Using Computational Analysis of Behavior To Discover Developmental Change In Memory-Guided Attention Mechanisms In Childhood

Amso, D.[1]*, Govindarajan, L.N. , Gupta, P., Placido, D., Baumgartner, H. , Lynn, A. , Gunther, K. , Sharma, T. , Veerabadran, V. , Thakkar, K., Kim, S., and Serre*[2]*, T.*

[1]Department of Psychology, Columbia University

[2]Department of Cognitive, Linguistic, and Psychological Sciences, Robert J. and Nancy D. Carney Institute for Brain Science, Brown University

*Correspondence: da2959@columbia.edu
thomas_serre@brown.edu

Keywords Automated behavioral analysis, spatial learning, memory, development, memory-guided attention, visual search, visual attention, top-down attention guidance .

STATEMENT OF RELEVANCE

A fundamental survival skill involves learning about space in a way that supports efficient attention and action in the real world. Studies of this skill, called memory-guided attention, offer conflicting evidence about its developmental course. Efficiency in memory-guided attention reduces the need to repeatedly expend expensive learning resources in previously visited spaces. At the same time, visual search anew can be more efficient than engaging memory and visual attention processes at once. The most adaptive strategy seems to depend on task dynamics. Naturalistic scenes and energy-costly task demands are more likely to engage memory during visual search in previously visited spaces in adults. As such, understanding the development of memory-guided attention may best be done in real world contexts. Yet, methodological restrictions have limited a fully embodied understanding of this skill in developing humans. We built a SmartPlayroom, outfitted with an array of video and depth sensor technologies, and combined with computer vision algorithms for automatically characterizing oculomotor and locomotor behavior.

- Children 4-9.5-year-old participated in a naturalistic memory-guided search task, while gaze fixation distribution and body movements were automatically tracked and analyzed with modern computer vision algorithms.

- We manipulated object placement and trial order such that nearby objects would be incidentally encountered during an initial search for reference objects and used computational models of top-down guidance in visual search to show that children distributed fixations consistent with the visual features of remembered reference objects when later searching for nearby objects.

- Our analyses revealed a fundamental difference across age: In younger but not older children, engaging memory supported faster visual search, but this required active body navigation during initial spatial learning.

- Overall, our study highlights the benefits of naturalistic experimental paradigms combined with modern computer methods for understanding mechanisms underlying the key operation of memory-guided attention.

Abstract

We tested 4-9.5-year-old children on a naturalistic memory-guided attention visual search task. We measured fixation distribution during a search using wearable eye tracking, and simultaneously recorded depth video data for each participant and used computer vision algorithms to track them during navigation. We manipulated object placement and trial order such that nearby objects would be encountered during initial search for reference objects. We used a computational model of top-down guidance for reference object visual features and examined the use of this top-down attention for reference objects during *subsequent* nearby object search. The data suggest that the value of physical navigation during initial spatial exploration for subsequent memory-guided attention, specifically in early childhood, is in its association with stronger visual representations of goal reference objects during spatial exploration. By middle childhood, visual search times were not impacted by memory engagement. 139 words

Introduction

Consider a child walking into her home with Grandmother. It is her birthday and she is not aware that her parents have planned a surprise party in the dining room. Twenty people stand and sit quietly around the dinner table and suddenly erupt into unexpected shouts. The immediate response to such a scene is confusion and uncertainty. This is quickly quelled by walking around the table to where Mom usually sits while scanning for her curly dark long hair, her physical presence at the expected location reliably reducing mounting anxiety. Children are asked to complete tasks like this every day. This skill is called memory-guided attention.

Objects generally do not occur in isolation but rather are contextually associated with other objects and locations in scenes **(**Oliva and Torralba, 2007; Võ, Boettcher and Draschkow, 2019). Brockmole and Henderson (2006) argue that memory-guided attention requires both implicit recognition of a scene or space, and also attention guidance processes. Visual search for objects in scenes has multiple influences, only some of which engage memory. These are bottom-up salience, top-down guidance, scene grammar, and retrieval of learned knowledge of the spatial structure of an environment (Brockmole and Henderson, 2006; Chun and Jiang, 1998, 1999; Goujon and Fagot, 2013; Wolfe and Horowitz, 2017; Wasserman, Teng, and Brooks, 2014). This work is concerned with the development of this hybrid attention and memory process in 4-9.5-year-old children, and of the conditions under which children are likely to engage it.

Chun and Jiang (1998) used spatial contextual cueing tasks to test memory-guided attention in adults. In these tasks, participants are asked to search for a target among arrays of distractors. The context is the spatial arrangement of targets and distractors. Relative target/distractor spatial arrangements in displays can either be random or repeated. Participants learned the relative spatial arrangement of the targets and distractors over repeated trials, as indicated by faster object detection reaction times on repeated relative to random visual search trials. In this way, learned spatial context cues target location and guides visual search. This contextual cueing effect obtains when only two items surrounding the target are repeated (see also Mack and Eckstein, 2011), as well as when all items in the display are repeated (Brady and Chun, 2007). This suggests that contextualizing an object in a scene can be done at both local and global spatial scales (Bar, 2004;

Oliva and Torrallba, 2007). Here we focus on whether children use memory-guided attention to search for incidentally encountered *local object co-occurrences*.

Previous studies using contextual cueing in middle childhood have produced mixed results (Couperus, Hunt, Nelson, and Thomas, 2011; Dixon, Zelazo, and De Rosa, 2010; Nussenbaum, Scerif, and Nobre, 2018; Vaidya, Huger, Howard, and Howard 2007; Yoshida, Darby, and Burling, 2011). Studies have found that children cannot reliably learn the spatial context for memory-guided attention (Vaidya et al., 2007), and others that children perform memory-guided attention tasks similarly to adults (Dixon et al., 2010), and still other studies have shown that children have better memory-guided attention than adults (Nussenbaum et al., 2018). These discrepancies across studies likely derive from differences in task demands. For example, both task length (Yoshida et al., 2011) and the number and nature of the distractors in a scene (Couperus et al., 2011; Nussenbaum et al., 2018) have been shown to influence memory-guided attention performance.

It is also unclear how the precise interaction of attention and memory development unfolds in childhood. Shimi, Nobre, Astle, and Scerif (2014) examined developmental differences, in children (7 and 11 year-olds) and adults, in using controlled visuospatial orienting mechanisms for the maintenance of items in visual short term memory. They found that poorer attention processes in younger children resulted in *less* benefit on visual short-term memory performance than found in older groups. At the same time, Nussenbaum et al., (2018) found that memory-guided attention was *better* in younger than older children and adults. These data points are not in conflict. Rather, taken together they suggest that poorer attention processes in young children may benefit from additional support offered by memory for spatial context during visual search.

The current state of understanding of this key skill limits opportunities to improve spatial memory disruptions in a host of neurodevelopmental disorders at a critical period in human development. Affected children are those that have, for example, 22q11.2 Deletion syndrome (Bearden et al., 2001), attention deficit hyperactivity disorder (Bedard, Martinussen, Ickowicz, and Tannock, 2004), Down syndrome (Clark, Fernandez, Sakhon, Spano, and Edgin, 2017), and Williams syndrome (Brown, Johnson, Paterson, Gilmore, Longhi, and Karmiloff-Smith, 2003). There are two critical issues highlighted by a rich adult literature that may shed additional light on the mixed findings in the developmental literature, and that we use to guide our empirical design.

First, Wolfe and Horowitz (2008) argued that there are conditions where *de novo* search for an object may be faster than memory retrieval followed by attention guidance (see also Wolfe and Horowitz, 2017). Thus, while experiments may manipulate scene context in a way that *allows* for memory-guided attention, as in the contextual cueing task, not engaging memory resources when visual attention alone is sufficient may be a more efficient strategy for finding target objects. Task demands may determine whether engaging memory processes during visual search is less efficient than engaging in *de novo* visual search on each trial (Kunar, Flusberg, and Wolfe, 2008; Vo and Wolfe, 2012; Vo and Wolfe, 2015). Specifically, multiple studies point to naturalistic visual search as more likely to benefit from memory engagement. Visual search that requires complex eye movements (Brockmole and Henderson 2006) or that requires some degree of navigation in virtual environments (Hollingworth and Henderson, 1998) has been shown to benefit from memory for repeated items in context (Solman and Kingstone, 2014). Naturalistic search involves head and body movements, egocentric and allocentric reference frames (Jiang, Won, and Mussack, 2014), and allows for proprioceptive and vestibular feedback shown to be important for spatial learning and memory (Chrastil and Warren, 2012; Li, Aivar, Kit, Tong, and Hayhoe, 2016). A recent study using virtual reality found superior memory for actively searched vs. memorized objects, as well as an increased use of memory when physically searching for multiple objects in a room in 3D (Helbing, Draschkow, and Vo, 2020). Importantly, there is an energy cost to moving head and body, and to the coordination of head, body, and eye movements (Ballard, Hayhoe, and Peltz, 1995; Hayhoe, Bensinger, and Ballard, 1998; Hardiess, Gillner, and Mallet, 2008; Solman and Kingston, 2015; Li Aivar, Tong, and Hayhoe, 2018). One possibility is that it is desirable to reduce this energy cost. Foulsham, Chapman, Nasiopoulos, and Kingstone (2014) found that top-down instructions in active visual search reduced the number of head and body movements made by participants, and resulted in faster target fixation. More generally, top-down instruction has been shown to support visual search. Chen and Zelinzky (2006) found that giving participants a preview of the target (top-down) speeded target detection and reduced attention to a color singleton distractor during visual search. In this way, memory for target object spatial location and local object co-occurrences may similarly be a means of exerting top-down guidance in complex search tasks.

The second related issue highlighted by adult studies is relevant to how to operationalize memory during visual attention guidance. Oliva and colleagues (2003) suggest mechanisms by which

spatial co-occurrence of objects (the ball is next to the pillow) may guide subsequent visual attention during search for an object (see also Torralba, Oliva, Castelhano, and Henderson, 2006). Top-down models of attention involve fixation distribution for scene features that approximate those of a defined target object. Retrieving information about local object co-occurrences to guide visual search should theoretically engage visual representations of the target object and the previously incidentally encountered neighbor. Top-down information from visual context has been found to modulate the saliency of image regions for fixation during object detection (Oliva et al., 2003). Hwang, Higgins, and Pomplun (2009) showed that top-down guidance computational vision models are an informative measure for understanding how saccades are distributed during visual search. Here we capitalize on this approach to ask whether top-down visual attention models can inform whether children engage visual memory during target visual search. For example, having previously encountered the ball next to the pillow might mean that a child distributes fixations to image regions consistent *with the ball* when later searching for the pillow.

In sum, the challenges we addressed in this work were (1) to design naturalistic task contexts in which it might be fruitful to use memory for local object co-occurrences to guide attention rather than to simply search *de novo*, and (2) to measure eye movements during active behavior in order to derive top-down guidance models. Here we use a naturalistic visual search task using an automated behavioral data collection space, which we call the SmartPlayroom (Figure 1). The SmartPlayroom is equipped with mobile eye tracking and depth and video sensor technologies, allowing precision in oculomotor and locomotor measurement. We adapted a visual search task procedure developed by Li et al. (2016) for both trial ordering and strategic object co-occurrence manipulation. Anchor objects have been shown to be valuable in supporting visual search for likely placement of other objects (Boettcher, Draschkow, Dienhart, and Võ, 2018). We ordered search trials such that the first six trials were designed to allow an opportunity for incidental learning about local object co-occurrence relations while children searched for and retrieved specific anchor objects, which we call Reference objects. In the following interleaved trials, children searched for and retrieved target objects that we had strategically placed either immediately Near these Reference objects, or relatively Far in the room (Figure 2). Trials started with the experimenter showing children a picture of the object they were to search for and retrieve in the room. Children were given no instructions on how to search the space. They could actively navigate the space in search of target objects, engaging head and body movements, or they could stand and scan until

the target was located and then retrieve it. In this way, we are able to examine the relationship between spontaneous head and body movement on visual attention during initial learning and subsequent memory-guided attention across early to middle childhood.

To summarize, we define memory-guided attention in the SmartPlayroom as the use of previous experience to guide current visual attention, and we operationalized it with computer vision models to provide measures of top-down attention guidance (Hwang et al., 2009; Peters, Iyer, Itti, and Koch, 2005; Zelinsky, Adeli, Peng, and Samaras, 2013; Zhang, Tong, Marks, Shan, and Cottrell, 2008). As in Hwang et al., (2009), top-down attention guidance values reflect the extent to which children distribute fixations, during visual search for the Near objects, in a manner that is consistent with the known visual features of the previously encountered paired Reference object. High values of top-down guidance on paired Near relative to Far object trials thus reflect the use of previous experience in the room to guide current visual attention. To best understand the developmental trajectory of memory-guided attention, we used our novel SmartPlayroom methods to explore the impact of top-down guidance values and physical movement strategies on initial learning Reference trials, and their impact on subsequent visual search reaction times.

Method

Participants. This study was conducted as part of a grant aimed at developing the SmartPlayroom behavioral data collection technology. Because of the novel multi-method exploratory nature of the work, there was not a clear effect size from which to calculate *a priori* power analyses. We used Li et al. (2016), from which the task was directly adapted, to determine the number of participants to include (*N*=42). A total of *N* = 52 children 4-9.5 years of age (*M* = 5.97, *SD* = 1.42, range = 4.01- 9.85 years) were tested in two separate sessions on visually similar versions of the task**,** one naturalistic in the SmartPlayroom and one screen-based on separate days and in counterbalanced session order. The screen-based version was included as part of our SmartPlayroom methods development plans. The tasks are not comparable in demands and the results of the screen-based tasks are not reported here. Of these, 16 participants completed both sessions but were excluded from final analyses for the following reasons: *N*=2 4-year-old children did not understand the task, *N*=2 4-year-old children refused to wear the portable eye tracking glasses, and *N*=12 5-9-year-olds for poor quality eye tracking data that could not be processed for reaction time or used in our computational algorithms. The final *N* = 36 children had *M* age = 6.21 years, *SD* = 1.47 (range 4.09 - 9.70 years). Parents reported that children were: 23 White Non-Hispanic, 8 White Hispanic, 2 White and Asian, 2 Black and White Non-Hispanic, and 1 Black Hispanic child. Family income-to-needs was *M* = 4.23, *SD* = 3.60. Average parent education is *M* = 16 years, *SD* = 2.5 years. Participants were recruited through birth records, from existing databases, or through community advertisements. All parents signed consent forms, children provided assent, and families were compensated for time and travel in accordance with University IRB-approved protocol.

Materials

Toy Objects. Reference objects were 6 3-dimensional geometric objects, each defined by a unique shape and color but similar in texture (orange rectangle, yellow triangular prism, blue cube, green cone, red sphere, pink cylinder). Near/Far objects were 12 familiar toys that were each roughly 3.5 inches in size. An additional 15 toys were included in the room as foils (i.e., 4 stuffed animals, 1 children's reading book, 1 soccer ball, 1 football, 1 small basketball, 3 bath toys, 1 stacking rings toy, 1 large lego, 1 wooden car, 1 toy ring chain).

Eye-Tracking Devices. The wearable SmartPlayroom eye tracker was a Positive Science Headgear Model DB9-CHG that included a scene camera view and an eye camera that utilized infrared LED to track the pupil and corneal reflection. Children wore a lightweight backpack to record data on a 2015 11-inch MacBook Air installed with 1.6 GHz Intel Core i5 processor running the Positive Science software Version 1.8.6.1.

SmartPlayroom Cameras, Recording, and Synchronization. The SmartPlayroom was fashioned with 6 Firefly MV 0.3 MP Color FireWire 1394a (Micron MT9V022) and 4 Xbox One Kinect 2.0 cameras, mounted on walls/ceilings to record behavior (see Figure 1). All computers used for data recording were time-synchronized using the standard Network Time Protocol (NTPv3 defined in RFC 1305) in the broadcast mode. In theory, such a protocol allows for the synchronization of all computer clocks within a few milliseconds of Coordinated Universal Time. Computers passively listened to time updates after an initial round-trip calibration exchange (conducted at the beginning of each session). These time updates were then used to timestamp the recorded data which were then synced during post-processing. Computers for video recording ran the Ubuntu 14 operating system with an open-source NTP client program (available at www.ntp.org). The computer used to record eye-tracking data used Mac OS X built-in NTP client. Depth sensor data were recorded on Windows PCs running the open-source NetTime (available at http://www.timesynctool.com).

Task Procedures

All children participated in both the SmartPlayroom naturalistic visual search and also a computerized testing session. Testing setting order was counterbalanced across participants such that about half participated in the computerized task first. The exact combination of Reference, Near, and Far toy object locations and pairings was counterbalanced across all children and was not repeated within children across testing settings. However, given the very different demands of settings, no direct comparison is made here. We additionally tested children on age-appropriate NIH Toolbox Flanker Task as a means of accounting for individual differences in visual distractor suppression abilities that were independent of our manipulation. The NIH Toolbox Flanker is a standard index of executive attention and the ability to suppress distraction (Weintraub et al, 2013; Zelazo et al., 2013, see SI Appendix B for details).

Memory-guided attention visual search task. During the first 6 Reference trials, the child was asked to find the colorful geometric shapes that are the Reference objects. The remaining 12 experimental trials included counterbalanced searches for targets that were placed 8 inches (20 cm) from the Reference objects (Near trials) and searches for targets that were placed 36 inches (91 cm) from the Reference objects (Far trials). Toy objects were counterbalanced across Near and Far conditions. The average distance from the starting point was equidistant across conditions. The Reference objects were shuffled within the first six trials while the Near and Far objects were presented randomly for the remainder of the trials (7-18).

Prior to beginning, children were fitted with the portable Positive Science eye tracker. After a short calibration session, children were then taken to the testing room. Children were instructed that they would be asked to find a toy object. They began each trial at the same starting circle (Figure 1). The experimenter then turned over a placard with the image of the toy they were to find on that trial. The placard was shown until the child began their search or for a maximum of 5 seconds. The placard was then turned over so that the child could no longer see it. The child was given 30 seconds to find the object. If the child retrieved the target within the 30 second time frame, the trial was marked as correct. If 30 seconds elapsed and the child had not found the object, or the child retrieved the incorrect object during the 30 second window, the experimenter said, “Let’s try again. Can you find this?” while showing the placard again. This trial would be marked incorrect. If the child still could not find the object and/or the additional 30 seconds had elapsed, the experimenter said, “Ok, let’s try another one. Can you come back to the start mat?” After the child retrieved the object and returned it to the experimenter, the object was replaced in its original location by a second experimenter that came in briefly from the periphery (outside the search space), and the next trial began. Trials wherein the participant visually inspects the scene thoroughly (while staying immobile) before heading straight toward the target object were identified manually.

SmartPlayroom Data Processing and Analysis

The processing of eye-tracking data was fully automated using computer vision methods. This included the development of two separate computer vision systems. A computer vision system was trained to automatically detect the placards used to indicate the toy to be searched at the beginning

of each trial. This was used to automatically segment experiments into individual trials and to associate individual trials to the toys being searched. Another computer vision system was trained to automatically detect any fixation on any of the toys being searched in the course of the experiment.

Computer vision algorithms. For both placard and object detection, we used the faster R-CNN (Ren, He, Girshick, and Sun, 2015) architecture previously shown to yield state-of-the-art accuracy for the detection of natural object categories (comparable results were also obtained with YOLO9000 (Redmon and Farhadi, 2017). We used in-house annotation software to gather ground-truth bounding boxes for each of the 18 placards and corresponding 18 objects in order to create an image dataset for training and testing the computer vision algorithms. The object dataset was created by manually labeling about 6K frames sampled from head-mounted eye tracker videos from randomly selected participants. Similarly, a placard dataset was created by labeling 3K frames from head-mounted eye tracker videos from 15 randomly selected participants. The number of available samples for training was increased using standard "image augmentation" methods such as flipping and blurring images, cropping sub-images at random and applying small affine transformations. Our final training dataset consisted of about 17K labeled frames for toys and 11K frames for placards. In addition to the toy categories, we also included an additional *background* class for non-target objects in the room, and both the placard and toy recognition systems were trained on a 19-way classification.

Our evaluation procedure for both the placard and toy detectors treated each object bounding box independently, and the accuracy was computed as the number of boxes correctly detected over the total number of boxes in the test dataset. An object was considered correctly detected if the predicted class label was correct and the predicted bounding box overlapped with the ground truth. In total, we had about 3K bounding boxes (from 14 videos) with placards and 5.5K bounding boxes (from 10 videos) for toys that were held out from training and used for evaluating the accuracy of the trained recognition algorithms. As an additional training step, we performed a round of bootstrapping which proceeded as follows: the trained detectors were re-applied to the training data frames, and predicted bounding boxes with zero overlap with any of the ground-truth boxes were identified as false alarms and were injected back into the training dataset as background exemplars. This extended dataset was then used for a second round of training known as "fine-

tuning". This procedure was instrumental in keeping the number of false positives in check. The standard Intersection-over-Union (IoU) measure was used as an indicator of the quality of the detections, and it is worthy to note that both our detectors achieved an IoU score of ~82% indicating significant overlap with the ground truth bounding boxes. Overall accuracy metrics for the placard and toy detectors were 97.2% and 97.9%, respectively (corresponding to the system's accuracy averaged across all 18 classes).

Automated parsing of experiments into trials. Videos from the head-mounted eye tracker were processed and placards were detected by the computer vision algorithms to automatically mark the start of each trial. The start of the subsequent trial was considered the end of the preceding trial. Additionally, an experimenter then inspected all trials manually and recorded the time at which the participant first grasped the target toy.

Automated annotation of eye fixations. Estimating the exact location of eye fixation in 3D space is inherently a challenging problem, given that the eye-tracking data reports 2D coordinates with respect to the eye camera. Thus, we defined a 3D "error window" (i.e., a region of uncertainty centered around each fixation). The window was constrained to shrink with the viewing distance to account for the fact that we are more confident that a large object at a short distance is more likely to be truly fixated compared to a small object at a further distance in the background which is more likely to fall close to fixation just by chance. To best measure the viewing distance for every toy (in the current field of view) from the subject, we employed a metric which was calculated as follows:

$Err_t = 10 \times \frac{\mathbb{B}_t(.)}{\mathbb{B}_t(1\mathrm{m})}$, where Err is the error window centered around the currently reported fixation for toy t. $\mathbb{B}_t(.)$ is the height of the current bounding box detection on toy t. $\mathbb{B}_t(1\mathrm{m})$ is the height of the bounding box detection on toy t when placed at a 1-m distance.

The scaling factor value of 10 (pixels per degree of visual angle), as well as bounding box detections at 1-m, were obtained in a dedicated calibration session. A toy t was considered to be fixated if its detected bounding box overlapped with the fixation zone centered on the eye tracker fixation readout, with width Err_t. In this way, this measure allows for multiple objects to be considered fixated simultaneously.

Eye-tracking metrics in the SmartPlayroom. We computed three eye-tracking metrics. For each visual search trial, we used machine annotations to compute the search response time (RT) as the time from trial initiation to the first fixation to the target object that preceded object retrieval. We also computed the number of fixations to a target object prior to the trial on which it was the object of the search (Table 1), as a measure of repeated prior sampling. Finally, we isolated the proportion of trials where a fixation was made to the Near object immediately before the Reference object was fixated and to Reference object immediately before the Near object was fixated. The latter measure served as a covariate for top-down guidance analyses as described in the results and SI.

Computational Models of Attention

Top-down attention guidance measures. As done in previous work (Hwang et al., 2009; Peters et al., 2005; Zelinsky et al., 2013; Zhang et al., 2008), we built a model of top-down guidance by considering the output of object classifiers (here the toy detectors developed for the automated annotation of fixations). Guidance estimates were based on target probabilities obtained from these detector outputs. To a first approximation, these detector outputs provide a measure of similarity between individual toys and image locations. These probability outputs o were linearized using the following formula: $y = \max(0, \log(o) + C)/C$, where $C = 15$ for our analysis.

For each frame and toy detector, we considered all detections above a threshold of 3.05×10^{-7}. This led to about 40-80 bounding boxes per frame. For any pixel, we considered all the potential bounding boxes that this pixel belonged to and assigned to the pixel the maximum score computed over all detections that overlapped with this location. When a pixel did not fall on any bounding box, the pixel was simply assigned a score of 0 for that particular toy. Computing these guidance scores across all pixels led to top-down attention guidance maps for each video frame and toy target (Figure 3). Predicted guidance values were extracted at fixations then averaged for each trial and participant. In this way, we predicted several top-down guidance scores for different trial types and target guidance predictions. This included guidance from a Reference object search for Reference search trials, Near search trials, and Far search trials.

Locomotor turn events and path trajectory data. We recorded depth video data for each participant using a Microsoft Kinect v2 and used computer vision algorithms to track participants' body joints using Microsoft Kinect Studio. Each recording was stored as a .xef file and participants'

trajectories were calculated from joints extracted using custom software to process the .xef files. We estimated the path followed by each participant to reach targets for individual trials by extracting the (x_t, y_t, z_t) coordinates of the body's center of mass to yield the trajectory $Z = \cup_{t=0\ldots T} (x_t, y_t, z_t)$.

We calculated "distance traveled" and "head and body turn" scores. To calculate distance traveled, paths were first discretized into a number of line segments, the endpoints of each marking a "turn" event. Turns were identified by thresholding angular deviation between successive time steps. The lengths of these line segments were then summed to yield the distance traveled values. Specifically, a deviation in path trajectory of 9 degrees or more was marked as a turn event. It is worthwhile to note that a turn event in general indicates a body turn. Although head turns occur during the course of a body turn, we do not count head-only as turn events.

Results

Analysis Plan. We used standard parametric analyses on averaged data by trial type (Reference, Near, Far). Table 1 shows descriptive statistics for key measures across participants and collapsed across trial types, as well as their correlations with age. Table 2 shows all correlations between measures used in subsequent analyses. We asked (1) how visual search RTs vary by trial type and age, and (2) whether Reference trial search strategy, through its impact on top-down guidance values, shaped subsequent RTs to Near - Far object search trials.

Visual Search Performance. Table 1 shows that overall accuracy, percent trials where children retrieved the correct object within the allotted time frame, was high and Table 2 shows this value was positively correlated with age. However, accuracy did not vary by trial type, $F(2,68) = .92$, $p = .40$, $\eta p2 = .03$ or by trial type and age, $F(2,68) = .94$, $p = .40$, $\eta p2 = .03$. As such, accuracy differences by age may reflect more general instruction following, for example, that are not likely to shed mechanistic light on memory-guided attention. Note that even when we divide age into younger ($N = 18$, 4.1 - 6.31-year-old children) and older ($N = 18$, 6.48 - 9.56-year-old children) groups, younger children have a high M accuracy = .93, $SD = .09$, as do older children $M = .97$, $SD = .05$. Indeed, accuracy is not the best measure of performance by design. When children retrieved the incorrect object or went beyond the allotted time, we marked the trial as incorrect but did ask them to try again until the correct Reference object was retrieved. Only 11 of all 648 trials

(36 participants x 18 trials each) were such that the object was never retrieved. We designed the task this way because object retrieval on Reference trials provided an opportunity for children to encounter Near/Reference spatial object relations in the SmartPlayroom. Accordingly, children had, on average, higher cumulative fixations (sum of all fixations to an object before it becomes the search target) on Near relative to Far objects, $t(35) = 5.80$, $p = .000$ (Table 1 for Means). Table 2 shows that these values were unrelated to RTs.

RTs for correct trials were measured as the time from trial initiation to the time of the fixation that immediately preceded walking toward the correct object to retrieve it. RT values within-subject were cleaned for outlier trials with extreme values (+/- 2 SDs from the participant's grand mean). Participants on average contributed eye tracking RT data for $M = 14.83$ of the 18 total trials ($SD = 2.25$ trials). This remainder trials reflected incorrect trials, eye tracker data loss (RT <100 msec or no data recorded), or outliers ($M = .97$ trial, $SD = .91$).

An ANCOVA comparing eye movement RTs to target object detection for each trial type (Reference, Near, Far) and age as a continuous variables showed only a significant interaction between trial type and age, $F(2,68) = 3.30$, $p < .05$, $\eta p2 = .09$. Helmert contrasts showed that the interaction was specific to performance on Reference relative to Near/Far trials, $F(1,34) = 5.04$, $p < .05$, $\eta p2 = .13$. There was no effect of Near relative to Far search trial RTs and age, $F(1,34) = .006$, $p = .94$, $\eta p2 = .000$. Visuospatial attention has been shown to show developmental change in this 4-9-year-old age-range (Amso and Scerif, 2015; Lynn, Festa, Heindel and Amso 2019). Consistent with this literature, these data indicate that on Reference visual search trials without the experience with our room (first 6 Reference trials), older children are faster to detect the target object than are younger children, $r(36) = -.34$, $p < .05$. However, these effects disappear for the Near and Far object search trials (all ps > .53), indicating that there may be benefit, to younger children, of learning and memory of the spatial arrangement of the objects in the SmartPlayroom (Figure 4). Another way to interpret the data in Figure 4 is that there is a great deal of variability in performance and no difference by trial type in the younger ($N = 18$, 4.1 - 6.31-year-old children) children, $F(2,34) = .36$, $p = .70$, $\eta p2 = .02$, but older children ($N = 18$, 6.48 – 9.56-year-old children) are *slower* for Near and Far trials after having completed the Reference trials, $F(2,34) = 3.69$, $p < .05$, $\eta p2 = .18$. Understanding individual differences in Near and Far trial RT, relative to Reference baseline search times by age, might shed light on these results.

Does visual memory for incidentally encountered Reference/Near object co-occurrences impact visual search? Here, we examined whether visual memory for the Reference object was associated with individual variability in eye movement RTs on Near relative to Reference and Far object search trials. See methods for computing top-down attention guidance for the Reference object values and also Figure 3. See also Supplemental Information (SI) Appendix A for control analyses.

Top-down attention guidance scores for Reference objects (TDG-Reference) *on Reference trials* are an index of how well children's specific fixation distributions mapped onto the visual features of the *specific* Reference object they had just been shown when searching for it. This measure corresponds to online goal oriented visual search on Reference trials. TDG-Reference values on Reference trials did not correlate with age (Table 2). Separately, TDG-Reference values on Near and Far trials reflect the extent to which children distributed fixations consistent with the features of the Reference object when *later* asked to search for the spatially Near/Far targets. This value corresponds to visual memory for the Reference object when searching for incidentally encountered objects on later trials. We computed TDG-Reference on Near object search trials and on Far object search trials. These values were not correlated with age but were correlated with each other (Table 2).

See Figure 2 for the specific Reference/Near pairings. A high TDG-Reference score on Near - Far search trials (TDG-Reference Near-Far) indicates a large incidence of fixations on locations in the SmartPlayroom that share visual features (i.e., shape, color) with the Reference object on *subsequent* searches for Near relative to Far placed objects. We always include this memory variable as a difference score between top-down guidance or fixation distribution consistent with Reference object features, on the paired Near relative to Far objects search trials. This eliminates the possibility, albeit unlikely, that we are indexing general fixation distribution similarity across trial types on the interleaved Near and Far object search trials. We can infer from this variable whether the spatial relations or local co-occurrences, among Near and Reference objects first encountered on Reference trials, shaped *subsequent* attention distribution, and thus served as a form of retrieved visual memory for the Reference/Near object co-occurrence in the room.

We used ANCOVA to examine RTs by trial type (Reference, Near, Far). We included age, TDG-Reference (Near-Far), and their interaction as continuous independent variables in the analysis.

We also incorporated into this analysis the proportion of trials per child where the Reference object was fixated *immediately* before target selection on Near trials. We describe in SI that this value may influence TDG-Reference on Near search trials, and account for this variance by including it in the analysis. Results showed only interactions of trial type and TDG-Reference (Near-Far), $F(2,62) = 5.75$, $p < .01$ ηp2 = .16, and trial type, TDG-Reference (Near-Far), and age, $F(2,62) = 4.57$, $p = .01$, ηp2 = .13. Helmert contrasts in the model showed that the effects of trial type and TDG-Reference (Near-Far) are significant for both Reference relative to Near/Far trials, $F(1,31) = 5.12$, $p < .05$, ηp2 = .14, and also for Near relative to Far trial RTs, $F(1,31) = 7.17$, $p = .01$, ηp2= .19. The same obtained for age, TDG-Reference (Near-Far), and trial type for the contrast of Reference relative to Near/Far trials, $F(1,31) = 4.65$, $p < .05$, ηp2 = .13, and for Near relative to Far trials, $F(1,31) = 4.40$, $p < .05$, ηp2 = .12.

Figure 5A shows that the *slopes* of the lines that describe the relationship between TDG-Reference (Near-Far), indicating memory use, and RTs by trial type differs by Age. TDG-Reference (Near-Far) is associated with faster RTs on Near/Far relative to Reference trials in younger children ($N = 18$, 4.1 - 6.31-year-old children), but slower RTs on Near/Far relative to Reference trials in older children ($N = 18$, 6.48 - 9.56-year-old children). The three-way interactions indicate that the relationship between memory-use (TDG-Reference Near-Far) and RTs differs by age. These data offer some evidence that faster RTs on Reference relative to Near/Far trials in older children (Figure 4) may reflect recruitment of memory for the Reference object.

Similarly, Figure 5B shows that, in younger but not older children, higher TDG-Reference on Near relative to Far object search trials, indicating use of the paired Reference trial visual features during visual search, was associated with relatively *faster* eye movement RTs to find Near relative to Far objects. In other words, visual memory for the previously encountered Reference-Near co-occurrences improved visual search for the Near relative to Far target objects only in younger children. These data offer some evidence that variability in younger children's performance in Figure 4 may also reflect recruitment of memory for the Reference object.

We thus asked whether other indices of learning and memory might also impact visual search performance. We found that neither the proportion of Reference trials where the Near object was fixated immediately before Reference object selection, $F(2,64) = .61$, $p = .55$, ηp2 = .02, nor total

incidental fixations on Near – Far objects before the object became the target of visual search, $F(2,64) = .72$, $p = .49$, $\eta p2 = .02$, explained RT performance by trial type and age.

In the introduction, we discussed the idea that certain task demands make it such that search is likely to benefit from memory engagement (Wolfe and Horowitz, 2017). Based on the literature reviewed in the introduction, and the variability in young children's data observed in Figure 4, we asked both whether higher TDG-Reference values on Near-Far search trials reflect different spatial search and learning strategies during Reference trial search in younger and older children, and whether these strategies might explain the relationship between TDG-Reference (Near-Far) and RTs in younger children.

Do initial search strategies on Reference trials support subsequent top-down guidance for Reference trials on paired Near target search? Recall that children are shown the target object on a placard before they are to find and retrieve it from the room (Figure 1). We observed that children spontaneously engaged one of two strategies. They either immediately began to navigate the space until the object was fixated and ultimately grasped for retrieval (we call this a Navigate-First strategy), or stood still and visually scanned the space until the target was fixated and then walked over to grasp and retrieve it (a Fixate-First strategy). Figure 6 illustrates example trials where children used these strategies. Most children spontaneously had both types of trials (26 children), with a smaller number exclusively using a single strategy (7 Fixate-First only, and 3 Navigate-First only). The proportion of Reference trials where children used a Navigate-First strategy was negatively correlated with age (Table 2). Younger children were more likely to use a Navigate-First strategy than older children. There was no relationship between the proportion of Navigate-First trials, age, and Flanker RT (see methods), $F(1,32) = .32$, $p = .57$, indicating that the choice of search strategy in the naturalistic space was unrelated to an independent index visual distractor suppression skill. Proportion of Navigate-First trials correlated with TDG-Reference on Reference trials (Table 2). Head and body turns were ostensibly higher on Navigate-First trials. Nonetheless, they are also possible on Fixate-First trials. We therefore calculated head and body turns separately per trial type. Head and body turns on Navigate-First trials (M=163., SD = 145.96) were higher than on Fixate-First trials (M = 108.65, SD = 66.14). Within the context of finding the Reference object in a cluttered space filled with distractors, this exploration strategy may bring changing distractors into view and thus also place greater demand on attention resources during search of

the Reference object. We tested this possibility by calculating the top-down attention guidance values on Reference object search trials where children used a Navigate-First strategy and correlated this with children's number of turns. In support of our prediction, we found making more physical turns while navigating in search of the Reference object was correlated with a higher top-down attention guidance values for that Reference object on Reference trials, $r(36) = .35$, $p < .05$.

Table 2 shows correlations that inform our final analyses. To reiterate, we previously showed that TDG-Reference (Near - Far) was associated with relatively faster RTs to Near objects in younger children (Figure 5, Table 2). The final analyses will address variables during initial search for the Reference object that support subsequent memory for the Reference object when searching for Near objects, here defined by higher TDG-Reference on Near relative to Far object search trials. Table 2 shows that TDG-Reference (Near-Far) is significantly correlated with a Navigate-First Reference trial visual search strategy and also the extent of initial target goal oriented attention, as measured by TDG-Reference on the first 6 Reference trials.

Thus, we next asked whether higher levels of goal-oriented attentional engagement, top-down guidance for the Reference object, on either Navigate-First or Fixate-First Reference trials (note these top-down attention guidance values are not correlated with each other, Table 2) are supporting *subsequent* use of top-down attention guidance for the Reference object on Near relative to Far object search trials. We conducted an ANCOVA with the dependent variable of top-down attention guidance on Near - Far object search trials. The predictors were age, the top-down attention guidance values on Reference object search trials where children used a Navigate-First strategy, top-down attention guidance values on Reference object search trials where the children used a Fixate-First strategy, and the interaction of each navigation variable with age. The analysis resulted in a main effect of top-down attention guidance values for the Reference object on Navigate-First Reference trials, $F(1,30) = 15.65$, $p = .000$, $\eta p2 = .34$, and this value interacted with age, $F(1,30) = 13.99$, $p = .001$, $\eta p2 = .32$. The analysis also resulted in a main effect of Fixate-First trials, $F(1,30) = 5.77$, $p < .05$, $\eta p2 = .16$, and this variable interacted with age, $F(1,30) = 8.22$, $p < .01$, $\eta p2 = .22$. The main effects indicate that higher top-down guidance values for the Reference object on Reference trials support higher values for top-down guidance for the Reference object on *subsequent* Near relative to Far object search trials. The interactions indicate

an age-related significant difference in the slope of the line that describes the relationship between the demand navigation places on goal-oriented search for Reference objects and subsequent top-down guidance for paired Near objects. The interactions are illustrated in Figure 7 by dividing the sample into two groups along the median age (N = 18, 4.1 - 6.31-year-old children, N = 18, 6.48 - 9.56-year-old children). Higher TDG-Reference on Navigate-First Reference trials is associated with subsequent higher TDG-Reference on Near relative to Far object trials, and the effect is larger in younger children (Figure 7). In contrast, higher TDG-Reference on Fixate-First Reference trials is associated with subsequent higher TDG-Reference on Near relative to Far object search trials. In this case, the effect is specific to older children (Figure 7). These data suggest that children use different strategies to search the space for goal Reference objects, both of which support target goal-oriented Reference search and that confer similar value for subsequent memory. However, for older children this does not translate into stronger visual search performance on Near relative to Far object search trials (Figure 5), and relative to the pattern shown in younger children, it slows visual search in comparison to their own baseline Reference RTs. In contrast, memory-guided attention improves visual search performance in younger children (Figure 5), with the variability explained by individual differences in Reference object goal orientation (TDG-Reference) during initial spatial exploration involving navigation.

Conclusions

Our data indicate that understanding the development of memory-guided attention is not about mapping a single trajectory of memory-guided attention across child development but rather the conditions and ages in which it is adaptive to engage memory processes during visual search. As noted by Wolfe and colleagues (Kunar, Flusberg and Wolfe, 2008), memory is only valuable under certain search conditions. In some cases, visual search *de novo* is more efficient than engagement of an additional cognitive process. Our results suggest there is (1) a developmental shift in reliance on physical navigation as a strategy for spatial exploration and (2) that visual attention or visual search benefits as a consequence of memory engagement subsequent to this exploration strategy in early childhood.

Like in Nussenbaum, Nobre, and Scerif (2018), younger children were more likely than older children to derive an attention orienting benefit from engaging memory, here measured by TDG-

Reference (Near-Far), both relative to their baseline visual search times on Reference trials and also in comparison to interleaved Far object search trials (Figure 5). There was not a significant TDG-Reference (Near-Far) by age correlation (Table 2). As such, TDG-Reference (Near-Far) values are not higher in younger children. Rather, the visual memory values across age are similar, but they are only supporting better visual search on Near relative to Far trials in younger children. Moreover, the slope of the line the describes the value of memory on baseline visual search differes by age, in that memory use has opposing effects on younger and older children, with some evidence that the trend involves slowing older children relative to their own baseline visual search RTs on Reference trials. Overall, these data might indicate that memory is scaffolding visual attention performance in *younger* children. Theoretically, in order for children to show an RT benefit for selecting Near relative to Far objects, they had to have encoded the local co-occurrence relations among Reference and Near objects *and* also the 6 Reference/Near objects' locations in the room. The spatial relation information is important for context (*I saw my wallet with my keys*) but should confer no value for search times if the location of the keys was also encoded. We examined how initial spatial exploration strategies impacted these results.

Children used different exploration strategies to search the space for Reference objects. Younger children were more likely to search by navigating the SmartPlayroom. Older children were more likely to stand and scan the space, find the target object, and then walk over to it and retrieve it (Table 2, Figure 5). These strategies were unrelated to an independent index of visual distractor suppression (Flanker task performance, Table 2). Figure 7 shows that top-down guidance computational values for the target Reference object during navigation supported better subsequent memory for the associated Near object in younger children. In contrast, top-down guidance computational values for the target Reference object during scan first (Fixate-first) strategy supported better subsequent memory for the associated Near object in older children. These data suggest that both spatial exploration strategies are effective for making and retrieving memory for Reference-Near object local co-occurrences. For younger children, Figure 5 shows that this memory supports faster visual search times on Near relative to Far object search. However, for older children this doesn't translate into stronger visual search performance on Near relative to Far object search trials (Figure 5). Thus, our data do not indicate an age-related difference in memory for the Reference-Near object co-occurrences. Rather, the developmental finding is about the conditions under which this memory is made, and whether its subsequent recruitment makes visual

search more efficient. In the broad picture, it seems this depends on both task conditions and the developmental state of visual attention orienting mechanisms in the service of visual search (Amso & Scerif, 2015)

Younger children also spontaneously navigated more than older children. It is not clear why this is. Within this search strategy, children with higher TDG-Reference values for the target Reference object were able to use the memory for the Reference object on associated Near trials (Figure 7) and improve their visual search times (Figure 5). Younger children who used a stand-and-scan (Fixate-first) Reference trial search strategy, did not derive this subsequent memory visual search benefit (Figures 7B and 4B). Thus, in early childhood, engaging memory supported faster visual search, but this crucially depended on how the initial memory was made. Specifically, the robustness of the visual representation of the goal object during initial Reference trial search, specifically when paired with active navigation, was critical to its subsequent use in improving visual search in young children. These data indicate that there are conditions that support spatial memory making in early childhood.

We can only speculate on older children's visual search performance. It is possible that older children are already performing optimally on the Reference trial visual search task and that the poorer performance on Near/Far relative to Reference trials reflects tedium with the task. The alternative is that it may be more efficient for them to search on each trial *de novo* (Wolfe and Horowitz, 2017). While this remains an open possibility, our current data do not yet clearly support this interpretation. Future studies can manipulate different levels of task difficulty and their impact on visual attention in middle childhood. It stands to reason that increasing visual search task distractors, room size, or object similarity might create more of a challenge for visual attention in older children and generate contexts in which memory is a boon for efficiency in visual search.

There is one alternative explanation for the older children's lack of reliance on memory to guide attention. This result is also consistent with developmental non-linearities in the benefit of contextual memory more broadly. Memory for items in context shows a great deal of developmental change in childhood (DeMaster, Pathman, and Ghetti, 2013; Edgin, Spanò, Kawa, and Nadel, 2014; Tummmeltshammer and Amso, 2017). Both human and animal data have shown that using spatial context improves subsequent memory in early childhood and again in

adolescence and adulthood, but not in middle childhood (Edgin et al., 2014). For example, Edgin et al (2014) had participants complete an object recognition task, indicating whether an object was old or new, after a learning session with objects embedded in scenes. They found that in young children (<4.5 years) and again in adolescence, presenting the item in the scene context improved subsequent recognition memory. However, in middle childhood, recognition memory was similar with or without the original scene context. In our data, older children had statistically similar values for TDG-Reference, our index of memory for the Reference object when later searching for nearby neighbors. This simply didn't improve search times for the Near relative to Far objects. Indeed, visual search performance relatively slowed on both Near and Far trials relative to Reference.

It is important to note that we did not directly manipulate how children searched on Reference trials, and as such these results can only be cautiously interpreted as association between exploratory patterns and later memory-guided attention efficacy. It is also unclear whether navigation during visual attention search for a target in a naturalistic space challenge top-down attention resources by bringing distractors into focus, and as a unintended consequence of focused attention enable deeper encoding about the local space, or whether head and body movements during navigation *narrow* the space of distraction allowing for better learning and memory for the target object and its surround. Future work will directly address this issue. The latter mechanism is consistent literature showing that active object handling supports stronger spatial representations in adults (Draschkow and Võ (2016). Visual search for an object while walking and making turns in a room requires focusing attention on changing path to avoid collisions or falls, for example. Previous work from infants shows that visual experience of the environment is altered by locomotor strategy (Kretch, Franchak, and Adolph, 2014). Data from elderly participants suggests that balance and fall avoidance are associated with focusing attention on the walking surface (de Melker Worms et al., 2017).

Our work would benefit from being placed in a larger literature on the value of navigation during development (see Newcombe, 2019 for full review and theoretical framework). Spatial cognition has long been linked to navigation in early childhood (Piaget and Inhelder, 1956), and developmental change in active exploration has been understood to be of value for attention and learning (Gibson, E.J., 1998). For example, emerging navigation, crawling and walking, has been linked with spatial learning in infancy (Clearfield, 2004). However, the exact mechanisms

supporting spatial learning during navigation early in development are multi-faceted, emerge at different times, and may be integrated together even later. These include, for example, physical skill development, perspective-taking development, egocentric and allocentric frame of reference, and landmark use for example (Newcombe, 2019). It is thus not surprising that young children and older children in our sample benefit differently from strategies used during initial spatial learning on Reference trials. It is most important that in either case, the association between initial behavior and subsequent memory-guided attention was via top-down attention guidance at both the encoding and retrieval stages of the task.

Table 2 shows that neither the cumulative number of fixations to an object before it becomes the target of search, nor whether the Near target was incidentally fixated immediately before the Reference object, a relational learning opportunity, were related to improvements in reaction time. They were also unrelated to TDG-Reference for the Near object. That is, neither having incidentally fixated the Near object before selection of the Reference nor having fixated it a large number of times over the course of other trials was associated with higher top-down guidance for the Near object on Near trials. These data indicate that these indices did not play a significant role in memory-guided attention in our task.

A limitation of this work is the small sample size and homogenous sample. Relevant to the latter, we cannot extrapolate or generalize our findings beyond the Western Educated Industrialized Rich and Democratic (WEIRD) sample of children tested. Relevant to the sample size, and given its exploratory nature and the copious number of measurements we were able to make, it was difficult to determine how to power such a study. Nonetheless, the work is grounded in theory, uses as a guide sample size from a published design (Li et al., 2016) and is sufficiently powered to extract meaningful results. We conclude by noting that spaces like the SmartPlayroom offer the quantitative precision required of strong science but without the limitations of deprived experimental testing environments (Bronfenbrenner, 1974). This study produced a copious amount of data from a variety of sensors and quantitative analysis would not have been possible without the automation offered by modern computer vision methods. While machine vision has rapidly become more prevalent for the analysis of animal behavior (Egnor and Branson, 2016), human studies have however been lagging behind. Here we show a benefit of this technology for understanding mechanisms underlying the key operation of memory-guided attention.

Acknowledgments

We would like to acknowledge Youssef Barhomi for contributing code for synchronizing all sensors and Gary Chien for contributing software for head tracking. We are grateful to Jeremy Wolfe and Melissa Le-Hoa Võ for their feedback on the manuscript. This work was funded by NIH (R21 MH 113870-01) to D.A. and T.S. Additional support to DA by NSF 1844476. Additional support to T.S. was provided by the Center for Computation and Visualization (CCV). We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program.

Author Contributions

DA, PG, HB, AL, KG, and TS designed the experiment. LNG, PG, KG, DP, TS, VV, KT, and SK collected and processed the data. DA, LNG, and TS analyzed the data and wrote the paper.

Declaration of Interests

T.S. serves as a scientific advisor for Vium, Inc.

Figure Legends

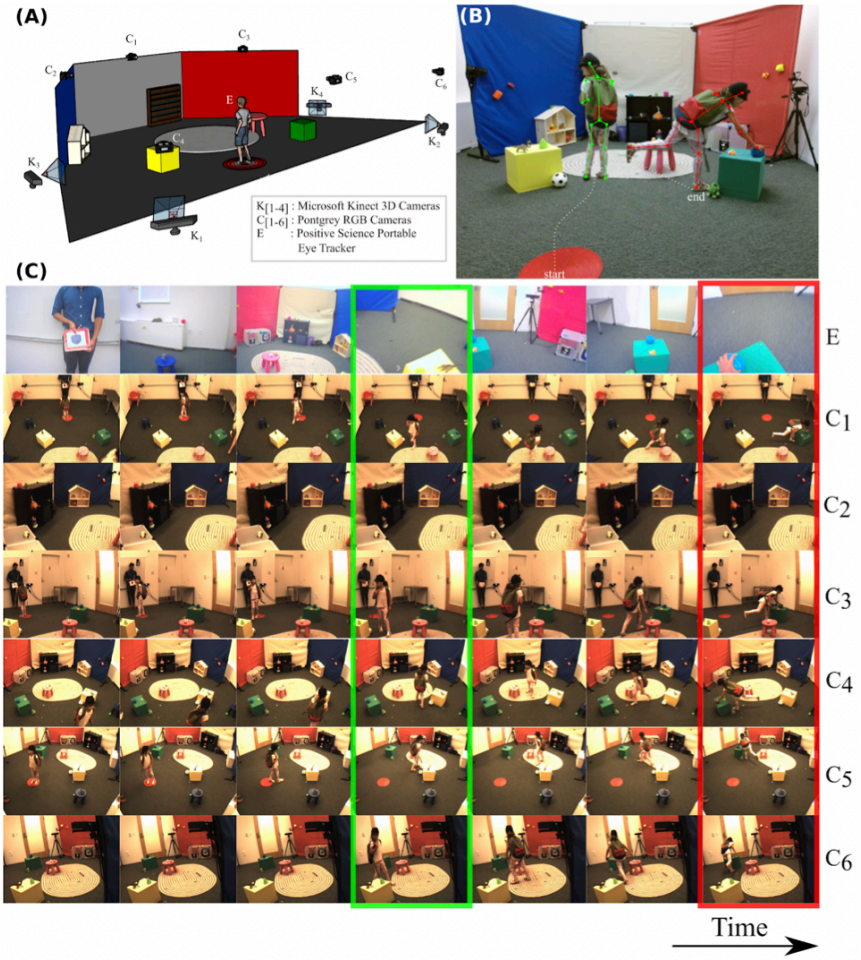

Figure 1: (A) Illustrates a 3D sketch of the SmartPlayroom. The participant wears a portable eye tracker (E). The room is equipped with 6 color cameras (C1-C6) and 4 Kinect depth sensors (K1-K4). (B) Depicts an example of a search trial. Children are shown a target object and are asked to search for and retrieve it for the experimenter. A second experimenter returns the object to its location after the trial is over. The start and endpoints are labeled accordingly. Body pose from the Kinect is shown for two search frames (in green and red). (C) Corresponding views from the eye tracker (E) and the RGB cameras (C1-C6) are highlighted in green and red. The time course of a single visual search trial from all RGB camera angles is shown from left (start) to right (end).

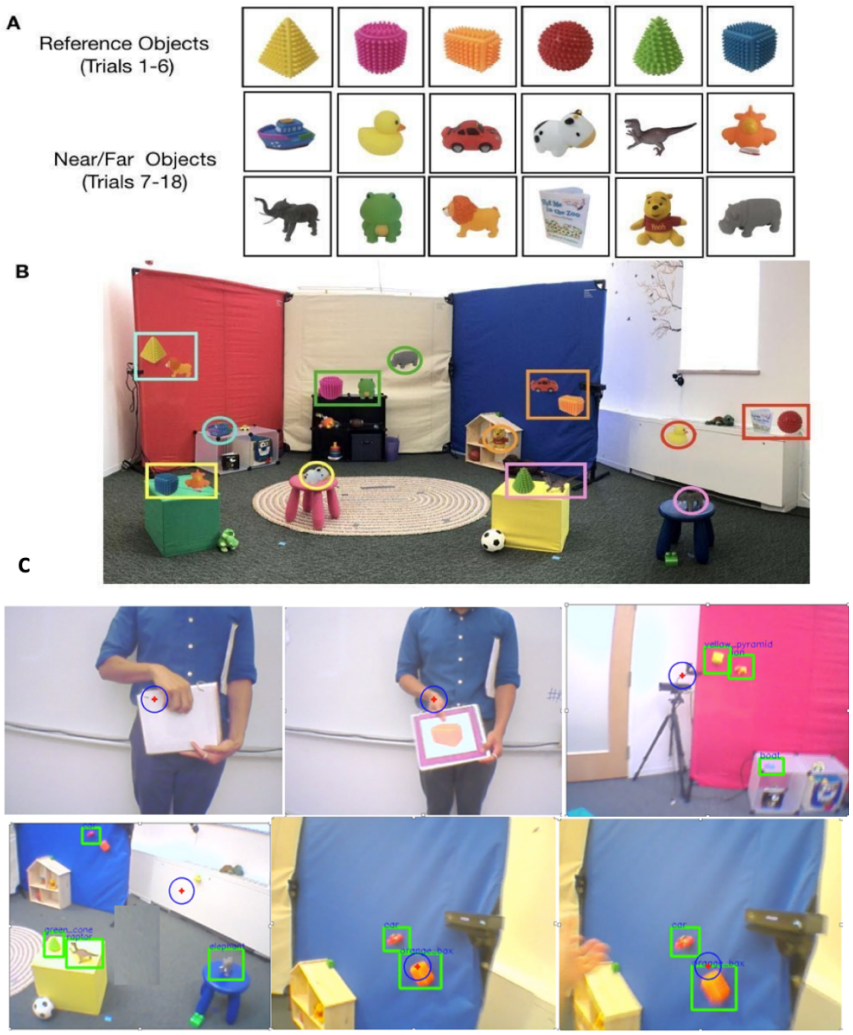

Figure 2. (A) Illustrates objects used on Reference search trials 1-6 (top panel). Reference objects were chosen to be distinctive with respect to color, shape, edges, etc. The remaining toy objects (middle and lower panels) are counterbalanced across Near and Far object search trials 7-18. **(B)** Depicts a sample object distribution in the SmartPlayroom. Rectangles are drawn around Reference/Near object pairs for illustration. Circles (same color) are drawn around the toys that serve as Far objects. Remaining objects are distractors or foils. Near objects are placed on the same surface and approximately 8 inches from the Reference. Far objects are placed approximately 36 inches from the Reference object. Objects are enlarged for viewer clarity and thus distances between objects are not as they were during the task.

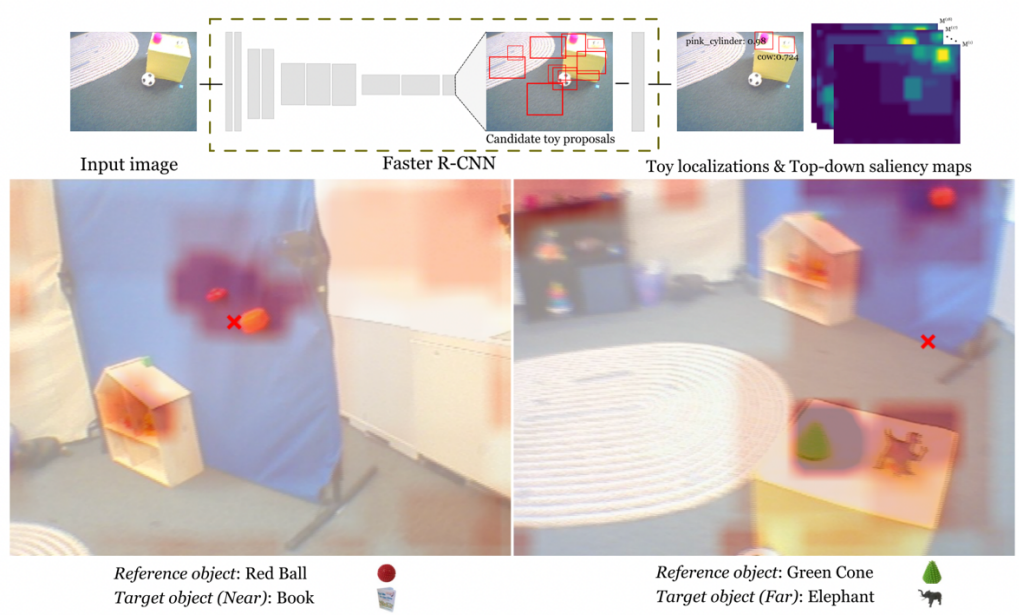

Figure 3. Overview of the method used to compute top-down guidance (TDG-Reference) measures. **Top row:** We built a computational model of top-down guidance for the Reference object by considering the output of the faster R-CNN trained to detect toys as a measure of similarity between individual toys and image locations. Guidance estimates were computed at image locations based on target probabilities obtained from these detector outputs and aggregated into top-down guidance maps for individual toys. **Bottom row:** TDG-Reference guides fixation on the near as opposed to far trials. First-person scene view overlayed with our top-down guidance map for the corresponding Reference object (TDS-R) for a sample Near (left) and Far (right) trial. The participant's eye fixation is marked with a red cross.

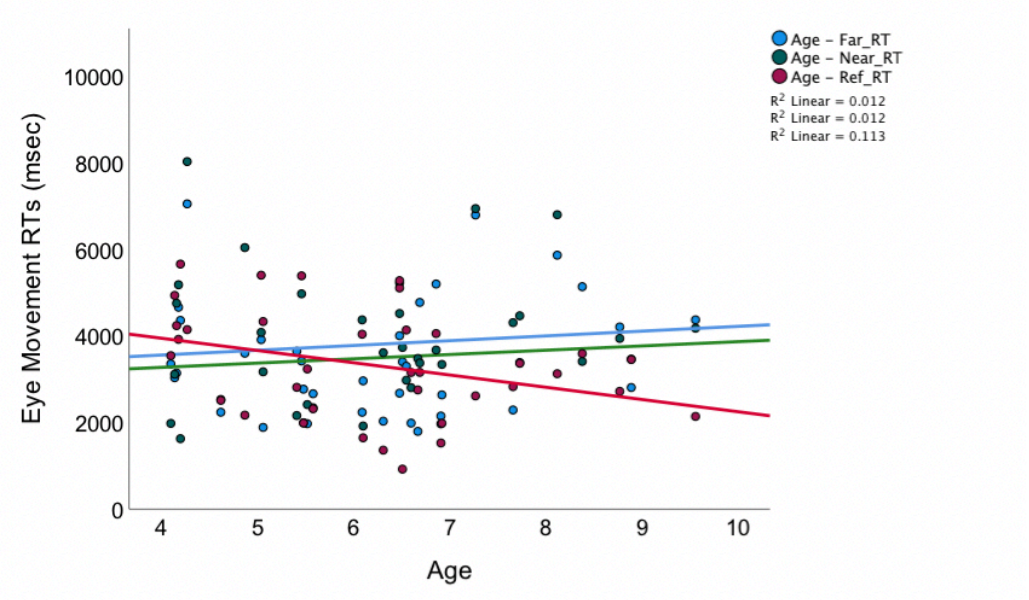

Figure 4. (A) Illustrates the relationship between Age and eye movement RTs to the target objects by trial type.

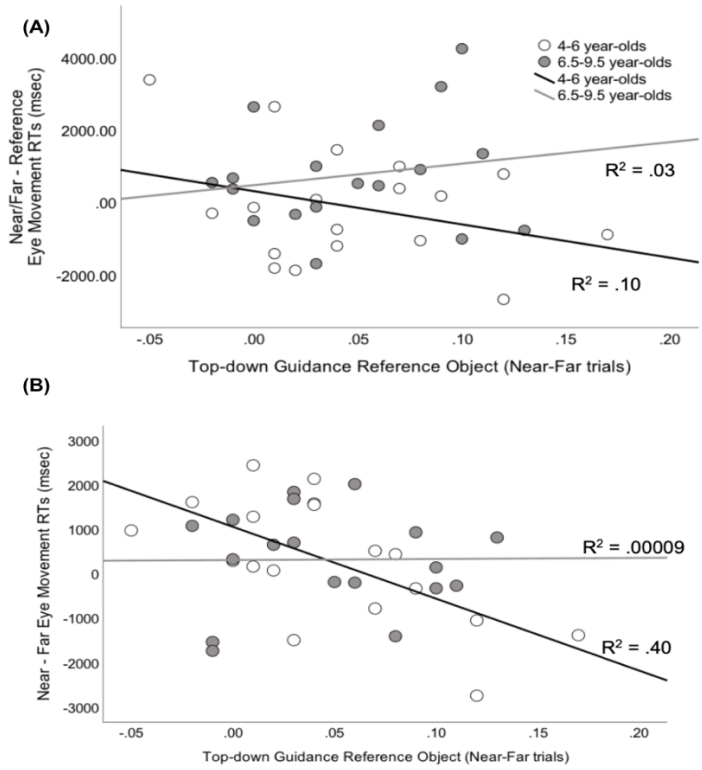

Figure 5. (A) Illustrates the relationship between TDG-Reference scores and eye movement RTs to the target objects in younger and **(B)** older children**.** The data are split along the median age for illustration only. Age is a continuous variable in all analyses.

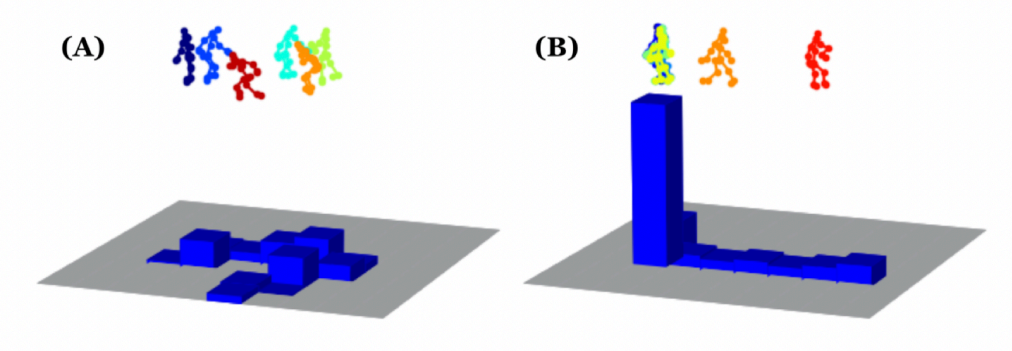

Figure 6. Participants use one of two strategies to locate the desired target object on Reference object visual search trials. **(A)** Representative trial for the Navigate-First strategy, wherein the participant navigates while visually inspecting the room for the target. **(B)** Representative trial for the Fixate-First strategy, wherein the participant is initially immobile while thoroughly visually inspecting the room, before proceeding to navigate to grasp the target. The top panels show the body skeletons obtained from the Kinect sensor for these examples, while the bottom panel shows a spatial occupancy grid for these two trials. Time is color-coded in the figure, from blue (start of the trial) to red (end of the trial).

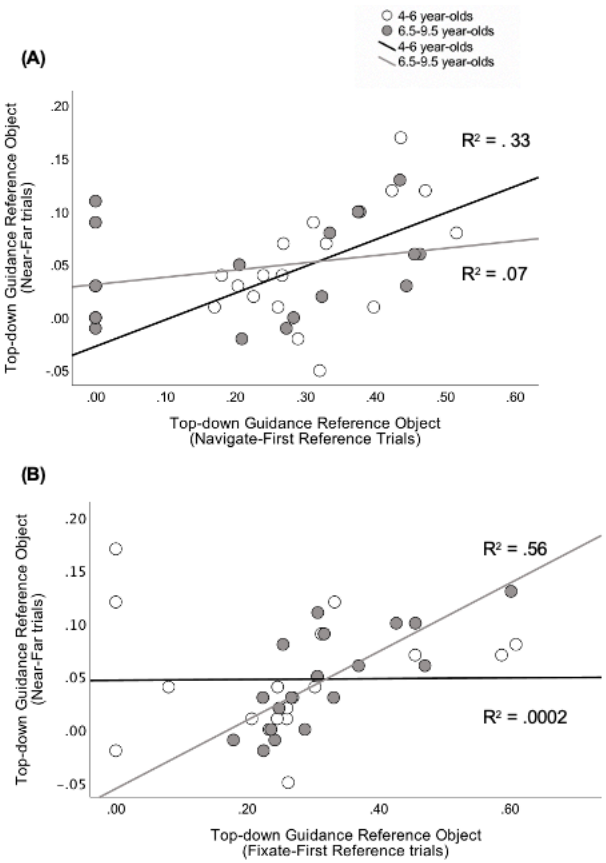

Figure 7. Depicts the impact of top-down attention guidance for the Reference object (TDG-Reference), during **(A)** Navigate-First Reference trials and **(B)** Fixate-First Reference trials, on subsequent engagement of TDG-Reference on Near relative to Far object search trials. Age is a continuous variable in all analyses but is split along the group median for illustration.

Tables

Table 1

Descriptive Statistics For Dependent Variables

	Mean	Standard Deviation	Range
Accuracy (percent correct trials)	.95	.07	(.72, 1.0)
Eye Movement RTs To Reference Targets	3324.99	1241.41	(923.60,5664.80)
Eye Movement RTs To Near Targets	3806.98	1510.11	(1628.33,8036.00)
Eye Movement RTs To Far Targets	3496.36	1339.61	(1797,7058.17)
Proportion of Navigate-First Trials	.44	.32	(0, 1)
Top-Down Guidance -Reference Targets/Trials	.31	.09	(.21,.56)
Top-Down Guidance -Reference Objects/Near Trials	.28	.07	(.19, .47)
Top-Down Guidance -Reference Objects/Far Trials	.24	.04	(.17, .32)
Average Prior Fixations to Near Target Objects	39.28	23.01	(4,137)
Average Prior Fixations to Far Target Objects	26.08	15.76	(2,83)
Proportion Trials Immediate Fixation Near/Reference	.34	.18	(0, .83)
Proportion Trials Immediate Fixation Reference/Near	.31	.20	(0, .83)
NIH Toolbox Flanker RT Score	1.79	.75	(.75, 3.53)

Table 2

Correlations between variables in analyses

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Age																		
2. Reference Trial RT	-.34*																	
3. Near Trial RT	0.11	0.19																
4. Far Trial RT	0.11	0.19	.62**															
5. Near - Far Trial RT	0.01	0.02	.54**	-0.32														
6. Proportion Navigate-First Reference Trials	-.36*	0.23	-0.22	-0.12	-0.14													
7. Number of Turns Navigate-First Reference Trials	-0.01	0.14	-0.01	-0.17	0.18	.40*												
8. TDG-Reference Object on Reference Trials	-0.01	0.04	-0.13	0.03	-0.18	.39*	-0.02											
9. TDG-Reference Object on Navigate-First Reference	-0.14	0.21	-0.05	0.00	-0.06	.61**	.35*	.70**										
10. TDG-Reference Object on Fixate-First Reference	0.27	-0.28	0.01	-0.06	0.08	-0.07	-0.07	.41*	0.25									
11. TDG-Reference Object on Near Search Trials	0.01	-0.10	-0.24	-0.06	-0.22	.41*	0.03	.80**	.58**	0.24								
12. TDG-Reference Object on Far Search Trials	-0.05	0.03	0.07	-0.01	0.10	0.18	0.15	.50**	.51**	0.08	.67**							
13. TDG-Reference Object on Near - Far Search Trial	0.05	-0.15	-.37*	-0.07	-.37*	.42*	-0.07	.69**	.38*	0.26	.82**	0.13						
14. Prior Fixations Near - Far Objects	-0.05	0.21	-0.08	0.08	-0.18	-0.12	-0.10	-0.13	-0.13	-0.23	-0.22	-0.26	-0.09					
15. Proportion Reference Trials Near Target Fixated	.35*	0.22	0.09	0.02	0.09	-0.04	0.17	-0.09	0.12	0.17	-0.17	-0.14	-0.11	0.16				
16. Flanker RT	-.52**	0.31	0.07	0.03	0.05	0.19	0.01	0.02	0.00	-0.32	-0.01	0.04	-0.05	0.05	-0.20			
17. Reference Trial Accuracy	0.27	-0.30	-0.18	-0.01	-0.21	-0.03	-.36*	-0.22	-.36*	-0.09	-0.04	-0.16	0.07	-0.16	0.24	-0.01		
18. Near Trial Accuracy	.35*	-.40*	0.29	-0.03	.38*	-.36*	0.04	-.45**	-0.31	0.32	-.39*	-0.16	-.40*	-0.21	0.12	-0.22	0.27	
19. Far Trial Accuracy	0.24	0.02	0.27	0.24	0.07	-0.30	0.06	-.45**	-0.32	0.05	-0.31	-0.23	-0.24	-0.08	0.07	-0.08	0.20	0.27

Supplemental Information

Appendix A: Additional Control Analyses for TDG-Reference

The proximity of Near and Reference objects is designed to elicit incidental fixations among these two stimuli. As expected based on object placement, children overall made more incidental fixations on Near objects than Far objects before they became the target of the search, $t(36)=5.80$, $p=.000$. This value reflects fixations to an object made at any point before it became the search target. With respect to the computation of TDG-Reference on Near object search trials, this proximity can result in incidental fixations to the Reference object immediately before the Near (but not Far) object is fixated. These fixations may artificially inflate TDG-Reference values on Near object search trials. This section describes full control analyses to ensure this is not the case. Nonetheless, to be conservative and account for this possibility, we calculated the proportion of Near trials in which the Reference object was fixated immediately before the Near object was found. This variable was used as a control continuous variable in a repeated measures ANCOVA examining the use of TDG-Reference by trial type.

To reiterate, a possible confound of the finding that TDG-Reference is higher for Near than Far trials is that the Near and Reference objects may be more likely to be in a child's field of view simultaneously than are Reference and Far objects. Thus, Reference objects may be incidentally fixated more on Near than Far trials, raising TDG-Reference scores. We ensure this is not driving our results in two separate analyses. First, if all the variance could be explained by the field of view, one would expect top-down guidance values for the Reference object on Reference trials and on Near trials to be statistically identical. However, children distributed fixations more consistent with topdown attention guidance for the Reference object on the Reference object search trials than on the Near object search trial, $t(35)=3.03$, $p=.005$. Second, we examined top-down attention guidance for the Reference object on Near object search trials versus the reverse top-down attention guidance for the Near object on Reference object search trials. If the top-down guidance values reflect only fixation distribution in the same field of view, these values should be statistically identical. A repeated measures ANCOVA comparing these top-down attention guidance values, with the continuous variable of the average proportion of trials in which children fixated the Near/Reference object immediately before the search target was found, resulted only in a significant main effect of trial type, $F(1,33)=5.90$, $p<.05$, $\eta p2=.15$. Children were more likely to distribute fixations consistent with top-down attention guidance for Reference object on Near

object search trials (*M*=.28, *SD*=.26) than for Near objects on Reference object search trials (*M*=.07, *SD*=.08).

Appendix B: NIH Toolbox Flanker Task and Data Processing

This task is part of a large cognitive assessment battery offered to streamline cognitive processing assessments across scientific laboratories. The Flanker is based on the Eriksen Flanker task (Eriksen and Eriksen, 1974), and the later use of that task in a developmental Attention Network Task battery (Rueda et al, 2004). Children are presented with a central target (fish or arrow depending on the age of the child) that is flanked on either side by two distractors (four total distractors) that are either pointing in the same (congruent) or different (incongruent) directions as the target. Participants press a button to indicate the direction that the central target is pointing. All task parameters and data processing were determined by the NIH Toolbox standard measures for both speed (RTs) and accuracy. Table 2 shows the correlations of relevant variables with this measure.

References

Amso, D.and Scerif G. (2015). The attentive brain: insights from developmental cognitive neuroscience. *Nature Reviews Neuroscience, 16,* 606-19.

Ballard, D. H., Hayhoe, M. M. and Pelz, J. B. (1995) Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*, 66–80.

Bar, M. (2004) Visual objects in context. *Nature Reviews Neuroscience, 5*, 617–629.

Bearden, C.E., Woodin, M.F., Wang, P.P., Moss, E., McDonald-McGinn, D., Zackai, E., Emannuel, Z., and Cannon, T.D. (2010). The neurocognitive phenotype of the 22q11.2 Deletion syndrome: selective deficit in visual-spatial memory. *Journal of Clinical and Experimental Neuropsychology, 23(*4), 447-464, DOI: 10.1076/jcen.23.4.447.1228

Bedard, A-C., Martinussen, R., Ickowicz, A., and Tannock, R. (2004). Methylphenidate improves visual-spatial memory in children with Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 43(3),* 260 – 268.

Boettcher, S. E. P., Draschkow, D., Dienhart, E., and Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye-movements during visual search. *Journal of Vision*, Vol. 18(13), 11. https://doi.org/10.1167/18.13.11

Brady, T. F. and Chun, M. M. (2007). Spatial constraints on learning in visual search: Modeling contextual cuing. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 33,* 798–815.

Brockmole, J. R., and Henderson, J. (2006). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, *59*(7), 1177–1187. doi: 10.1080/17470210600665996.

Brockmole, J.R., Henderson, J.M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition,13*(1):99–108.

Brown, J.H., Johnson, M.H., Paterson, S.J., Gilmore, S.J., Longhi, E., and Karmiloff-Smith A. (2003). Spatial representation and attention in toddlers with Williams syndrome and Down syndrome. *Neuropsychologia, 41*(8), 1037-46.

Chen, X., and Zelinsky, G.J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*(24), 4118-33.

Chrastil, E. R., and Warren, W. H. (2012). Active and passive contributions to spatial learning. *Psychonomic Bulletin and Review*. doi: 10.3758/s13423-011-0182-x.

Chun, M. M., and Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, *36*(1), 28–71. doi: 10.1006/cogp.1998.0681.

Chun, M. M., and Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*(4), 360–365. doi: 10.1111/1467-9280.00168.

Clark, A.C., Fernandez, F., Sakhon, S., Spano, G., and Edgin, J.O. (2017). The medial temporal memory system in Down syndrome: Translating animal models of hippocampal compromise. *Hippocampus, 27*(6), 683-691.

Clearfield, M. W. (2004). The role of crawling and walking experience in infant spatial memory. *Journal of Experimental Child Psychology*, *89*(3), 214–241. doi: 10.1016/j.jecp.2004.07.003.

Couperus, J. W., Hunt, R. H., Nelson, C. A., and Thomas, K. M. (2011). Visual search and contextual cueing: Differential effects in 10-year-old children and adults. *Attention, Perception, and Psychophysics*, *73*(2), 334–348. doi: 10.3758/s13414-010-0021-6.

de Melker Worms, J. L. A., Stins, J. F., Wegen, E. E. V., Verschueren, S. M., Beek, P. J., and Loram, I. D. (2017). Effects of attentional focus on walking stability in elderly. *Gait and Posture, 55*, 94–99. doi: 10.1016/j.gaitpost.2017.03.031

DeMaster, D., Pathman, T. and Ghetti, S. (2013). Development of memory for spatial context: Hippocampal and cortical contributions. *Neuropsychologia, 51*(12), 2415-26. doi: 10.1016/j.neuropsychologia.2013.05.026.

Dixon, M. L., Zelazo, P. D., and De Rosa, E. (2010). Evidence for intact memory-guided attention in school-aged children. *Developmental Science*, *13*(1), 161–169. doi: 10.1111/j.1467-7687.2009.00875.x.

Draschkow, D., and Võ, M. L.-H. (2016). Of "what" and "where" in a natural search task: Active object handling supports object location memory beyond the object's identity. *Attention, Perception and Psychophysics*. doi:10.3758/s13414-016-1111-x pdf

Egnor, S. E. R., and Branson, K. (2016). Computational Analysis of Behavior. *Annual Review of Neuroscience*, *39*(1), 217–236. doi: 10.1146/annurev-neuro-070815-013845.

Foulsham, T., Chapman, C., Nasiopoulos, E. and Kingstone, A. (2014). Top-down and bottom-up aspects of active search in a real-world environment. *Canadian Journal of Experimental Psycholology, 68*, 8–19.

Gibson, E. (1998). Introduction: Visually Controlled Locomotion and Orientation. *Ecological Psychology*, 10(3), 157–159. doi: 10.1207/s15326969eco103and4_1

Goujon, A., and Fagot, J. (2013). Learning of spatial statistics in nonhuman primates: Contextual cueing in baboons (Papio papio). *Behavioral Brain Research, 247*, 101-109. doi:10.1016/j.bbr.2013.03.004.

Hardiess, G., Gillner, S. and Mallot, H. A. (2008). Head and eye movements and the role of memory limitations in a visual search paradigm. *Journal of Vision, 8*(7), 1–13.

Hayhoe, M. M., Bensinger, D. G. and Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research, 38*, 125–137.

Helbing*, J., Draschkow*, D., and Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. Cognition, 196, 104147.

Hollingworth A, Henderson J. (1998) Does Consistent Scene Context Facilitate Object Perception? *Journal of Experimental Psychology General*. 127(4), 398–415.

Hwang, A. D., Higgins, E. C., and Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5), 25.1–18.

Jiang, Y. V., Won, B., Swallow, K. and Mussack, D. (2014). Spatial reference frame of attention in a large outdoor environment. J*ournal of Experimental Psychology, Human Perception and Performance, 40,* 1346–1357.

Kretch, K. S., Franchak, J. M., and Adolph, K. E. (2014). Crawling and Walking Infants See the World Differently. *Child Development, 85*(4), 1503–1518. doi: 10.1111/cdev.12206

Kunar, M. A., Flusberg, S. J., and Wolfe, J. M. (2008). The role of memory and restricted context in repeated visual search Percept Psychophys, 70(2), 314-328.

Li, C.-L., Aivar, M. P., Kit, D. M., Tong, M. H., and Hayhoe, M. M. (2016). Memory and visual search in naturalistic 2D and 3D environments. *Journal of Vision*, *16*(8), 9. doi: 10.1167/16.8.9.

Li, C.-L., Aivar, M. P., M., Tong, M. H., and Hayhoe, M. M. (2018). Memory shapes visual search strategies in large-scale environments. *Scientific Reports, 8*, 4324.

Lynn, A., Festa, E., Heindel, W., and **Amso, D.** (2019). What underlies visual selective attention development? evidence that age-related improvements in visual feature integration influence visual selective attention performance. *Journal of Experimental Child Psychology.*

Mack, S. C. and Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision, 11*, 1–16.

Newcombe, N. S. (2019). Navigation and the developing brain. The Journal of Experimental Biology, 222(Suppl 1). doi: 10.1242/jeb.186460

Nussenbaum, K., Scerif, G., and Nobre, A. C. (2018). Differential Effects of Salient Visual Events on Memory-Guided Attention in Adults and Children. *Child Development*. doi:10.1111/cdev.13149.

Oliva, A., Torralba, A., Castelhano, M.S., and Hendersson, J.M. (2003). Top-down control of visual attention in object detection. *Proceedings of IEEE Conference on Image Processing, 3*, 14-17.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *TICS, 11*(12), 520-527.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Piaget, J. and Inhelder, B. (1956). The child's conception of space. London: Routledge and Kegan Paul.

Redmon, J., and Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6517–6525).

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 91–99). Curran Associates, Inc.

Rueda, M., Fan, J., Mccandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., and Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, 42(8), 1029–1040. doi: 10.1016/j.neuropsychologia.2003.12.012

Shimi A., Nobre A.C., Astle D., Scerif G. (2014). Orienting attention within visual short-term memory: development and mechanisms. *Child Development, 85*(2):578-92.

Solman G.J.F., Kingstone A. (2014). Balancing energetic and cognitive resources: Memory use during search depends on the orienting effector. *Cognition, 132*(3):443–454.

Torralba, A., Oliva, A., Castelhano, M. S. and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features on object search. *Psychological Review, 113***,** 766–786.

Tummeltshammer, K., and Amso, D. (2017). Top-down contextual knowledge guides visual attention in 6- and 10-month-old infants. *Developmental Science*. doi: 10.1111/desc.12599.

Vaidya, C. J., Huger, M., Howard, D. V., and Howard, J. H. (2007). Developmental differences in implicit learning of spatial context. *Neuropsychology, 21*(4), 497-506. doi:10.1037/0894-4105.21.4.497.

Võ ML-H, Wolfe JM. (2012). When does repeated search in scenes involve memory? Looking at versus looking for objects in scenes. *Journal of Experimental Psychology, Human Perception and Performance, 38*(1), 23–41.

Võ ML-H, Wolfe JM. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Science, 1339*(1), 71-81.

Võ, M. L.-H., Boettcher, S. E., and Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. Current Opinion in Psychology, 29, 205-210.

Wasserman, E. A., Teng, Y., and Brooks, D. I. (2014). Scene-based contextual cueing in pigeons. *Journal of Experimental Psychology: Animal Learning and Cognition, 40*(4), 401-418. doi:10.1037/xan0000028.

Weintraub S, Bauer PJ, Zelazo PD, Wallner-Allen K, Dikmen SS, Heaton RK, Tulsky DS, Slotkin J, Blitz DL, Carlozzi NE, Havlik RJ, Beaumont JL, Mungas D, Manly JJ, Borosh BG, Nowinski CJ, Gershon RC. I. *NIH Toolbox Cognition Battery (CB): introduction and pediatric data. Monogr Soc Res Child Dev. 2013 Aug*;78(4):1-15

Wolfe, J.M., and Horowitz, T.S. (2017). Five factors that guide attention in visual search. *Nature Human Behavior, 005*8.

Yoshida, H., Darby, K., and Burling, J. (2011). Cued Attention and Learning of Spatial Context in Children. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33. https://escholarship.org/uc/item/4ph9x4pc.

Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S. II. NIH Toolbox Cognition Battery (CB): measuring executive function and attention. *Monogr Soc Res Child Dev. 2013 Aug*;78(4):16-33 (PubMed abstract)

Zelinsky, G. J., Adeli, H., Peng, Y., and Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628). doi: 10.1098/rstb.2013.0058.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 1–20.