The development of visual working memory capacity during early childhood
Vanessa R. Simmering*
Department of Psychology, University of Wisconsin–Madison, 1202 W. Johnson St., Madison, WI 53706, USA

Abstract
The change detection task has been used in dozens of studies with adults to measure visual working memory capacity. Two studies have recently tested children in this task, suggesting a gradual increase in capacity from 5 years to adulthood. These results contrast with findings from an infant looking paradigm suggesting that capacity reaches adult-like levels within the first year. The current study adapted the change detection task for use with children younger than 5 years to test whether the standard version of the task was too difficult and may have underestimated children’s capacity. Results showed that 3- and 4-year-olds could successfully complete this modified task and that capacity increased roughly linearly, from 2 or 3 items during this period to 3 or 4 items between 5 and 7 years. Furthermore, performance did not differ significantly between the modified version and a replication of the standard version with 5- and 7-year-olds. Thus, there is no evidence that previous research with the change detection task underestimated children’s capacity. Further research is needed to understand how performance relates across the infant looking task and change detection to provide a more complete picture of visual working memory capacity over development.

Introduction
Visual working memory (VWM) provides a critical foundation for our understanding of the visual world around us. As we move our eyes to survey our environment, VWM provides a bridge between the fleeting perceptual representation formed in a single fixation and our enduring long-term...
representation of familiar objects, people, and places. Without the ability to represent visual information in working memory during eye movements, our experience would be a series of disjointed snapshots. Decades of research have revealed that VWM has a severely limited capacity of just 3–5 simple items in young adults (Cowan, 2010). Beginning with Luck and Vogel’s (1997) seminal article, the change detection task has been a common method for testing VWM capacity. In this task, a small number of items are briefly presented in a memory array. After a short delay, a test array is presented in which either all of the items match the memory array (“same” trials) or 1 item has changed (“different” trials). This array remains visible until the participant responds by pressing a key corresponding to his or her judgment of same or different. Capacity can be estimated from performance in this task using a formula proposed by Pashler (1988), which subtracts the effect of “guessing” (i.e., responding different on no-change trials) from correct performance on change trials. Although dozens of studies have tested adults’ capacity with this task, typically reported at 3 or 4 simple objects, only a few have tested the development of VWM capacity in this way.

Capacity is a concept of considerable interest within psychology because it has been proposed as the foundation for a range of intelligent behaviors, including language comprehension (e.g., Just & Carpenter, 1992), executive control in rule use (e.g., Marcovitch, Boseyoski, Knapp, & Kane, 2010), and performance on standardized tests of scholastic performance (e.g., Cowan et al., 2005). For instance, Cowan et al. (2005) conducted a wide-ranging study in which capacity was estimated from numerous verbal and visual working memory tasks, all of which showed moderate to strong correlations with measures of working memory and standardized test scores. With regard to VWM specifically, capacity increased gradually over development, from approximately 3.5 items at 7 and 8 years, to 4.4 items at 9 and 10 years, to 4.8 items at 11 and 12 years, to 5.7 items for college students (note that estimates are averaged across two experiments reported by Cowan et al.). The estimates for adults in this study are slightly higher than those typically reported in the literature, but they follow from the estimates derived for children in the same task.

These findings suggest that capacity can provide an index of general cognitive ability, as measured by a range of tasks as well as strong continuity over development beginning during the grade school years. Riggs, McTaggart, Simpson, and Freeman (2006) expanded on Cowan et al. (2005) findings by testing younger children in the change detection task. They found similar increases at an earlier stage of development, with capacity estimated to be 1.52 items at 5 years, 2.89 items at 7 years, and 3.83 items at 10 years. Taken together, the studies by Cowan and colleagues and Riggs and colleagues suggest continuity of VWM from 5 years to adulthood.

Studies with infants, however, have revealed contradictory results. In a preferential looking task modeled after the Luck and Vogel (1997) change detection paradigm, Ross-Sheehy, Oakes, and Luck (2003) found that VWM capacity increased rapidly during early infancy, reaching adult-like levels by the end of the first year. In this task, infants were seated in front of two video displays, each containing a small number of items that blinked on and off over a 20-s trial. On the no-change display, the items remained the same after each blink (similar to “same” trials in change detection), whereas one item changed with each blink on the change display (similar to “different” trials in change detection). Capacity was estimated as the highest number of items per display at which infants showed a reliable preference for the change display. Ross-Sheehy and colleagues estimated capacity to be 1 item at 6.5 months, 3 or 4 items at 10.5 months, and at least 3 items at 12.5 months (12-month-olds were not tested beyond 3 items per display) using this change preference task, suggesting that capacity reached adult-like levels within the first year.

Riggs et al. (2006) suggested three possible explanations behind the developmental inconsistency between the infant change preference task and the change detection task used with older children.

---

1 Cowan (2001; see also Cowan et al., 2005) proposed a modified formula for tasks on which a single item is cued on test. This formula, which produces lower estimates than Pashler’s (1988) formula, was used to calculate the results reported from Cowan et al. (2005); all other capacity estimates from change detection discussed here were calculated using Pashler’s (1988) formula.

2 Additional infant studies have been proposed to assess VWM capacity through other methods (e.g., Feigenson & Carey, 2003; Rose, Feldman, & Jankowski, 2001). However, these paradigms differ in critical ways from the change detection and change preference tasks discussed here, and they may rely on cognitive processes in addition to VWM (e.g., motor planning, episodic memory, long-term memory). Due to such differences, these studies are not included for comparison here.
First, they argued that infants may be able to show a reliable change preference without remembering all of the items; a further study by Oakes, Messenger, Ross-Sheehy, and Luck (2009), however, showed that this is not likely to be the case. In this study, Oakes and colleagues tested 6-month-olds in a version of their change preference task in which all 3 items on the change display were replaced with every blink; therefore, if memory for 1 item is sufficient to drive a preference at a set size of 3, 6-month-olds should reliably prefer the change display in this condition. The 6-month-olds showed no preference in this experiment; thus, memory for a single item within the multi-item displays was not sufficient to drive a robust change preference.

A second potential explanation proposed by Riggs et al. (2006) was that looking behavior in the change preference task is more passive than the verbal decisions required by the change detection task. As a result, they suggested that 2 tasks may tap into different VWM processes, although they did not specify which processes these may be. As a third possibility, Riggs and colleagues stated that VWM capacity could reach adult-like levels during infancy and that the developmental improvements in the change detection task during childhood could result from other cognitive factors besides capacity. Developmental capacity research in general has shown that different tasks produce different estimates, suggesting that the details of the task context may either mask or boost capacity (see Simmering & Perone, 2011, for a discussion).

These potential explanations all point to the central question raised by these studies: What is the source of developmental change in performance? This question must be applied to both tasks because they are both purported to rely on VWM capacity. Is capacity changing rapidly during the first year, as evidenced by performance in the change preference task, only to be masked later in the change detection task? Or are changes in the infant task reflecting other processes, whereas change detection reveals capacity more directly? It is also possible that changes in both tasks are due to cognitive factors other than capacity.

Three articles following Ross-Sheehy et al. (2003) have provided empirical evidence to explain performance changes during infancy. First, Oakes, Ross-Sheehy, and Luck (2006) showed that the ability to bind colors to locations in the preferential looking task coincided developmentally with the increase in capacity from 1 to 2 or more items at 7.5 months of age. As described above, Oakes et al. (2009) showed that young infants did not prefer the changing stream at a set size of 3 even when all 3 items in the display changed with every blink. The authors interpreted this result as further evidence that infants must be able to bind colors to locations in order to remember more than 1 item simultaneously. Lastly, Ross-Sheehy, Oakes, and Luck (2010) showed that young infants will prefer change displays of 3 items if the item that changes is cued, essentially reducing the memory load from 3 items to 1 item.

These studies illustrate that the explanation for changes in performance during infancy relies primarily on emerging abilities to properly individuate and/or attend to items across presentations that allow them to detect the changes occurring. Perone, Simmering, and Spencer (2011) challenged this conclusion by showing that a computational model, with no mechanism for binding colors to space, can capture younger and older infants’ performance in the critical conditions from Ross-Sheehy et al. (2003) and Oakes et al. (2009). The model proposed by Perone and colleagues (2011) suggests that infants’ emerging ability to detect changes in displays with more than 1 item depends on their ability to stably remember the items on the no-change display as well as on the change display. Moreover, their model produced robust change preferences for displays of 3 items while holding only 1 or 2 items in memory at a time. Rather than the number of items being the defining factor producing a change preference, it is the stability of these representations and how memory interacts with the dynamics of fixation that produces the pattern of results from Ross-Sheehy et al. (2003) and Oakes et al. (2009). Thus, they conclude that VWM abilities do increase between 6 and 10 months of age but that performance on the change preference task does not directly reveal the capacity of VWM.

The computational results by Perone and colleagues (2011) suggest that developmental improvements in the change preference task do not solely reflect changes in capacity, although capacity may increase slightly during the first year. Even with this alternative explanation, infants’ capacity would need to be between 1 and 2 items by 10 months, which is comparable to the estimate for 5-year-olds derived by Riggs et al. (2006). Is it possible that capacity increases between 6 and 10 months, only to plateau until 5 years, and then begins to increase again? Although this developmental trajectory
seems unlikely, further empirical evidence is needed to evaluate the possibility. Critically, no previous studies have tested capacity in either the change preference task or change detection task between 1 and 5 years of age. A primary goal of the current study was to fill this empirical gap by adapting the change detection task for use with children younger than 5 years. Studying younger children may provide further evidence for the protracted trajectory revealed by Riggs et al. (2006) and Cowan et al. (2005), in contrast to the rapid increase suggested by infant studies. A second question addressed by the current study is whether the change detection task used by Riggs and colleagues was too difficult for children, potentially underestimating capacity with burdensome task demands. As such, a task that is adapted to young children's abilities may provide a better index of VWM capacity during childhood.

To address these two questions, I conducted a pilot study to probe the youngest age that could reliably perform a more child-friendly version of the task. To modify the change detection task for young children, I used the study by Riggs et al. (2006) as a starting point. In their study, the authors modified the standard task in three ways relative to the task used with adults (as used in Cowan et al., 2005; Luck & Vogel, 1997, and others). First, Riggs and colleagues extended the duration of the memory array from 250 to 500 ms to allow more time for encoding. Second, the total number of trials required of each participant was reduced to 12 (6 change and 6 no-change) at each set size of 1 to 5 items (compared with, e.g., 40 trials per set size [Luck & Vogel, 1997]). Finally, the trials were blocked by set size and ordered such that the easier set sizes (1 and 2 items) were first, followed by the higher, more difficult set sizes.

Using these task details as a starting point, pilot data with 3- and 4-year-olds revealed two specific problems experienced by young children that prompted additional modifications. First, young children had difficulty in understanding the required comparison between the memory and test arrays within a trial. For instance, some children would compare the items within the test array, whereas other children would compare the items in the test array on one trial with the memory array on the subsequent trial. To help children understand the appropriate comparison between arrays, the experimenter described the task as a card-matching game and then presented the items in each array on a rectangular background to identify them as a “card”. To reduce interference across trials, the spatial location of the cards alternated between the left and right sides of the screen from trial to trial. Fig. 1 shows an example of two trials using the alternating side presentation of the cards.

The second problem children had was viewing and encoding all of the stimuli during the short presentation of the memory array (500 ms). Children tended to look around the experimental room between trials, so extra display time was needed to insure that they oriented to the computer screen and to all items within the display. This was addressed in three ways. First, each trial was initiated with a key press by the experimenter, who waited until children were attending and ready to continue. Second, the duration of the memory array was increased further, from 500 ms to 2 s, for 3- and 4-year-olds. Lastly, rather than presenting the items in randomly selected locations on each trial, they were clustered near the center of the card at fixed locations on each trial (see Fig. 1). This reduced the need for children to scan the screen to find all of the items during encoding.

With these modifications, children as young as 3.5 years could reliably perform the change detection task. Four age groups of children (3.5, 4.25, 5.0, and 7.5 years) were selected to participate. All participants completed the modified “card” version change detection task. In addition, a second group of 5- and 7-year-olds completed a replication of Riggs et al. (2006) task to allow for a direct test of whether these modifications increased capacity estimates relative to previous studies. This comparison serves not only to insure that the modified task presented here provides a measure of VWM comparable to the standard version of the change detection task but also to replicate Riggs and colleagues’ findings with 5-year-olds, an age group that has not otherwise been tested in change detection.

3 I chose to keep the stimulus presentation duration at 500 ms for older children to remain consistent with Riggs et al. (2006) paradigm for replication purposes, although Cowan, AuBuchon, Gilchrist, Ricker, and Saults (2011) found that increasing the stimulus presentation time did not significantly affect 7-year-olds’ performance on a related task.
Method

Participants

In total, 56 children completed the modified "card" version of the change detection task: 14 3-year-olds ($M = 3$ years 4.68 months, $SD = 2.35$ months), 14 4-year-olds ($M = 4$ years 2.56 months, $SD = 1.77$ months), 14 5-year-olds ($M = 5$ years 0.36 months, $SD = 0.78$ months), and 14 7-year-olds ($M = 7$ years 6.16 months, $SD = 3.94$ months). A second group of 14 5-year-olds ($M = 5$ years 3.28 months, $SD = 5.43$ months) and 14 7-year-olds ($M = 7$ years 5.45 months, $SD = 1.92$ months) participated in a replication of the task used by Riggs et al. (2006) for comparison. An additional 22 children participated but were excluded from analyses for the following reasons: did not understand the task (8 children: 7 3-year-olds and 1 5-year-old), chose to end before completing a set size of 3 (10 children: 4 3-year-olds and 6 5-year-olds), experimenter error (1 3-year-old), equipment failure (2 children: 1 3-year-old and 1 5-year-old), and vision abnormality (1 7-year-old). All participants reported normal or corrected-to-normal vision and no family history of color-blindness. Informed consent was obtained from parents of child participants. Participants were recruited from a volunteer database maintained by the university and received a toy for participating.

Apparatus

In both versions of the task, the change detection task was explained to children using flashcards ($3 \times 3$ inches) that showed set size 1 (SS1), SS2, and SS3. For most participants ($n = 63$), the primary task was completed on an 18-inch CRT display connected to a Macintosh G4 computer, with stimulus presentation controlled by Matlab 5.2 (MathWorks) using the Psychophysics Toolbox extensions (Version 2) (Brainard, 1997; Pelli, 1997). The remaining participants completed the task on a 15.4-inch widescreen Dell Latitude E6500 laptop computer, with stimulus
presentation controlled by Matlab 10.2 using the Psychophysics Toolbox extension (Version 3) (Kleiner, Brainard, & Pelli, 2007). Stimulus presentation was adjusted to appear approximately the same size on both screens.

Children were seated approximately 26 inches from the screen. In the standard version of the task, stimulus arrays consisted of 1–5 colored 1-inch squares presented at random locations on a gray background. For the modified version, the background on the computer monitor was black and the colored squares were enclosed within a gray rectangular frame (5.75 inches tall \times 4.75 inches wide). The gray frame was centered vertically on the screen on every trial; each block began with the frame centered horizontally on the left half of the screen and alternating sides over trials (see Fig. 1). Within the frame, stimuli could appear at any of five equally spaced positions in a 3-inch diameter circle around the center of the frame. For SS2, SS3, and SS4, the stimuli appeared in neighboring positions; all five positions were filled for SS5. For each set size, the positions were chosen randomly for the first trial but remained constant across the 12 trials in that set size. Colors were drawn randomly without replacement from the set of eight colors used by Luck and Vogel (1997): black, blue, cyan, green, red, violet, white, and yellow.4

Procedure

A female experimenter guided children through the task. She described the task as a matching game in which children needed to help a teddy bear find cards that matched (the teddy bear was typically used only for younger children; older children were sufficiently motivated without it). For training, the experimenter demonstrated the task using flashcards showing 1, 2, or 3 colored squares within a gray frame. Flashcards were placed on either the right or left side (alternating across trials) of a large sheet of cardstock in the experimenter’s lap. The first card was shown for approximately 2 s, and the child was instructed to “look at the picture and remember the colors”; the terminology “picture” was used to build on children's experience playing commercially available picture-matching card games. The card was then removed, and after a brief delay the second card for the trial was shown in the same location. The experimenter then asked the child if the two cards matched. After the child responded, the experimenter placed both cards side-by-side and praised or corrected the child as needed, explicitly pointing out the changed item to insure that the child understood the proper comparison. For all children, the flashcard trials were presented in the same order: SS1 no-change, SS1 change, SS2 change, SS2 no-change, SS3 change, and SS3 no-change.

Once the child understood the task, the experimenter began the computerized version. The first block consisted of practice trials, with the memory array presented for 2 s, a delay of 900 ms, and the test array presented until a response was entered. Children responded verbally, and the experimenter entered the response on a keyboard. When the response was entered, a chime played if the response was correct. This positive feedback was included to help children stay motivated. If children seemed fatigued, they were offered a break between set size blocks.

The practice block consisted of eight trials in random order: four trials to SS1 and four trials to SS2 (half of the trials at each set size were change trials and half were no-change trials). Between each block of trials, the child was offered a short break. The order of test blocks was semirandomly assigned (see Riggs et al., 2006) such that the first two blocks were either SS1 or SS2; for 3- and 4-year-olds, the third block was always SS3; for 5- and 7-year-olds, SS3, SS4, and SS5 were randomly ordered across Blocks 3, 4, and 5. A different order was used for younger children because they often chose to end their participation before SS4 and SS5; note that these children were still included in analyses (only those ending before SS3 were excluded from analyses, as described above). These high-set-size blocks were difficult for all children, and young children tended to become discouraged as their performance.

4 Note that Luck and Vogel (1997) also used a brown color, but pilot data suggested that children found this color to be too similar to red to reliably distinguish; as such, we chose to exclude it from our color set. RGB values for the colored squares and background were as follows: black (0, 0, 0), blue (0, 0, 255), cyan (0, 255, 255), green (0, 255, 0), gray (150, 150, 150), red (255, 0, 0), violet (238, 130, 238), white (255, 255, 255), and yellow (255, 255, 0). See Vogel, Woodman, and Luck (2001) for approximate CIE values.
Results

Modified "card" task

Participants' responses were classified as correct rejections (correct no-change trials), hits (correct change trials), misses (incorrect change trials), or false alarms (incorrect no-change trials). Fig. 2 shows the distributions of these response types for each age group separately. Note that correct rejections and false alarms sum to 1.0 (all no-change trials) and hits and misses sum to 1.0 (all change trials). As this figure shows, correct rejections and hits were the most common responses across ages and set sizes. In addition, performance decreased (i.e., errors increased) with set size, especially for younger children. It is important to note that, although young children performed worse than older

---

5 Due to a programming error, the randomization of trials in the modified version of the task used with 7-year-olds occasionally presented unequal numbers of change and no-change trials within the test blocks. Overall, the average number of change and no-change trials differed by less than 0.5, with change trials being slightly more frequent. Given the similarity between the results of different versions with 7-year-olds (described below), there is no reason to believe that this randomization error influenced children's performance.
children (whose results were similar to adults’ typical performance [e.g., Luck & Vogel, 1997]), all age
groups showed the same general pattern, with misses increasing more than false alarms over set sizes.
This suggests that young children were performing the task appropriately even though they were not
performing well; if they were responding randomly, we would expect higher rates of errors overall
with more similar performance across trial types rather than the typical increase in misses that adults
show as set size nears and surpasses capacity.

Performance was analyzed by computing overall percentage correct (collapsed across change and
no-change trials) separately for each set size and age group. As Fig. 3A shows, performance was high-
est for 7-year-olds and lowest for 3-year-olds, and it decreased markedly across set size for all age
groups. Because many of the younger children did not complete SS4 and SS5, overall analyses com-
pared only SS1, SS2, and SS3. An analysis of variance (ANOVA) with set size (1, 2, or 3) as a within-par-
ticipants factor and age (3, 4, 5, or 7 years) as a between-participants factor revealed significant main
effects of set size, $F(2, 104) = 48.95, p < .001, \eta^2_G = .207$, and age, $F(3, 52) = 10.61, p < .001, \eta^2_G = .257$,
which were subsumed by a significant Set Size $\times$ Age interaction, $F(6, 104) = 2.77, p < .05, \eta^2_G = .044$.

To understand how performance is changing over development, I first followed this interaction
with separate analyses across set sizes for each age group. These analyses should reveal the point
at which children’s performance drops significantly from ceiling (although young children’s “ceiling”
is low, with 3-year-olds scoring only ~88% correct at SS1). For 3- and 4-year-olds’ data, I performed
one-way ANOVAs with set size (1, 2, or 3) as a within-participants factor. These analyses revealed

![Fig. 3. Results from the modified version of the task. (A) Percentage correct performance (collapsed across change and no-
change trials) separately for each set size and age group. (B) Capacity estimates across set sizes (bars) and mean maximum
capacity estimates (line) separately for each age group. yr, years. Note that capacity can, at maximum, equal the set size in each
block. Error bars in both panels show 95% confidence intervals.](image-url)
significant effects for both age groups (3 years: \( F(2, 26) = 13.97, p < .001, \eta^2_g = .242 \); 4 years: \( F(2, 26) = 18.28, p < .001, \eta^2_g = .404 \)), and follow-up Tukey HSD tests (\( p < .05 \)) showed that performance was lower in SS3 than in SS1 and SS2, although SS1 and SS2 did not differ from each other (see Fig. 3A).

Because all 5-year-olds completed at least SS4, the one-way ANOVA on their data included set size (1, 2, 3, or 4) as a within-participants factor. This analysis revealed a significant main effect, \( F(3, 39) = 21.40, p < .001, \eta^2_g = .488 \). Tukey HSD tests showed no significant difference between SS1 and SS2, followed by significant decreases to SS3 and then SS4 (see Fig. 3A). All 7-year-olds completed all five set sizes, allowing for a one-way ANOVA with set size (1, 2, 3, 4, or 5) as a within-participants factor. This analysis also revealed a significant main effect, \( F(4, 52) = 17.49, p < .001, \eta^2_g = .449 \). Tukey HSD tests showed that SS1, SS2, and SS3 did not differ from each other but were higher than SS4 and SS5 (which also did not differ from each other). Thus, 3- to 5-year-olds' performance dropped when arrays surpassed 2 items, whereas 7-year-olds' performance did not drop until set size surpassed 3 items. This suggests that 7-year-olds are able to remember arrays of 3 or more items better than younger children.

Next, I compared age groups separately for each set size to see which showed significant improvements in age. Each one-way ANOVA with age (3, 4, 5, or 7 years) as a between-participants factor revealed significant effects (SS1: \( F(3, 52) = 4.46, p < .01, \eta^2_g = .204 \); SS2: \( F(3, 52) = 5.59, p < .01, \eta^2_g = .244 \); SS3: \( F(3, 52) = 10.29, p < .001, \eta^2_g = .373 \)). Tukey HSD follow-up tests showed that, in each case, 3-year-olds performed significantly worse than 7-year-olds; in addition, for SS3 only, 3- and 5-year-olds' performance, as well as 4- and 7-year-olds' performance, differed significantly (see Fig. 3A). Similar to the follow-ups of the set size effects for the different age groups, this suggests that 3- and 4-year-olds performed comparably, with some improvement at 5 years and further improvement by 7 years. An additional one-way ANOVA comparing 5- and 7-year-olds at SS4 also yielded a significant effect of age, \( F(1, 26) = 10.44, p < .01, \eta^2_g = .287 \), indicating superior performance by 7-year-olds at this set size (see Fig. 3A).

In addition to comparing percentage correct, capacity (\( K \)) for each age group was estimated using Pashler's (1988) formula, \( K = SS \times (H - FA)/(1 - FA) \), based on the hit (\( H \)) and false alarm (\( FA \)) rates for each set size (SS). Fig. 3B shows capacity estimates across set sizes (bars) separately for each age group; note that capacity can, at maximum, equal the set size in each block. Because of this limit on capacity estimates, it is not reasonable to analyze these data across set sizes. Moreover, calculating an average capacity estimate across blocks would result in an artificially low number by including set sizes below capacity. To avoid this limitation, I chose to use each participant's highest estimate for each set size block that the participant had completed (see, e.g., Olsson & Poom, 2005). An alternative approach is to take the estimate from the highest set size completed by each participant. Inspection of the data, however, revealed that this may underestimate young children's capacity if their performance was particularly poor at higher set sizes. Specifically, for 6 of the 3-year-olds, SS3 performance yielded lower estimates than SS2 or SS1 performance; moreover, only 1 4-year-old, 4 5-year-olds, and 8 7-year-olds had their highest estimate at the highest set size they had completed. As such, I chose to take the highest estimate regardless of which set size it was derived from.

Mean “maximum” capacity estimates are also shown in Fig. 3B (line). As can be seen in this figure, capacity estimates tend to be similar across set sizes that are near or beyond capacity for that age. For example, 3-year-olds show comparable mean capacity estimates for SS2 and SS3, and the maximum estimate for this age group is approximately 2 items; similarly, mean capacity estimates for 5-year-olds seem to peak at around SS3, which corresponds to the maximum estimate of just under 3 items. For 7-year-olds, capacity estimates continued rise from SS1 to SS3, were similar between SS3 and SS4, but then increased slightly to peak at around 4 items in SS5, which is near the maximum estimate for this age group. Across age groups, maximum capacity estimates were on average 0.58 items lower than the set size from which they were estimated, suggesting that estimates were not limited by the set sizes tested here.

I analyzed mean maximum capacity estimates in a one-way ANOVA with age (3, 4, 5, or 7 years) as a between-participants factor, which revealed a significant main effect, \( F(3, 52) = 22.18, p < .001, \eta^2_g = .561 \). Follow-up Tukey HSD tests showed that capacity did not differ between 3- and 4-year-olds or between 4- and 5-year-olds; however, capacity estimates were significantly lower for 3-year-olds compared with 5-year-olds and for all younger age groups compared with 7-year-olds. Thus, capacity
increased between 3 and 5 years and again between 5 and 7 years. This effect mirrors the differences in percentage correct described above, providing further evidence that children’s performance followed a similar pattern to adults’ typical performance; because hits figure more prominently in capacity estimates than correct rejections, capacity and percentage correct follow similar patterns only when performance does not differ dramatically between change and no-change trials. Thus, if children were responding randomly, or were strongly biased to one response type, capacity estimates and percentage correct performance would likely diverge.

Comparison with standard task

To determine whether the changes made in the modified version affected children’s capacity estimates, a second group of 5- and 7-year-olds completed a replication of Riggs et al. (2006) task. Fig. 4 shows the percentage correct performance and capacity estimates across age groups in this task with the comparable data from the modified version of the task. As this figure shows, results were similar across different versions of the task, with slightly better performance in the modified version. Following the

---

6 The method employed here differed in two minor ways from Riggs et al. (2006) procedure. First, colors were selected from the set of eight used in the “card” version instead of nine used by Riggs and colleagues. Second, colors were selected without replacement.
analyses above, I first analyzed percentage correct in a three-way ANOVA with set size (1, 2, or 3) as a within-participants factor and age (5 or 7 years) and version (modified or standard) as between-participants factors. This analysis revealed no significant main effect or interactions with version ($p_s > .10$). Next, I analyzed mean maximum capacity in a two-way ANOVA with age (5 or 7 years) and version (modified or standard) as between-participants factors, which also revealed no main effect or interaction with version ($p_s > .39$). Thus, there is no evidence that the modifications described above inflate estimates of capacity.

**Discussion**

This study sought to answer two questions about the development of VWM capacity. First, it explored VWM capacity before 5 years of age by testing 3- and 4-year-olds in the modified version of the change detection task. Results from the modified version of the task suggested continuity over development, with change detection performance improving and capacity increasing from 2 to 3 items between 3 and 5 years to approximately 4 items by 7 years. Second, it addressed whether the change detection task used by Riggs et al. (2006) was too difficult for children by comparing the modified task with a replication of their method with 5- and 7-year-olds. The comparison showed that the modified version does not overestimate capacity even though it is easier for young children to perform.

Although there was no significant difference from the standard version of the task, estimates of children's capacity in the modified version of the task were higher than those reported by Riggs et al. (2006) and Cowan et al. (2005). The reasons for this difference are not entirely clear: even in the replication of Riggs and colleagues' design reported here, estimates were nearly 3 items for 5-year-olds and 4 items for 7-year-olds (compared with their estimates of only 1.52 and 2.89 items, respectively, reported by Riggs et al. and of 3.5 items reported for 7- and 8-year-olds by Cowan et al.). One possible source of the difference is that both Riggs and colleagues and Cowan and colleagues allowed for repetition of colors within the array, whereas the current design did not. These researchers have suggested that allowing for repetition of colors requires more specific encoding of items in memory (see also Vogel et al., 2001); however, pilot data for the current study suggested that young children found these repetitions to be confusing. Cowan and colleagues noted that participants performed slightly worse (~0.5 item lower $K$) when the new item on change trials duplicated a different item than had been present in the memory array. Moreover, such repetitions emphasize feature-space binding, which may be limited in young children (Simmering, Johnson, Patterson, & Spencer, 2009; see also Cowan, Naveh-Benjamin, Kilb, & Saults, 2006). Although this does not directly address whether color repetitions within the memory array affect performance, independent of the changing item, it is suggestive that these types of changes in methodology can affect capacity estimates. Further research is needed to directly test whether repeating colors within an array affects capacity estimates.

A second, more likely reason for the higher capacity estimates in the current study is differences in the method of estimating capacity. As noted above, Cowan (2001) proposed a formula modification for the version of the task in which an item is cued at test (as in Cowan et al., 2005); it differs from the original formula (Pashler, 1988) in that the dividing term ($1 - FA$), is eliminated. Because this term is always less than 1, the result from Pashler's formula will always be larger than the result from Cowan's formula. As such, similar levels of accuracy lead to different capacity estimates through different computations. Although this could explain why the 5-year-olds in the current study showed similar capacity to the 7- and 8-year-olds in Cowan et al. (2005) study, it does not address the difference from Riggs et al. (2006). In their study, no item was cued at test, thereby precluding the use of Cowan's formula; rather, they used Pashler's formula, as was the case here. However, the section on capacity in their article (p. 23) does not describe how they arrived at a single estimate for each age group, that is, whether they averaged across set sizes or took the maximum for each participant. Estimation from the data presented in Riggs and colleagues' Fig. 1 and the description in the section on capacity suggest that calculations were averages across set sizes. Averaging across set sizes for the data presented here resulted in mean capacity estimates of 1.82 items for 5-year-olds and 2.31 items for 7-year-olds, which are much closer to those from Riggs and colleagues, further supporting this interpretation. Furthermore, 5- and 7-year-olds' hit rates in Riggs and colleagues' Fig. 1 are comparable to those in the
current study (see Fig. 2); however, they did not report children’s correct rejection rates, so it is not known whether those trials were also similar across studies.

Even with slightly higher estimates in the current task, 3- and 4-year-olds’ capacity as measured by change detection still falls short of the adult-like 3 or 4 items that infants showed in the change preference task (Ross-Sheehy et al., 2003). The modified version of the change detection task presented here allowed younger children to complete a task that had previously been too challenging; however, the capacity estimates it yielded do not reach 3 or 4 items until between 5 and 7 years of age, much later than the 10 months suggested by Ross-Sheehy and colleagues (or possibly 7.5 months, as in Oakes et al., 2006). Thus, it seems that the difficulty with the adult change detection task is not the explanation of the conflicting findings by Ross-Sheehy and colleagues and Riggs et al. (2006). Recall that Riggs and colleagues suggested three possible explanations for the differences across tasks. Their first suggestion, that infants are not remembering all of the items, has been ruled out by further infant experiments (Oakes et al., 2009).

The second possible explanation proposed by Riggs et al. (2006) was that the looking task is tapping a different (more passive) memory system than the change detection task. This possibility is not addressed by the data presented here. It is unclear whether Riggs and colleagues consider only the response type (looking vs. verbal) to differentiate these potential memory systems or whether the manner in which stimuli are presented leads to different types of encoding. Further studies would be needed to identify whether VWM can be subdivided in such a way, and how different behavioral tasks may depend on these subsystems.

The third explanation proposed by Riggs et al. (2006) was that capacity could reach adult-like levels during infancy, but developmental improvements in other cognitive processes were driving behind change detection performance. As described in the Introduction, Perone and colleagues (2011) suggested that performance in the infant change preference task is determined by more than just capacity. Rather, the stability of memory representations and the interaction of memory and fixation within the change preference task influence infants’ behavior. Thus, developmental changes in performance in both tasks could be explained by cognitive processes other than just VWM capacity. This raises an important challenge for VWM theorists to put forth explanations of performance that specify how various cognitive processes come together to produce the behavior measured in each task. Only by reaching this level of specificity can we begin to understand the relationship between different tasks used at different points in development.

Although the data presented here cannot fully test this hypothesis, I propose a related (but distinct) explanation of the difference across tasks: Both tasks may rely on the same underlying VWM system, but the different task structures and behavioral measures lead to legitimately different estimates of capacity. This explanation builds on perspectives such as dynamic systems and ecological theories that emphasize the importance of considering behavior in context (e.g., Plumert, 2008; Thelen, 2000). Numerous examples from other domains—including Piaget’s A-not-B task (e.g., Zelazo, Reznick, & Spinazzola, 1998), conservation of number (e.g., Mehler & Bever, 1967), and theory of mind (e.g., Siegal & Beattie, 1991)—have shown that changing the task context allows children at younger ages to show developmental progress not apparent otherwise. An important implication of this approach is that neither the infant looking task nor the change detection task provides a “pure” measure of capacity; rather, capacity limits arise in the context of the behavior they serve (see Simmering & Perone, 2011, for a related discussion).

The study presented here indicates a need for continued exploration into the development of VWM capacity and understanding how capacity estimates are influenced by the tasks designed to test them. Further studies are needed to compare performance directly across these tasks to determine the source of the developmental inconsistency from infancy to early childhood. Combining evidence from the change preference task and the change detection task suggests a relatively protracted developmental trajectory, with capacity increasing across many years during childhood. Only by understanding the entire range of cognitive processes that contribute to performance in these laboratory tasks can we begin to find the sources of developmental changes in these processes—including and in addition to VWM capacity—as well as their influence on how the human visual system functions in real-life activities.
Acknowledgments

A subset of the data presented here was part of the author's dissertation, which was funded by a fellowship from the University of Iowa Graduate College. Thanks are owed to the families and students who participated in this research. Thanks also go to Jeffrey S. Johnson and Rob Olson for programming assistance, to Rebecca Patterson, Valerie Vorderstrasse, and Chelsey M. (Patterson) Wood for assistance with data collection, to John P. Spencer for input and guidance during the conceptualization of the project, and to Jessica Horst and Sammy Perone for reading earlier versions of the manuscript.

References