX.

Beck, D. M., Schaefer, C., Pang, K., & Carlson, S. M. (2011). Executive function in preschool children: Test–retest reliability. *Journal of Cognition and Development*, *12*(2), 169–193. https://doi.org/10.1080/15248372.2011.563485

Bernier, A., Carlson, S. M., Deschênes, M., & Matte-Gagné, C. (2012). Social factors in the development of early executive functioning: A closer look at the caregiving environment. *Developmental Science*, *15*(1), 12–24. https://doi.org/10.1111/j.1467-7687.2011.01093.x

Bernier, A., Carlson, S. M., & Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development*, *81*, 326–339. https://doi.org/10.1111/j.1467-8624.2009.01397.x

Best, J. R., & Miller, P. H. (2010). A Developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Blakey, E., & Carroll, D. J. (2018). Not all distractions are the same: investigating why preschoolers make distraction errors when switching. *Child Development*, *89*, 609–619. https://doi.org/10.1111/cdev.12721

Blakey, E., Visser, I., & Carroll, D. J. (2016). Different executive functions support different kinds of cognitive flexibility: Evidence from 2-, 3-, and 4-year-olds. *Child Development*, *87*, 513–526. https://doi.org/10.1111/cdev.12468

Bolenz, F., Reiter, A. M. F., & Eppinger, B. (2017). Developmental changes in learning: computational mechanisms and social influences. *Frontiers in Psychology*, *8*, 2048. https://doi.org/10.3389/fpsyg.2017.02048

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children’s mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293. https://doi.org/10.1207/S15326942DN1903_3

Choi, K., Kirkorian, H. L., & Pempek, T. A. (2017). Understanding the transfer deficit: Contextual mismatch, proactive interference, and working memory affect toddlers’ video-based transfer. *Child Development*, *89*(4), 1378–1393.https://doi.org/10.1111/cdev.12810

Diamond, A. (1990). The development and neural bases of memory functions as indexed by the AB and delayed response tasks in human infants and infant monkeys. *Annals of the New York Academy of Sciences*, *608*, 266–317. https://doi.org/10.1111/j.1749-6632.1990.tb48900.x

Diamond, A., Prevor, M., Callender, G., & Druin, D. (1997). Prefrontal cortex cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development*, *62*(4), i–206. https://doi.org/10.2307/1166208

Dumontheil, I., & Klingbert, T. (2012). Brain activity during a visuospatial working memory task predicts arithmetical performance 2 years later. *Cerebral Cortex*, *22*, 1078–1085. https://doi.org/10.1093/cercor/bhr175

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, *132*(1), 33. https://doi.org/10.1037/0033-2909.132.1.33

Espy, K. A., & Bull, R. (2005). Inhibitory processes in young children and individual variation in short-term memory. *Developmental Neuropsychology*, *28*(2), 669–688. https://doi.org/10.1207/s15326942dn2802_6

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*(1), 31. https://doi.org/10.1037/0033-2909.134.1.31

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Science*, *3*, 410–419. https://doi.org/10.1016/S1364-6613(99)01388-1

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, *18*, 1–16. https://doi.org/10.1002/acp.934

Hammond, S. I., Müller, U., Carpendale, J. I., Bibok, M. B., & Liebermann-Finestone, D. P. (2012). The effects of parental scaffolding on preschoolers' executive function. *Developmental Psychology*, *48*(1), 271. https://doi.org/10.1037/a0025519

Hostinar, C. E., Stellern, S. A., Schaefer, C., Carlson, S. M., & Gunnar, M. R. (2012). Associations between early life adversity and executive function in children adopted internationally from orphanages. *Proceedings of the National Academy of Sciences*, *109*(Supplement 2), 17208–17212. https://doi.org/10.1073/pnas.1121246109

Huber, B., Yeates, M., Meyer, D., Fleckhammer, L., & Kaufman, J. (2018). The effects of screen media content on young children's executive functioning. *Journal of Experimental Child Psychology*, *170*, 72–85. https://doi.org/10.1016/j.jecp.2018.01.006

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*, 233–253. https://doi.org/10.1111/j.2044-835X.1998.tb00921.x

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year-olds: A family affair? *Developmental Neuropsychology*, *28*, 645–668. https://doi.org/10.1207/s15326942dn2802_5

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology*, *43*, 1447–1459. https://doi.org/10.1037/0012-1649.43.6.1447

Joh, A. S. (2016). Training effects and sex difference in preschoolers' spatial reasoning ability. *Developmental Psychobiology*, *58*, 896–908.

Johansson, M., Marciszko, C., Brocki, K., & Bohlin, G. (2015). Individual differences in early executive functions: A longitudinal study from 12 to 36 months. *Infant and Child Development*, *25*, 533–549. https://doi.org/10.1002/icd.1952

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Levine, S. C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex difference in spatial ability. *Developmental Psychology*, *35*, 940–949. https://doi.org/10.1037/0012-1649.35.4.940

Levine, S. C., Ratliff, K., Cannon, J., & Huttenlocher, J. (2012). Early puzzle play: A predictor of preschoolers' spatial transformation skill. *Developmental Psychology*, *48*, 530–542. https://doi.org/10.1037/a0025913

Logie, R. H. (1995). *Visuo-spatial working memory*. Hove, UK: Lawrence Eribaum Associates.

Luciana, M., & Nelson, C. A. (1998). The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. *Neuropsychologia*, *36*, 273–293. https://doi.org/10.1016/S0028-3932(97)00109-7

Mandolesi, L., Petrosini, L., Menghini, D., Addona, F., & Vicari, S. (2009). Children's radial arm maze performance as a function of age and sex. *International Journal of Developmental Neuroscience*, *27*, 789–797. https://doi.org/10.1016/j.ijdevneu.2009.08.010

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Moser, A., Zimmermann, L., Dickerson, K., Grenell, A., Barr, R., & Gerhardstein, P. (2015). They can interact but can they learn? Toddlers' transfer learning from touchscreens and television. *Journal of Experimental Child Psychology*, *137*, 137–155. https://doi.org/10.1016/j.jecp.2015.04.002

Müller, U., Liebermann-Finestone, D. P., Carpendale, J. I., Hammond, S. I., & Bibok, M. B. (2012). Knowing minds, controlling actions: The developmental relations between theory of mind and executive function from 2 to 4 years of age. *Journal of Experimental Child Psychology*, *111*, 331–348. https://doi.org/10.1016/j.jecp.2011.08.014

Okamoto-Barth, S., & Call, J. (2008). Tracking and inferring spatial rotation by children and great apes. *Developmental Psychology*, *44*, 1396. https://doi.org/10.1037/a0012594

Overman, W. H., Pate, B. J., Moore, K., & Peuster, A. (1996). Ontogeny of place learning in childen as measured in the radial arm maze, morris search task, and open field task. *Behavioral Neuroscience*, *110*, 1205–1228.

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal lobe lesions in man. *Neuropsychologia*, *20*, 249–262. https://doi.org/10.1016/0028-3932(82)90100-2

Pinto-Hamuy, T., & Linck, P. (1965). Effect of frontal lesions on performance of sequential tasks by monkeys. *Experimental Neurology*, *12*(1), 96–107.

Riggins, T., Cheatham, C. L., Stark, E., & Bauer, P. (2013). Elicited imitation performance at 20 months predicts memory abilities in school-aged children. *Journal of Cognition and Development*, *14*, 593–606. https://doi.org/10.1080/15248372.2012.689392

Roman, G. D., Ensor, R., & Hughes, C. (2016). Does executive function mediate the path from mothers' depressive symptoms to young children's problem behaviors? *Journal of Experimental Child Psychology*, *142*, 158–170. https://doi.org/10.1016/j.jecp.2015.09.022

Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, *28*, 207–218.

Subiaul, F., Zimmermann, L., Renner, E., Schilder, B., & Barr, R. (2016). Defining elemental imitation mechanisms: A comparison of cognitive and motor-spatial imitation learning across object- and computer-based tasks. *Journal of Cognition and Development*, *17*, 221–243. https://doi.org/10.1080/15248372.2015.1053483

van den Bos, W., Bruckner, R., Nassar, M. R., Mata, R., & Eppinger, B. (2018). Computational neuroscience across the lifespan: Promises and pitfalls. *Developmental Cognitive Neuroscience*, *33*, 42–53. https://doi.org/10.1016/j.dcn.2017.09.008

Zimmermann, L., Moser, A., Grenell, A., Dickerson, K., Yao, Q., Gerhardstein, P., & Barr, R. (2015). Do semantic contextual cues facilitate transfer learning from video in toddlers? *Frontiers in Psychology*, *6*, 561. https://doi.org/10.3389/fpsyg.2015.00561

Zimmermann, L., Moser, A., Lee, H., Gerhardstein, P., & Barr, R. (2017). The ghost in the touchscreen: Social scaffolds promote learning by toddlers. *Child Development*, *88*, 2013–2025. https://doi.org/10.1111/cdev.12683

How to cite this article: Zimmermann L, Frank HE, Subiaul F, Barr R. Applying computational modeling to assess age-, sex-, and strategy-related differences in *Spin the Pots*, a working memory task for 2- to 4-year-olds. *Dev Psychobiol*. 2020;00:1–12. https://doi.org/10.1002/dev.22016