Effects of 2D vs 3D Landmark Geometry and Layout on Young Children's Recall of Locations from New Viewpoints

James Negen[1], Hannah E. Roome[2], Samantha Keenaghan[1] & Marko Nardini[1]

[1]Durham University, South Road, Durham, DH1 3LE

[2]University of Texas at Austin, 110 Inner Campus Drive, Austin, TX 78705

Acknowledgments

Supported by grant 220020240 from the James S. McDonnell Foundation 21st Century Science Scholar in Understanding Human Cognition Program. Thanks to Colin Lever for helpful discussions and comments.

ABSTRACT

Spatial memory is an important aspect of adaptive behavior and experience, providing both content and context to the perceptions and memories that we form in everyday life. Young children's abilities in this realm shift from mainly egocentric (self-based) to include allocentric (world-based) codings around four years. However, information about the cognitive mechanisms underlying acquisition of these new abilities is still lacking. We examined allocentric spatial recall in 4.5-8.5 year olds, looking for continuity with navigation as previously studied in 2-to-4-year-olds and other species. We specifically predicted an advantage for 3D landmarks over 2D ones and for recalling targets 'in the middle' versus elsewhere. However, we did not find compelling evidence for either of these effects, and indeed, some analyses even support the opposite of each of these conclusions. There were also no significant interactions with age. These findings highlight the incompleteness of our overall theories of the development of spatial cognition in general and allocentric spatial recall in particular. They also suggest that allocentric spatial recall involves processes that have separate behavioral characteristics from other cognitive systems involved in navigation earlier in life and in other species.

Spatial memory is the metaphorical hook on which our everyday experiences hang – when having a complex interaction with another person, adapting to a new structure of potential rewards, or even having a simple percept of a new sight, sound, or smell, our perceptions come attached to locations that help us contextualize and organize our experiences. Remembering locations reliably is also crucial for many everyday tasks, both on small scales (finding keys on a cluttered desk) and large (giving directions to a friend). Schemes for remembering spatial locations can be parsed into *allocentric*, meaning world-based and relative to reference points that stay in the same place as the organism moves, and *egocentric*, meaning self-based and relative to the organism (Piaget & Inhelder, 1956). Allocentric memories are generally more useful since they remain stable even after an organism becomes disoriented or leaves an environment and returns later, though the two work together in many situations (e.g. Burgess, 2006). Allocentric spatial cognition is widely found to be more difficult (e.g. Burgess, Spiers & Paleologou, 2004; King et al., 2002; Shelton & McNamara, 2001), to develop later in life (e.g. Acredolo, 1978; Negen, Heywood-Everett, Roome & Nardini, 2017; Newcombe, Huttenlocher, Drummey & Wiley, 1998; Piaget & Inhelder, 1956), and to depend on distinct neural substrates (*see* Bird & Burgess, 2008 *for review*). Our recent studies using novel cognitive modelling analyses (Negen & Nardini, 2015; *see also* Nardini, Burgess, Breckenridge & Atkinson, 2006) and virtual reality methods (Negen et al., 2017) have shown that allocentric spatial recall emerges shortly after the fourth birthday. In this article, our main goal is to better characterize those new allocentric spatial memory skills in middle childhood, specifically 4.5-8.5 years old.

After reviewing effects seen in other spatial cognition tasks, we chose to focus on two potential factors that we expected to moderate performance: (1) an advantage for 3D landmarks, where objects in the environment with some appreciable height might allow greater precision of allocentric recall than 2D markings on the ground (e.g. Lee & Spelke,

2008); and (2) an advantage for targets ‘in the middle’ of a landmark array, allowing greater precision for those targets versus ones that lay elsewhere (e.g. Ankowski et al., 2012; Nardini et al., 2009). Since so little is known about what factors moderate performance in allocentric spatial memory in this age range, testing these two effects serves to address a major gap in the existing literature.

In the rest of the paper, we use navigation as a point of comparison for our allocentric spatial recall task here. The difference in practice is just one of response modality: recall is defined as pointing to a remembered location, whereas navigation is moving to a remembered location. Navigation tasks are much more common in the literature (e.g. Hermer & Spelke, 1994). However, recall tasks provide a stricter measure of children’s allocentric reasoning, removing any egocentric methods of succeeding at the task (see Negen et al. 2017; c.f. Stürzl, Cheung, Cheng & Zeil, 2008). Using a recall task allows us to further characterize the cognition that it elicits for comparison with navigation. We hypothesized that two major performance effects in previous navigation studies would also be present in our spatial recall task since both are directed by the goals of understanding and working with spatial information. Alternatively, effects on performance might be very dissimilar since the act of navigating always gives an opportunity to gather and use egocentric information as the organism moves around (e.g. Stürzl, Cheung, Cheng & Zeil, 2008), and a spatial recall task might be more related to scene perception or mental rotation skills. Supporting this, the ability to navigate to remembered locations after being disoriented emerges much earlier, from 1.5-2.0 years old (Hermer & Spelke, 1994), versus 4 years old for a task like ours (Negen et al., 2017). By using virtual reality to force participants to make responses without the aid of any egocentric information, we are both (a) characterizing allocentric spatial recall in greater detail and (b) seeing how similar the use of allocentric information by itself is to

the various processes that support navigation, not just in terms of when they emerge but also what determines performance.

2D and 3D Landmarks

Studies regarding navigation suggest a strong advantage for 3D objects over 2D objects, both in behavioral performance during early childhood and neural coding. These findings have been part of a wider line of theory development that clarifies the importance of boundaries (which must be 3D, at least in the sense that term is generally used) in guiding navigation across a wide variety of species, tasks, short-term timeframes of reconciling incoherent spatial perceptions, and relatively long-term stages of early development (e.g. Gallistel, 2017; Lee, 2017; Xu, Regier & Newcombe, 2017).

Of particular interest here, there is strong evidence that 3D landmarks are privileged in an early-developing form of human navigation: reorientation. In the basic paradigm, adapted from work with rodents (Cheng, 1986), children between two and four years old are put in a rectangular enclosure and watch a toy being hidden in one of the corners (Hermer & Spelke, 1994). They are then gently disoriented by turning with their eyes closed and allowed to search for the toy. Several studies have seen children at the same age succeed when the enclosure is a 3D rectangle of walls and fail when it is a 2D marking on the ground (Lee & Spelke, 2008; 2010; 2011). This behavior still stands even when visual features in the 2D/3D cases are very similar (Lee & Spelke, 2010) or the 3D case is visually subtle compared with the 2D (Lee & Spelke, 2011). The one caveat is that the 3D items need to still have some extended length along the ground – tall, thin columns do not allow reorientation to happen at the same age as walls that are longer, but shorter in height (e.g., Lee & Spelke, 2008). This suggests that a navigation-relevant cognitive system already exists at 2-4 years old with little or no sensitivity to 2D objects, which could provide an advantage for 3D landmarks in an

allocentric spatial recall task later in life if navigation and allocentric spatial recall are strongly continuous.

In addition, neuroscience studies suggest that navigation relies heavily on specialized *boundary vector cells* (e.g. Lever et al., 2009) and that these cells are not sensitive to the presence of 2D landmarks. These interface with *grid cells* and *place cells* (e.g. Moser, Kropff & Moser, 2008) to form an integrated system supporting navigation. Recent work suggests that the boundary vector cells are only sensitive to the presence of 3D objects, not 2D ones (Poulter, Lee & Lever, personal communication; in preparation), and that 3D geometry is primary in recovering hippocampal representations after disorientation (Keinath, Julian, Epstein, & Muzzio, 2017). Despite potential issues with using rodents to understand the neural underpinning of navigation in humans, neuroimaging studies with adult humans have begun to provide evidence for comparable organization (e.g. Doeller, Barry & Burgess, 2012; Ekstrom et al., 2003).

All of these results showing a 3D advantage come from navigation tasks or from freely-behaving mammals that are exploring a space. Since these paradigms always allow egocentric information to be gathered and applied through the act of movement, it is not clear that they will apply to a task that forces allocentric information alone to be used. One of our aims that motivated a study of allocentric spatial recall is to see if the results suggest differences between the cognition underlying navigation and recall. This is done by seeing if the 3D advantage is also present in a task like the one described here.

In The Middle

Performance at navigating to targets “in the middle” may arise earlier and be easier than more flexible kinds of navigation. Evidence for this claim comes from a widespread theoretical interest in the development and use of multiple landmarks to create a coordinated and durable spatial memory, rather than relying on singular beacons that may more easily

shift or be mis-remembered, in both humans (e.g. Piaget & Inhelder, 1956; Uttal, Sandstrom & Newcombe, 2006) and animals (e.g. Collett, Cartwright, & Smith, 1986). This has led to many studies that involve the presentation of a target in relation to an array, followed by the deformation of that array in a way that preserves specific ways of encoding the target but damages others, in order to investigate what systems of relations are encoded and when.

Unfortunately, studies that directly compare targets in the middle versus elsewhere in the exact same circumstances are lacking, but comparing results across studies is still illuminating. Ankowski et al. (2012) found that two-year-olds frequently search by nearby landmarks around the space when a target is 'in the middle', but at 3 years old most searches were closer to the midpoint than any of the four landmarks. As part of the developmental time course, from the age of 3-5 years old, children will readily generalize learning about the middle of one simple geometric environment to a variety of other simple geometric environments (Tommasi and Giuliano, 2014). Children aged 4-5 years old also seem able to navigate to the middle of two landmarks despite the landmarks being moved further apart (Uttal, Sandstrom & Newcombe, 2006). Overall, this suggests that between the ages of 2-5 years, children are acquiring some reliable skill at navigating to remembered targets 'in the middle'.

In contrast, children aged 4-5-years old show poorer performance at navigating to places that are marked out by other relations to landmarks. Children at 4-5 years failed to consistently navigate to a target hidden at either of two sides of a landmark with a distinctive face (Nardini et al., 2009), even performing significantly *below* chance in one condition at 4 years. Children under 5 also show lower performance at an un-cued radial arm maze (Overman, Pate, Moore & Peuster, 1996), especially in terms of the last four searches, when compared to children over 5 years. Children at 4 years can remember a target in a distinctive container with very high accuracy, but failed to distinguish two identical containers when the

single distinctive and two identical containers were in a stable regular triangle (Lee, Shusterman & Spelke, 2006), forcing them to remember left/right information instead of just being "in the middle". The overall pattern is that children under 5 years are able to navigate to targets in the middle of several identical landmarks, but have much greater difficulty navigating to targets with other spatial relations to landmarks such as left vs right.

Further nuancing the possible effects of targets 'in the middle' is the distinction between memory precision and memory bias. A long line of studies have found a *category adjustment effect* or *prototype effect*, where responses are biased towards prototypical placements in some kind of space – the middle being the most common and most early-developing prototype (e.g. Holden, Curby, Newcombe & Shipley, 2010; Huttenlocher, Hedges & Duncan, 1991; Huttenlocher, Newcombe & Sandberg, 1994). The usual explanation is that the prototype serves as the continuous stand-in for a categorical form of memory. For example, if you just remember that a glass was "on the table", that does not specify where on the table, so you begin looking for the glass in the continuous space by letting the center of the table stand in as the best example of being "on the table". If both the categorical information and some additional metric information are available, the two are combined by Bayesian (or Bayesian-like) inference and the result is a small but consistent bias towards the middle. Based on this kind of theory, we could expect to see some bias to respond in the middle for a spatial recall task even if the continuous component of memory is not more accurate there. Our analyses will have to account for this distinction.

In terms of language development, it is of interest that the phrase "in the middle" arises much earlier than many other parts of spatial language, which might also suggest that 'in the middle' is a particularly easy spatial relation to work with. Ankowski et al. (2012) provided a clear demonstration of early language comprehension in a study that used an expansion paradigm in navigation. Children were trained to navigate over to a location that

was marked by four posts that were close together. They were then shown the same four posts spread apart. Two-year old children who were explicitly told the target was being hidden "in the middle" navigated over to the middle of the expanded post array. In contrast, none of them navigated to the middle in an un-cued control condition. In addition, children failed a test of 'left' versus 'right' until six years old (Cox & Richardson, 1985), and performed poorly on 'in front' and 'behind' until then as well. The fact that this piece of language develops so early could reflect an early fluency with the concept of 'in the middle' as children navigate.

Since we hypothesized that navigation and allocentric spatial recall would show strong continuity, we expected to see a similar advantage for targets 'in the middle' in our allocentric recall task here. This again means that we are both characterizing recall better and allowing a comparison between recall and navigation.

The Present Study

Our task is an immersive virtual reality task modified from a recently-validated test for the development of allocentric spatial recall (Negen et al., 2017). Children are shown a target in immersive virtual reality and then 'teleported' gently to a new position in the environment. Following this, participants then had to point to where they thought the target was in the environment. Children under 4 years were able to successfully point towards the target when they were either teleported to the encoding viewpoint, allowing egocentric view matching[1] (e.g. Stürzl et al., 2008), or allowed to walk to a new viewpoint, allowing egocentric self-motion strategies (e.g. Newcombe, Huttenlocher, Drummey & Wiley, 1998), but not when teleported to a new viewpoint, which defeats both and requires allocentric

[1] Another strategy that might seem even simpler could also help to a small extent. The participant could practice putting their body in a specific position to point to the target when it was visible, then try to return their body to that position on recall. This could be done even with their eyes closed at the recall timepoint. Since the participant walked in between encoding and recall, they couldn't be guaranteed to have their body oriented in the same direction as encoding, so the use would be limited, but it could potentially be an additional method to reinforce the memory in these trials that allow egocentric strategies. This method is also defeated by having to teleport to a new viewpoint.

spatial recall. Children over 4 years achieved success at all conditions. This result, a classic egocentric-to-allocentric shift, validates the method by establishing that it is able to capture well-known and robust effects in terms of spatial cognition.

We modified that method here to ask further questions about allocentric spatial recall. Specifically, we removed the test trials that allowed any egocentric strategy (i.e. a matching viewpoint or the presence of helpful self-motion information) and focused all of our available testing time on trials where participants were teleported to a new viewpoint. We also changed the landmarks and the walkway to objects that allowed a simpler 2D vs 3D manipulation – specifically, a flat checked mat versus a checked box. We also changed the selection of targets to include one that was exactly 'in the middle' of the center points of the two landmarks and others counterbalanced around the environment (rather than random targets in an area). Otherwise, the procedure remained very similar. In brief, the task involved the participant watching a small animated character (a mouse) dig underground, but the participant is then 'teleported' to a new viewpoint, and has to point at the character's location by referencing either a pair of 2D (flat) or 3D (raised) symmetric landmarks. Based on previous work (Negen & Nardini, 2015; Negen et al., 2017), we tested children aged 4.5 to 8.5 years old and expected allocentric recall from even the youngest year. These changes allowed us to move over from looking for an egocentric-to-allocentric shift in younger participants, towards examining effects that moderate performance with an allocentric frame in participants who had already developed that skill.

Based on a hypothesis of continuity with navigation effects seen in children aged 2-4 years and in other mammals such as mice, we expected two factors to further moderate performance. First, precision will be higher when there are 3D visual landmarks versus similar 2D landmarks. Second, there will be an advantage for remembering locations that are in the middle of two landmarks. We also expected an effect of age (older participants being

more accurate) and rotation (shorter teleports leading to more accuracy), which we analyzed to check the basic validity of the task.

Method

Participants

Participants were 15 children aged 4.5-5.5 years, 16 children aged 5.5-6.5 years, and 17 children aged 6.5-8.5 years. All participants were recruited in and around the Durham, UK area, with full consent from a parent or guardian. Children were eligible for inclusion if their parents reported no developmental concerns or disorders. Demographics were not collected. There was no paid incentive, although children were offered a small reward at the end of the study. The study had approval from the local ethics committee.

The youngest age range chosen was taken from children's competency at allocentric spatial recall from 4 years old (Negen et al., 2017). Therefore, we expect the same type of recall to be evident at some level for children aged 4.5 years old. The breakdown of age groups was based on the belief that developmental changes in spatial precision would start to slow markedly around seven years of age (e.g. Piaget & Inhelder, 1956). Thus, we expected the two years of development between 6.5-8.5 years to look more similar at that point. Finally, the chosen sample size of at least 15 participants, with a full dataset, in each age group, was based on the reliable success of navigation studies finding 2D vs 3D effects with similar participant numbers (e.g. Lee & Spelke, 2011) but vastly fewer trials.

Apparatus

This experiment was made possible by the use of 16 motion tracking cameras (Vicon Bonita, Vicon, Oxford, UK), which can track the movement of infrared markers with millimeter accuracy at up to 240 Hz when calibrated. This allowed us to track the precise positions of a head-mounted display (Oculus Rift, Oculus, Menlo Park, CA, USA), along

with a response (pointing) device and a real 3D box, in real time for a highly-responsive and high-framerate (90 Hz) rendering.

The two landmarks provided in the environment were red and yellow checked rectangular shapes which were either 2D (width 30 cm x length 120 cm) or 3D (same width and length, height 30 cm) depending on the condition. In the 3D condition the red landmark was present in both the actual and virtual environments. This allowed participants to stand on the landmark and so experience it as a real object, both inside and outside of the virtual reality; see below. Each landmark had two axes of symmetry and there was an orienting skybox present (see Figure 1). Because of the symmetry of the landmarks and the infinite-distance rendering of the skybox, target locations could only be identified precisely by using at least two of these cues, which should encourage children to examine their relation and help bring out (a) relational language such as ‘in the middle’ and (b) encoding relative to geometry, which is a major theoretical aspect of reorientation behavior (e.g. Lee & Spelke, 2011).

The environment had two landmarks set at right angles. Two landmarks were used, as by Uttal, Sandstrom, and Newcombe (2006), rather than four (e.g. Ankowski et al., 2012), to make it less ambiguous which point is ‘in the middle’. With just two landmarks, there is a single target (D) that is directly between the center points of the landmarks and has the lowest possible average distance to them. If we had used four landmarks, there would actually be four more points that were ‘in the middle’ of pairs of landmarks. This way, there is a single point that is objectively ‘in the middle’ to compare to all other targets. (We return later to a discussion of which targets may feel subjectively ‘in the middle’ as well.) We then placed the landmarks at right angles so that we could arrange for a target (H) that was equidistant from the centers of the two landmarks, and also had a sense of alignment with the landmarks

(being on the minor axes of both), but not on the line between their centers or directly on any line connecting any of their points, to see if that has similar effects on precision.

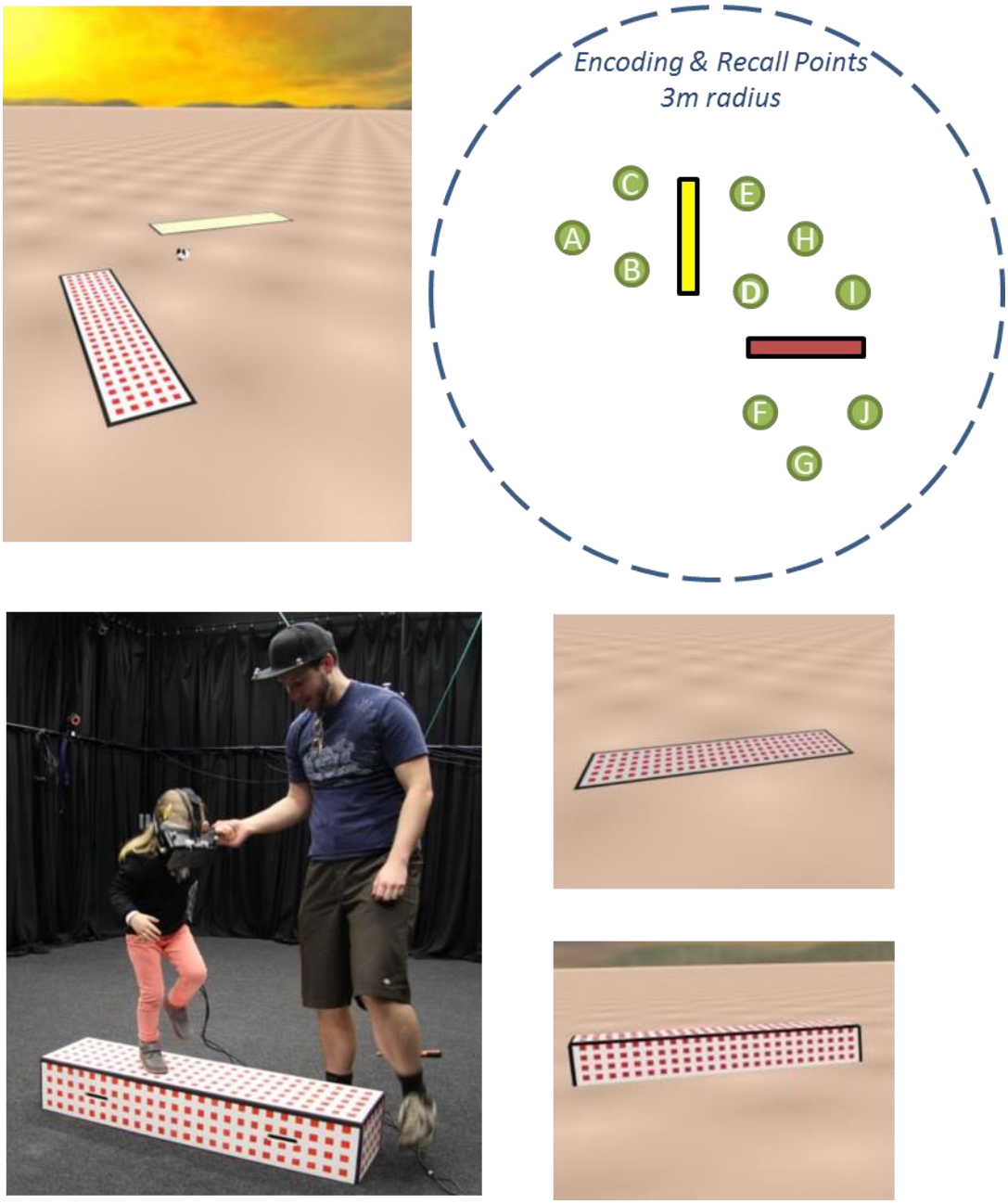

Figure 1. This was the virtual environment. Top left: a screenshot of the target item (mouse) and the two 2D landmarks. Top right: a diagram of the target locations, plus the invisible circle of viewing points at which participants stood while encoding targets and responding. Rotation magnitudes were in random order and the view was not reset between trials, so they could encode or recall from almost any point on the circle. Bottom left: a photograph of a participant interacting with the actual 3D box. Bottom and middle right: a comparison of the 2D versus 3D landmarks.

There was a red disc on the floor for participants to stand on, referred to by the experimenter as their “teleporter”. The experimenter narrating the experiment and giving instructions was represented virtually by a floating red ball with wings that was tracked to a hat that he was wearing. There was also a light checkerboard shading on the ground to help give a sense of optic flow and distance when moving and viewing the surface. The participants had a small ‘magic wand’ made from a screwdriver handle and some PVC cylinders which they used to point to signal their responses. When responding, a small red laser projected from the virtual wand to the floor, where a small inverted white cone pointing at the floor made it as clear as possible where they were actually responding.

Trial parameters and randomization. There were a total of 10 possible target locations (Figure 1). The critical ‘in the middle’ target, D, which was directly at the center point of a line connecting the centers of the two landmarks, had 4 trials per block. Another two targets were placed at reflections over each of the landmarks’ centers, C and J, also with a total of 4 trials per block (2 each). Similarly, a target that was at the intersection of the two minor axes, H, had 4 trials, and its two reflections, A and G, had 4 trials as well (2 each). An additional four targets near the remaining corners of the two landmarks, B, E, F, and I, shared 8 trials evenly. There was one block per condition (2D vs 3D) totaling 48 usable trials per child. The possible rotations were 30, 40, 50, 60, 105, 125, 145, or 165 degrees. In half of the trials this teleportation would be to the left, and half to the right.

Rotations were counterbalanced across targets so that each of three target categories (in the middle and its reflections; on minor axes and its reflections; and the remaining corners) were sampled under 90 degrees half the time and over 90 degrees half the time within each block, and so that rotations were used evenly, but otherwise random. Condition order (2D vs 3D) was counterbalanced across subjects, but all subjects saw both conditions. Within each block, no two trials in a row had the exact same target but the order was

otherwise random. The direction of the teleport was even within each block but otherwise random.

Procedure

Warmup. Before the 2D condition participants were instructed to walk over the flat red landmark, and before the 3D condition the participant stepped onto (Figure 1) and jumped off the red 3D landmark (box) with the experimenter, both inside and outside of the virtual reality. Once the headset was fitted, participants were shown how to point accurately at the target and they were given 4 trials to practice. They were required to point within 50 cm to continue. It was also demonstrated that they would have to remember the location of the target after it had 'dug' underground. They were given another 4 trials with the same criterion. Finally, they were shown the teleportation. This was done by fading the screen to black, moving the rendering camera, and then fading back up again. This proceeded from small teleports to longer ones, with a green cross indicating where they were going before each trial. The experimenter drew special attention to the green cross and emphasized that nothing in the environment would move except for the participant, their "teleporter", and the experimenter. The experimenter avatar 'teleported' along with the child so as to not serve as a useful spatial reference. The data collection trials were introduced by telling participants that the rest of the game would be just like the last trial, except they would not have a green cross to tell them where they are going first.

Data collection trials. On each trial, the mouse appeared from underground (it never moved across the surface) and the participant was allowed to view it as long as they wanted. The mouse would then 'dig' underground leaving no visible trace of its location, and the participant would be 'teleported' to a new position around a 3 m radius circle. That circle's center was denoted as (0, 0) in the program, with the center of the two landmarks laying at (-.3125, 0.625) and (.9375, -0.625) – so the center of rotation was offset by about 30 cm from

the ‘in the middle’ target. That circle itself was never shown explicitly to participants. The view rotated as well, so a child facing left of the center before teleportation would still be facing left after. The total teleport time was about 2.5 seconds (plus or minus screen refresh time) regardless of the distance of the teleport. The participant would then point to where they thought the mouse was, using the motion-tracked “wand”. A response was recorded when the pointing was stable over a period of two seconds. At this point the mouse would re-appear at its previous location and the experimenter would give verbal feedback (e.g. “Good job, it was by the red one, on that side, and on that end!” for a correct or nearby response, or “Hmm, it was on that side, but it was on the other end” for an incorrect response). After giving this feedback, the mouse again ‘dug’ underground. The view was not reset; the encoding viewpoint for the next trial was the recall viewpoint for the last trial.

A break was given between testing blocks (plus any more requested by participants). At the start of the second block, the new landmarks were introduced (i.e. they walked over the flat red landmark or stepped onto the red 3D landmark) but the other warmup trials were not repeated.

Results

First, we checked that performance was above chance for the youngest group as expected. Using the bootstrapping analysis from our previous work (Negen et al., 2017), all 15 children in the youngest age group scored above the expected chance level, significantly as a group, $p < .001$, and 11 of them as individuals (remaining *p*s: 0.10, 0.44, 0.13, and 0.25). All of the older children were significantly above chance as individuals as well. Having met the goal of eliciting allocentric spatial recall from each of the age groups, we proceeded with the planned analysis.

Analysis Plan. We chose to work with the raw error (distance from response to correct location) transformed into log units because the raw errors deviated strongly from a

normal distribution. In the next sections, we will use an ANOVA approach to look for two expected control effects, specifically an ordered age effect and a rotation effect, since we would doubt the validity of the task without these. This ANOVA also lets us look for age interactions to determine the best way to deal with the two hypothesized effects. Afterwards, we will examine the effect of 2D vs 3D landmark types with the same ANOVA calculations and some follow-up analyses, hypothesizing a 3D advantage. Following, to test if there was a hypothesized 'in the middle' advantage, we will use the ANOVA results and supplement them with sensitivity/specificity analyses and a modelling method.

Expected Effects of Age and Rotation

Since age group is the only purely between-subjects variable in the design, we looked at it separately first. We calculated an average score for each child. A one-way ANOVA found a strong expected effect of age group (4.5-5.5, 5.5-6.5, and 6.5-8.5 years old), $F(2,45) = 24.1165$, $p < .001$, demonstrating a developmental decrease in log-error. There was a significant repeated contrast for the younger versus middle group, $p < .001$, and the middle versus older group, $p = .046$. Following this, we ran a 3 (Age groups) x 8 (Rotation levels) x 10 (target locations) x 2 (2D vs 3D landmarks) x 2 (First or second testing block) ANOVA as our main analysis[2]. See Figure 2 for a general overview of the results and Table 1 for a full listing of the ANOVA calculations.

[2] We considered a repeated-measures or mixed-model analysis as our primary analysis but ultimately decided to present it as a secondary analysis in an appendix. The study was designed to give us a wide sampling of targets, rotations, blocks, and landmark types inside each age group. However, there were 10 targets x 8 rotations x 2 blocks x 2 landmark types, meaning that each child could encounter a total of 320 experimental conditions. Each child's actual dataset has 48 non-warmup observations that were sampled systematically across that possible space. As such, a repeated-measures approach would either be explicitly (e.g. multiple imputation) or implicitly (e.g. converting to z-scores or entering a subject variable) inferring most of the data. We also limited results to main effects and 2-way interactions for similar reasons.

Table 1. ANOVA to test expected age and rotation effects plus hypothesized 3D advantage (landmark type) and 'in the middle' advantage (target) and interactions.

	df	*SS*	*F*	*η2*	*p*
Age Group	2	150.58	115.833	0.079	**<.001**
Landmark Type	1	0.83	1.283	0.000	0.258
Block	1	4.84	7.445	0.003	**0.006**
Target	9	135.54	23.169	0.071	**<.001**
Rotation	7	96.08	21.117	0.051	**<.001**
Age Group*Landmark Type	2	0.69	0.531	0.000	0.588
Age Group*Block	2	3.24	2.492	0.002	0.083
Age Group*Target	18	18.64	1.593	0.010	0.054
Age Group*Rotation	14	6.49	0.713	0.003	0.763
Landmark Type*Block	1	18.43	28.350	0.010	**<.001**
Landmark Type*Target	9	4.18	0.714	0.002	0.697
Landmark Type*Rotation	7	4.75	1.044	0.002	0.398
Block*Target	9	4.41	0.754	0.002	0.659
Block*Rotation	7	2.86	0.628	0.002	0.733
Target*Rotation	63	48.34	1.180	0.025	0.159
Error	2151	1398.11			
Total	2303	1899.78			

Note. Significant effects are in bold. The significant effects in order of largest to smallest effect size are age group, target, rotation, landmark type x block, and block.

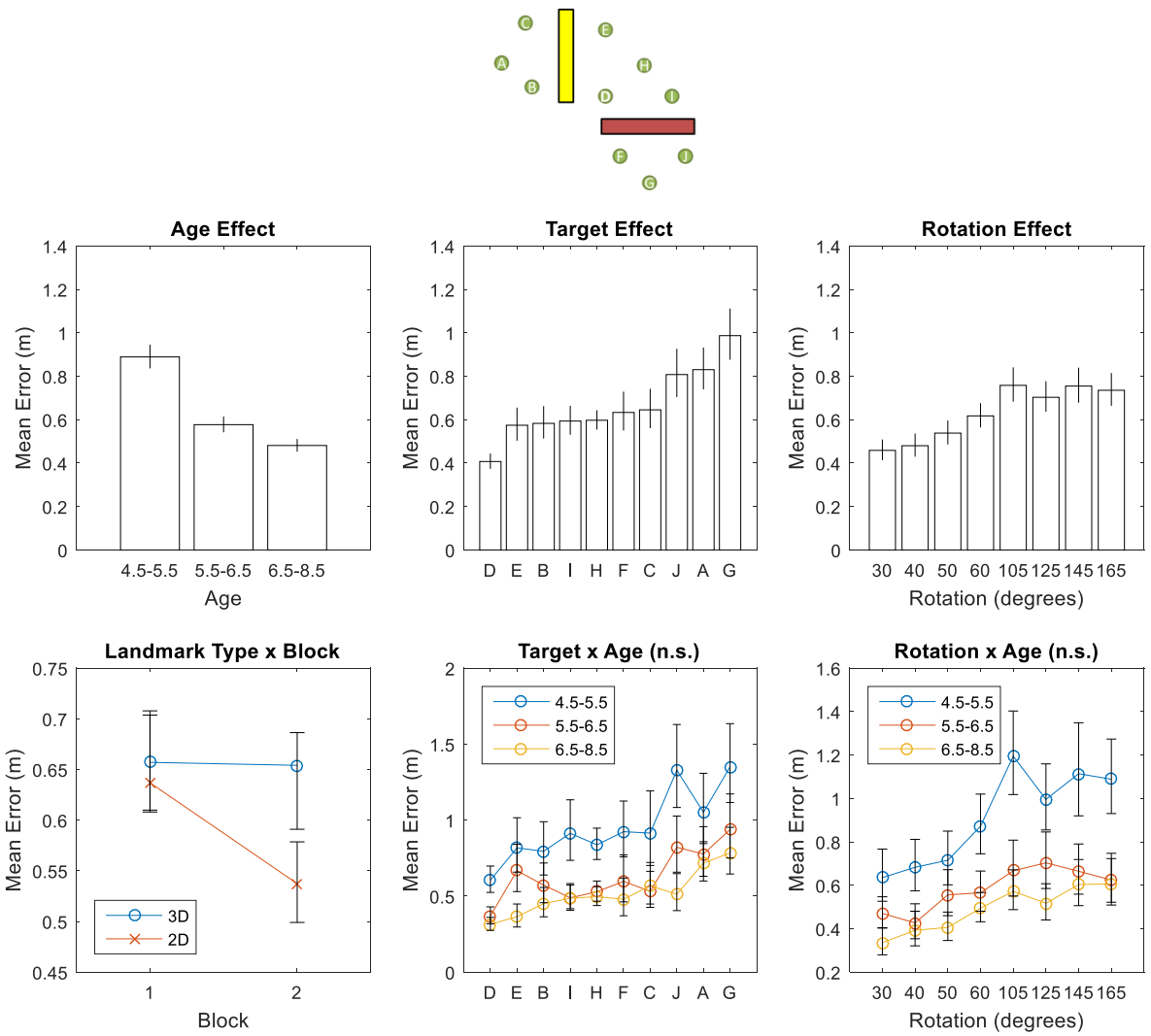

Figure 2. All scales on the y-axis have been re-converted to linear units (after being averaged in log units). Bars are 95% confidence intervals, calculated independently for each bar. There was a strong age effect (top left); a strong target effect (top middle); a strong but slightly non-linear rotation effect (top right); an interaction between landmark type and block with an accompanying main effect of block but not landmark type (bottom left); and no interaction between age and either target (bottom middle) nor rotation (bottom right).

As expected, larger rotations led to larger average log-error. We also found the effect of age group again. Therefore, both expected effects were present in the expected direction, which we interpreted as evidence that the task was sensibly measuring spatial recall performance.

2D vs 3D Performance

We did not find a main effect of 2D vs 3D landmarks. We did, however, find both a main effect of testing block with lower log-error in the second block (0.66m vs 0.59m), plus

an interaction between landmark type and testing block. A pair of post-hoc t-tests with a Bonferonni correction showed a significant effect of landmark type in the second block, t(1150) = 3.1743, p = .0015, but not the first block, t(1150) = .0848, p = .9325. As such, we not only failed to find a hypothesized advantage for 3D landmarks, but instead found a possible advantage for 2D landmarks that seems to require a level of familiarity with the task to emerge – though this should be interpreted with caution since the main effect of landmark type was not significant. There was also no significant interaction with age group (0.84m in 2D vs 0.93m in 3D for 4.5-5.5 year olds; 0.53m vs 0.63m for 5.5-6.5 year olds; 0.48m for both for 6.5-8.5 year olds).

We wanted to be sure that the pragmatics of pointing were not masking an actual 3D advantage. To check if this interaction effect is driven by the 3D boxes blocking the view of the child's response on some trials, we looked at the subset of trials where the target was on the same side of all the landmarks as the child when they were responding. The results were similar: there was no significant effect of block or landmark type, ps = .8369, .3192, but the same interaction appeared with lower average log-error for 2D landmarks than 3D landmarks in the second block, $F(1,703) = 6.2182$, $p = .0129$. It is not clear why the block effect was not significant, but one simple explanation is the reduction in statistical power. On balance, this suggests that the 3D landmarks occluding their view is not the main reason why 2D performance was better in the second block.

A number of supplementary analyses were also performed. These were variations on the primary ANOVA (e.g. including the distance between the recall viewpoint and the target as a factor) and a bootstrapping analysis to look at the spatial layout of responses in the 2D vs 3D conditions. Since the findings did not change our interpretation of the results regarding our main hypothesis, we have placed them in Appendices 2-6 for any reader who may be interested in the additional details.

Looking for an 'In the Middle' Effect

There was a strong main effect of target location (see Table 1), which did not interact with age. Our interest was not just in a general target effect but specifically an advantage for the target 'in the middle'. Figure 3 provides a map of the varying performance levels, confirming that the 'in the middle' target did indeed have the least log-error. These results so far are consistent with the hypothesis that targets 'in the middle' (target D) are easier to remember. It also speaks to the sensibility of our measure that targets further from the landmarks generally had higher log-error (Figure 3).

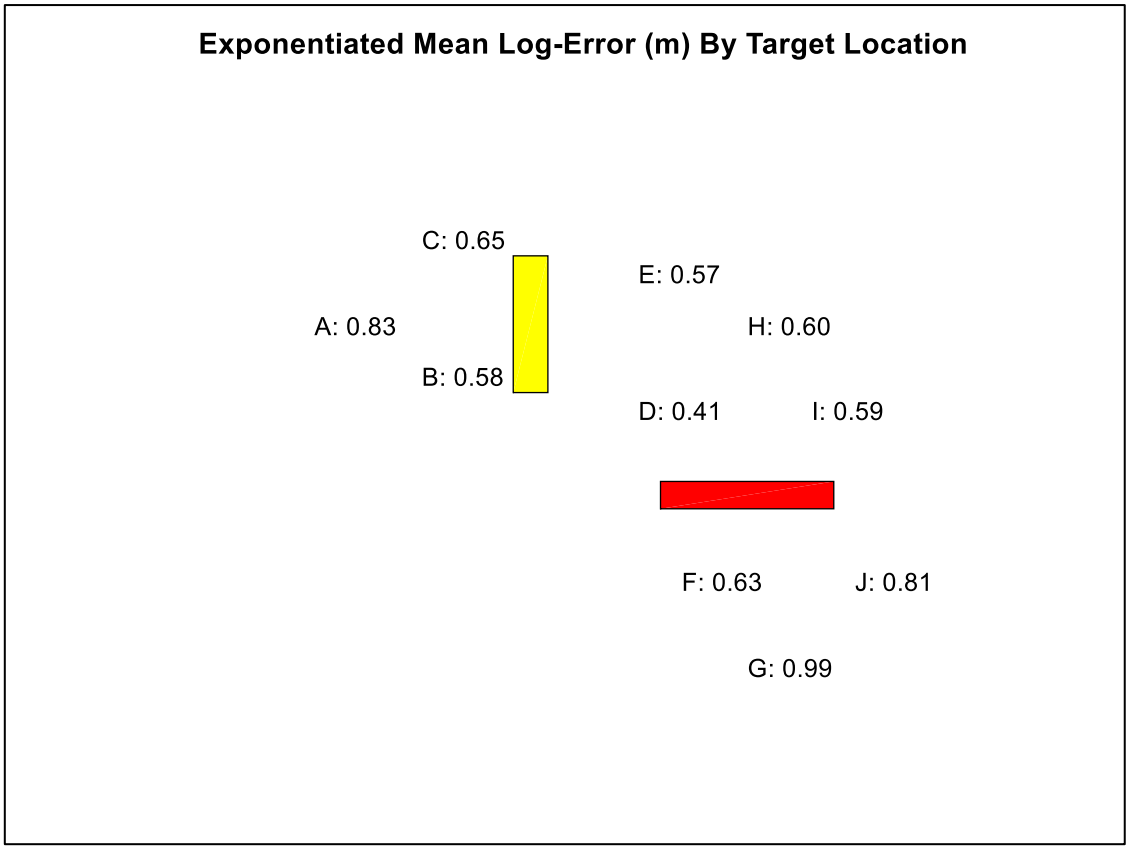

Figure 3. Error at each of the ten target locations, in meters, after taking the average of their logarithm and exponentiating them. The target D 'in the middle' had the lowest average log-error, as we hypothesized.

From the ANOVA above and Figure 3, it is already clear that participants were more accurate when they were asked to point at target D, which falls most clearly 'in the middle' (i.e. the one that is on the center point of a line segment joining the center points of the two

landmarks). However, does this necessarily mean that they were better at remembering that a target was there? Is it possible that some other process is creating this effect?

To find out, we considered whether there was an overall preference to point near the ‘in the middle’ target, D, regardless of the correct target. Table 2 shows what percent of responses fall nearest each target’s location for each actual target. The numbers down the diagonal are the rate of correct responses. If all errors were randomly placed on other target locations (e.g. they were just as likely to point at all four corners of the red box when they were not sure), then the columns should all sum to about 100%. The ‘in the middle’ target D sums to 189% and the two closest to it (targets B and F), across the major axes of the two boxes, are at 142% and 132%. The two farthest from the landmarks sum to 43% and 46%. This shows that responses are systematically biased to be towards the center of the arena in this dataset. This is a potentially serious issue for the interpretation of better performance on the ‘in the middle’ target. As an extreme example, imagine that children remembered and pointed directly to the correct location 50% of the time, but for the other 50%, they pointed directly at the ‘in the middle’ target. The accuracy at that one target would be 100% and 50% at all others, despite the fact that they are not genuinely any better at remembering targets there. Diagnostic of this would be a very high false positive rate, where they pointed to the ‘in the middle’ target when it was not actually there.

To get a closer look at this issue, we started by examining sensitivity, specificity, and positive predictive value (see Table 3). Sensitivity (true positives / (true positives + false negatives)) was the highest at the ‘in the middle’ target, D. However, specificity (true negatives / (true negatives + false positives)) was the lowest for that target and the positive predictive value (true positives / (true positives + false negatives)) for that target was in the middle of the range. In other words, participants did not do better at that target on metrics that specifically punish pointing at that target when the target is not actually there. However, this

does not necessarily mean that they were not *both* tending to point very often at the 'in the middle' target when unsure *and* more likely to remember when that was the correct location to recall – to look for that, we needed a method to separately assess their memory precision and their tendencies to guess in different places.

Table 2. Confusion Matrix between the targets and the target that responses were nearest.

		Response										
		A	B	C	D	E	F	G	H	I	J	Sum
Target	A	20%	33%	18%	8%	5%	6%	2%	2%	3%	1%	100%
	B	5%	48%	13%	11%	7%	3%	1%	4%	4%	2%	100%
	C	4%	26%	41%	6%	7%	3%	3%	5%	2%	2%	100%
	D	1%	5%	2%	65%	6%	6%	1%	7%	6%	2%	100%
	E	3%	7%	5%	24%	41%	5%	0%	7%	5%	1%	100%
	F	3%	9%	4%	10%	1%	45%	6%	4%	5%	13%	100%
	G	3%	4%	4%	7%	3%	31%	18%	5%	7%	17%	100%
	H	2%	5%	2%	22%	14%	4%	2%	36%	13%	2%	100%
	I	1%	2%	2%	24%	3%	6%	2%	12%	41%	5%	100%
	J	2%	3%	4%	11%	6%	23%	5%	3%	8%	33%	100%
Sum		46%	142%	96%	189%	93%	132%	43%	86%	96%	78%	

Table 3. Sensitivity, Specificity, and Positive Predictive Value by Target ("In the Middle" Highlighted).

	A	B	C	D	E	F	G	H	I	J
Sensitivity	0.20	0.49	0.43	0.67	0.43	0.46	0.18	0.36	0.43	0.34
Specificity	0.98	0.91	0.95	0.85	0.94	0.91	0.98	0.95	0.93	0.96
PPV	0.45	0.32	0.43	0.48	0.38	0.32	0.42	0.57	0.37	0.42

We used a Bayesian cognitive model to separately estimate the contributions of biased guessing behavior and an unbiased memory recall process. The model assumes that every trial is either completed by a guessing process or a memory process. When using the guessing process, responses are completely unrelated to the target, but the preference to guess near different targets is allowed to vary (a multinomial distribution with 10 possible outcomes). When using the memory process, the chance of pointing to each target is a function of how

close it is to the correct target (subject to exponential decay), with farther targets being less likely. This is controlled by a precision parameter. An additional parameter controls how often each process is used. Before fitting to the data, there is no built-in bias towards any particular rate of using the memory process, nor in favor of any particular place to prefer guessing. Details and code appear in Appendix 1. The results are presented in Figure 4.

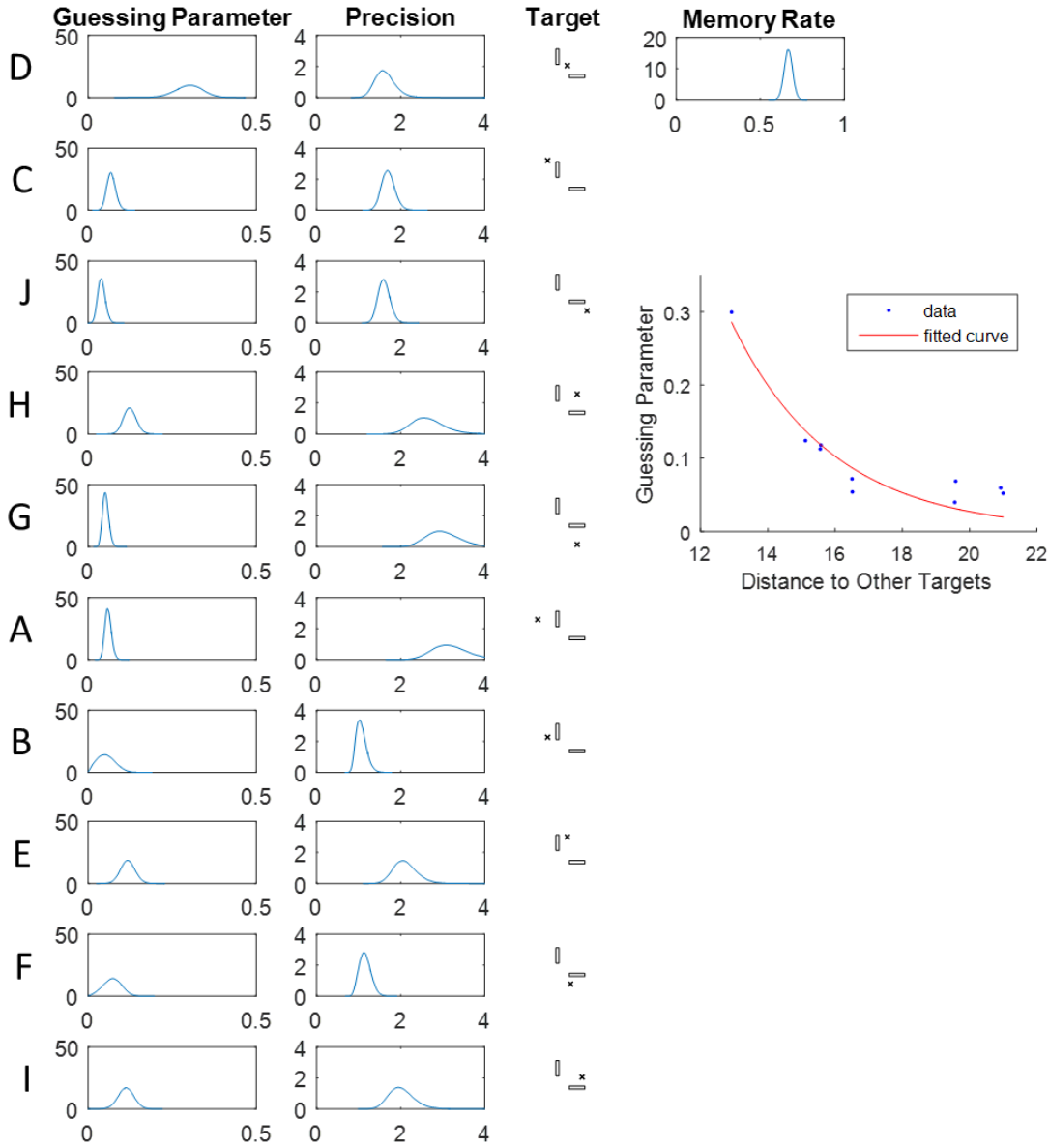

Figure 4. Posterior distributions over model parameters. The guessing parameter is predicted roughly as an exponential curve of the average distance to all other targets (i.e. how good of a guess it is with no trial-specific information). The precision is largely in three clusters which correspond to the nature of the geometric relation to the two landmarks (e.g. highest precision when they are on a minor axis of at least one landmark) and not particularly to which targets are "in the middle".

The model fit was good overall, with a correlation between predicted confusion matrix and observed confusion matrix of r = .90. The posterior mean memory rate was about 2/3 – the model inferred that for every trial where participants guessed (i.e. responded in a

way that was totally unrelated to the target), there were about two where they remembered (with some imprecision, of course).

When using the guessing process, children seem to be guessing at the ‘in the middle’ target much more often than any other target, roughly 35% of the time (as seen by looking at the peak of the distribution for target D in the Guessing Parameter column of Figure 4). This is a sensible strategy for guessing because it is the point that minimizes average error over all possible targets. Indeed, there is a roughly exponential relation between the posterior means of the guessing parameters and the average distance to all of the other targets.

The critical results in terms of actual memory precision do not support the expected advantage for a target ‘in the middle’. The posterior distributions over memory precision fall roughly into three clusters (sets of posterior distributions with a reasonable level of overlapping concentrations). A higher precision means that participants were more likely to point at the correct target and that the drop in probability of pointing is steeper as targets became further from the correct target. As you scan down the precision column of Figure 4, there are three distributions that are farther right. Those are the three targets with the greatest distance to a landmark (A, G, H), which had the highest precision. The trials including the ‘in the middle’ target and its reflections over the centers of the landmarks, plus the other two targets on the interior near the landmarks (C, D, E, I, J), had the next highest precision, falling down the center of the precision column. The two targets that were outside the landmarks and not a reflection of the ‘in the middle’ target over the landmark centers (B, F) had the lowest, falling to the left. In short, there was not any consistent measurable advantage in terms of memory precision for the target that was ‘in the middle’ (D), nor generally for targets on the interior of the landmark array versus the exterior (e.g. A and G vs H). Instead, it seems that the best precision is found on along the minor axes of the two landmarks,

followed by targets that at least lay on an intersection of lines through two edges of the landmarks.

The reader may also have some intuition that target H would be classified by some as ‘in the middle’, but the results for that target are not particularly good evidence for an ‘in the middle’ advantage either. It did not score highest in accuracy, sensitivity, or specificity, and the associated posterior distribution for its memory precision overlaps heavily with both A and G (which are clearly non-middle), with H having the lowest mode of the three. It did have the highest observed posterior predictive value, but it is not clear that this reflects any genuine memory advantage.

In summary, the modelling analysis, which allowed us to separately estimate the tendency to guess near certain targets and the precision of memory, suggests that apparent advantages for the ‘in the middle’ trials are potentially an artefact of a biased guessing process rather than evidence of a genuine advantage in terms of spatial recall.

Appendix 7 shows the results from applying the model to each of the three age groups separately, which do not show any particular difference that would change our main interpretation here; the memory rate and the precision parameters all generally increased with age, but the pattern of high guessing and middling precision for target D remained the same.

Discussion

The first expected effect, that all age groups would perform above chance when forced to use allocentric representations to recall locations, was supported by the data in line with our previous work (Negen et al., 2017). Children aged 4.5 years and above were able to systematically use allocentric representations, independent of any egocentric strategies during spatial recall, despite the potentially-disruptive effects of encoding soon after losing orientation (e.g. Knierim, Kudrimoti & McNaughton, 1995). Yet, neither of the two factors that we expected to moderate performance had compelling effects. Based on the analysis

conducted, there were three opportunities to find an advantage for 3D landmarks over 2D landmarks, but none of them followed the trend in the predicted direction. We actually found a significant post-hoc advantage for the 2D landmarks in the second testing block. We did find an accuracy advantage for the target that was 'in the middle', but when we applied a modelling method to separate out memory precision and tendency to guess at different targets, the result is that 'in the middle' is a common place to guess and is not subject to an especially-high memory precision. The effect of these two parameters did not interact significantly with age and the model results do not suggest any major shift in how target locations differ when the model is applied to each age group separately. Thus, these effects appear to be relatively consistent from 4.5 to 8.5 years old.

Placed together, these results emphasize how different and how separate allocentric spatial memory in middle childhood is from navigation in early childhood and in other species such as mice. Based on studies of neural encoding and reorientation behavior (which is a navigation task), we predicted a 3D advantage, but actually found a (limited) 2D advantage. Based on studies that have shown trends in navigation results in 2-to-5-year-olds, plus early language development, we predicted a memory precision advantage for targets that could be described with the short phrase "in the middle", but we found it to have middling memory precision. The only consistency is a tendency to bias responses toward the target 'in the middle', which is generally consistent with category adjustment models (e.g. Huttenlocher, Hedges & Duncan, 1991) rather than specific developmental navigation systems. This suggests that allocentric spatial recall in middle childhood may rely on fundamentally different mechanisms to those underlying seemingly-related tasks like reorientation or early linguistic spatial encoding. This kind of allocentric spatial recall also seems not to make use of neural representations of space related to boundaries that we currently know about from animal and human (adult) studies.

2D vs 3D Landmarks. The first expected effect, that 3D landmarks would be more useful for children at 4-8 years than 2D landmarks, found no support in our data. We considered whether a 3D advantage may be present but masked by simple pragmatics of the task. The 3D boxes have some height and thus occlude a little bit of the surface of the response area. This means that sometimes a child might want to respond somewhere that they cannot see, which may be less accurate. However, the interaction effect and the trend in means remained the same when we excluded trials where the child was not on the same side of the landmark as the target, making that explanation unlikely.

A 2D advantage was evident only during the second testing block, and it is not entirely clear why this advantage occurred. It could be a simple type II error in the first block, or it could be a more complex phenomenon, such as building up a useful prior over the spatial layout of targets that is easier to apply in the 2D landmark case. A third possibility for a 2D advantage in the second testing block is that sometimes simplicity is vital. The 3D landmarks include a greater number of surfaces to which locations could be referenced (vertical as well as horizontal). As the experiment was done in fully immersive virtual reality, they also provide potentially richer information about their own spatial layout, for example via stereopsis and motion parallax. In other contexts, young children fail to consistently select the most informative spatial references (Negen & Nardini, 2015; Nardini et al., 2006), and fail to judge spatial layout by appropriately combining multiple available cues to depth (Dekker et al., 2015; Nardini, Bedford, Desai & Mareschal, 2010). It is possible that immaturities in selecting and combining the available information appropriately meant that the additional information masked the useful signal with unnecessary noise on some trials. It would be interesting to test whether the additional 3D information starts to advantage spatial precision at older ages.

The difference in terms of 3D advantage between this task and reorientation tasks could be dependent on at least three specific factors: (1) the nature of the spatial cognition being demanded, (2) the nature of the response given, and (3) the age range. For the first, reorientation tasks are possible to complete in principle by some kind of view-matching (Stürzl, Cheung, Cheng & Zeil, 2008) – though there is evidence that depth processing is involved (Lee, Winkler-Rhoades & Spelke, 2012), so it would need to be a 3D "view" (likely excluding some features other than 3D surfaces) rather than a 2D view (like a typical camera would record). For the second, we allowed children to make a continuous response anywhere in the region around the landmarks via pointing (to prevent egocentric information from being gathered during movement) while they stood in one place. Reorientation tasks typically allow a child to move themselves to one of four pre-determined possible responses, two of which are perceptually identical. An advantage for 3D information may assert itself while moving via mechanisms such as motion parallax. Although, it does appear that children rarely change their minds once they start moving towards a corner in those tasks. It is also possible that pointing puts a higher executive function load on the participants (since movement must be inhibited and the decision space is larger), or is more related to additional individual differences such as mental rotation skills (e.g. Shepard & Metzler, 1971). Further, the typical reorientation task effectively eliminates a large range of possible responses for the child, and it is not clear what effect this has. For the third, the age range tested here is older than most reorientation studies. Although our age range does overlap somewhat with one study that has found a 3D advantage (Lee & Spelke, 2008), we did not find a landmark type by age interaction. Additional research would be needed in order to clarify exactly which factor(s) are necessary to cause the difference in behavioral patterns.

These results also bring up questions about how to interpret fMRI studies with adults. Ferrara and Park (2016) found that the retrosplenial cortex (RSC) is specifically responsive to

the presence of ‘walls’ (approximately 47cm in height) instead of ‘curbs’ in a visual display providing supporting evidence for a neural representation of boundaries. This is taller than the 3D landmark we showed the children – but those children are also significantly shorter than the adults, especially the youngest ones, so it is likely to represent a comparable impediment to movement (at least under purely visual analysis). This brings up questions about what that coding in RSC is used for, if it appears earlier in life, and if it is sensitive to the same parameters (scaled or unscaled to participant height). It would also be useful to understand how this interacts with developing hippocampal neurons (e.g. Wills & Cacucci, 2014). Moving that type of study into development could be very productive.

‘In the Middle’ Targets. The second expected effect, that there would be an advantage of being ‘in the middle’, was supported by a first-approximation look at raw error rates but can likely be explained by strategic or prior processes rather than actual superior memory. We found no particular memory advantage for the target ‘in the middle’ after accounting for the biased guessing behavior. We fitted a Bayesian cognitive model with a separate memory and guessing process to the data. The rate of guessing around the ‘in the middle’ target was very high, around 35%, but the memory precision was clearly lower than all three targets that were further from the landmarks, including two that are not even between the two landmarks. The precision parameter for each target is potentially explained by the availability of convenient placement in line with the axes and edges of the two landmarks. It does not, however, have any apparent effect of being on the interior of the two landmarks, nor ease of naming – one target that is “on the side of the red box farther away from the yellow” has much higher precision than the target that is “in the middle” (or even “near the corner”). As such, we cannot interpret the findings here as evidence that ‘in the middle’ is an easier spatial recall problem at this age, despite the many reasons to hypothesize that it might be.

The correct broader interpretation of the bias towards the center could take several forms. It could be an effect that is specific to this task and context, such as an explicit strategy – the target 'in the middle' is generally the best place to guess if you can estimate the distribution of targets over the course of the experiment. It could even be a simple prior belief that a mouse would typically prefer to hide near scene features rather than in a relatively-barren stretch of an environment. The bias found here could reflect something larger and less specific to this task as well. This could be an example of an extreme category adjustment effect (e.g. Huttenlocher et al., 1991), where the center serves as a prototype for categorical knowledge (e.g. "it was near the landmarks").

Many of the questions about 'in the middle' could be clarified a great deal by studies that present younger children with the opportunity to navigate to a target 'in the middle' and to targets elsewhere in otherwise identical settings. A simple bias to respond 'in the middle', found here, could explain some of the effects described in the introduction. Three navigation studies have trained children to respond 'in the middle' in one set of task parameters and then observed them responding 'in the middle' after those parameters change (Ankowski et al., 2012; Sims & Getner, 2008; Tommasi and Giuliano, 2014). However, they did not train any children to respond anywhere else, and they did not test what children would do without the training. It is not clear that their post-training behavior reflects the generalization of learning during the training rather than a bias to respond 'in the middle' in a new situation. Examining children's biases and priors in spatial memory tasks with different geometries and contexts, and also how they evolve with experience in each particular task at different ages, could be a fruitful line of future research. More generally, an understanding of how guessing behavior functions should be of interest for a wide variety of cognitive and perceptual tasks (Jones, Kalwarowsky, Braddick, Atkinson & Nardini, 2015).

Impact on Theory

Taken as a whole, these results highlight the incompleteness of our theories of how the whole of spatial cognition develops and functions in humans. We were not able to produce a novel, testable, verified prediction about allocentric spatial memory in middle childhood from (a) the study of early navigation performance, (b) the study of mammalian spatial neurons, or (c) the study of early spatial language development. An account of the detailed behavioral characteristics of developing allocentric spatial memory in middle childhood may have to be built from first principles rather than relying on those other areas of study for similarities. For example, we found that some targets were remembered with greater precision than others, but our explanations for why those specific targets were remembered better are all post-hoc.

To explain the cognitive processes involved in allocentric spatial memory, we may have to look to areas of study that are not traditionally classed with navigation behavior under the broader umbrella of spatial cognition. It could possibly be more related to mental rotation (e.g. Shepard & Metzler, 1971), for example. This certainly fits with the finding that rotation had a strong effect, which did not interact with age, though the effect does seem to decelerate above 105 degrees in the current dataset rather than remain linear (see Figure 2). It is also unclear whether there is something specific about the details of the task that make it similar to mental rotation in that aspect or if that would generally be found in any similar task. Another possible example is the process of scene perception (e.g. Henderson & Hollingworth, 1999), which occurs largely in brain regions outside the hippocampus (e.g. O'Craven & Kanwisher, 2000). Generally speaking though, studies in that area deal with recognizing similar scenes from different perspectives, often focusing on how saccades and other eye movements find the relevant features for recognition, rather than classing points inside the scene by their relative accuracy – so it also is not clear how to explain these data with those theories. A third example is a kind of categorization process (e.g. Shafto, Kemp, Mansinghka & Tenenbaum,

2011), though we would also need a theory of how 3D spatial relations are computed to be the inputs of the categorization. In general, it is possible that we should be analyzing allocentric spatial memory in middle childhood as a case of discontinuous conceptual development (e.g. Carey, 2004, 2009; Spelke, Lee & Izard, 2010), though it will require much further study to determine that conclusively.

Future Study

To continue examining this quandary, we suggest further study to identify the factors that explain differing levels of performance in terms of allocentric spatial recall during this period of development. The model results suggest that geometric convenience is a potential factor, and that strategic or prior choices have a strong influence on overall behavior, and that being interior versus exterior to the array is not largely relevant for the actual memory component. Knowing more about what kind of landmarks or landmark arrays are superior, the potential relevance or irrelevance of strong orienting cues, the influence of cognitive factors such as executive function, and even results of similar tasks in other age ranges, could help gain insight into the development under examination here. In studying this, we encourage other researchers to ideally use modelling methods to explicitly remove the effect of biased guessing behavior, or at least to examine sensitivity and specificity, since these can reveal patterns that are quite different and more nuanced than raw accuracy. We still only know a few facts about what behavioral features need to be included in an accurate model of how allocentric spatial memory functions in middle childhood.

References

Acredolo, L. P. (1978). Development of spatial orientation in infancy. *Developmental Psychology*, *14*(3), 224.

Ankowski, A. A., Thom, E. E., Sandhofer, C. M., & Blaisdell, A. P. (2012). Spatial language and children's spatial landmark use. *Child Development Research, 2012*.

Bird, C. M., & Burgess, N. (2008). The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, *9*(3), 182-194.

Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in cognitive sciences*, *10*(12), 551-557.

Burgess, N., Spiers, H. J., & Paleologou, E. (2004). Orientational manoeuvres in the dark: dissociating allocentric and egocentric influences on spatial memory. *Cognition*, *94*(2), 149-166.

Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus*, *133*(1), 59-68.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, *23*(2), 149-178.

Collett, T. S., Cartwright, B. A., & Smith, B. A. (1986). Landmark learning and visual-spatial memories in gerbils. Journal of Comparative Physiology A, 158, 835–851.

Cox, M. V., & Richardson, T. R. (1985). How do children describe spatial relationships?. *Journal of Child Language*, *12*(3), 611-620.

Dekker, T. M., Ban, H., van der Velde, B., Sereno, M. I., Welchman, A. E., & Nardini, M. (2015). Late development of cue integration is linked to sensory fusion in cortex. *Current Biology*, *25*(21), 2856-2861.

Doeller, C. F., Barry, C., & Burgess, N. (2012). From cells to systems: grids and boundaries in spatial memory. *The Neuroscientist*, *18*(6), 556-566.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184-188.

Ferrara, K., & Park, S. (2016). Neural representation of scene boundaries. *Neuropsychologia*, *89*, 180-190.

Gallistel, C. R. (2017). Navigation: Whence Our Sense of Direction?. *Current Biology*, *27*(3), R108-R110.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, *50*(1), 243-271.

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature*, *370*(6484), 57.

Holden, M. P., Curby, K. M., Newcombe, N. S., & Shipley, T. F. (2010). A category adjustment approach to memory for spatial location in natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 590.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological review*, *98*(3), 352.

Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive psychology*, *27*(2), 115-147.

Jones, P. R., Kalwarowsky, S., Atkinson, J., Braddick, O. J., & Nardini, M. (2014). Automated Measurement of Resolution Acuity in Infants Using Remote Eye-

TrackingAutomated Acuity in Infants. *Investigative ophthalmology & visual science*, *55*(12), 8102-8110.

Keinath, A. T., Julian, J. B., Epstein, R. A., & Muzzio, I. A. (2017). Environmental geometry aligns the hippocampal map during spatial reorientation. *Current Biology*, *27*(3), 309-317.

King, J. A., Burgess, N., Hartley, T., Vargha-Khadem, F., & O'Keefe, J. (2002). Human hippocampus and viewpoint dependence in spatial memory. *Hippocampus*, *12*(6), 811-820.

Knierim, J. J., Kudrimoti, H. S., & McNaughton, B. L. (1995). Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, *15*(3), 1648-1659.

Lee, S. A. (2017). The boundary-based view of spatial cognition: a synthesis. *Current Opinion in Behavioral Sciences*, *16*, 58-65.

Lee, S. A., Shusterman, A., & Spelke, E. S. (2006). Reorientation and landmark-guided search by young children: Evidence for two systems. *Psychological Science*, *17*(7), 577-582.

Lee, S. A., & Spelke, E. S. (2008). Children's use of geometry for reorientation. *Developmental science*, *11*(5), 743-749.

Lee, S. A., & Spelke, E. S. (2010). A modular geometric mechanism for reorientation in children. *Cognitive psychology*, *61*(2), 152-176.

Lee, S.A., Spelke, E.S., 2011. Young children reorient by computing layout geometry, not by matching images of the environment. Psychon. Bull. Rev. 18, 192–198

Lee, S. A., Winkler-Rhoades, N., & Spelke, E. S. (2012). Spontaneous reorientation is guided by perceived surface distance, not by image matching or comparison. *PloS one*, *7*(12).

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, *29*(31), 9771-9777.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS — a Bayesian Modelling Framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, *31*.

Nardini, M., Bedford, R., Desai, M., & Mareschal, D. (2010). Fusion of disparity and texture cues to slant is not mandatory in children. *Journal of Vision*, *10*(7), 494-494.

Nardini, M., Burgess, N., Breckenridge, K., & Atkinson, J. (2006). Differential developmental trajectories for egocentric, environmental and intrinsic frames of reference in spatial memory. *Cognition*, *101*(1), 153-172.

Nardini, M., Thomas, R. L., Knowland, V. C., Braddick, O. J., & Atkinson, J. (2009). A viewpoint-independent process for spatial reorientation. *Cognition*,*112*(2), 241-248.

Negen, J., Heywood-Everett, E., Roome, H. E., & Nardini, M. (2017). Development of allocentric spatial recall from new viewpoints in virtual reality. *Developmental Science*.

Negen, J. & Nardini, M. (2015). Four-year-olds use a mixture of spatial reference frames. *PLOS ONE* 10(7): e0131984.

Newcombe, N., Huttenlocher, J., Drummey, A. B., & Wiley, J. G. (1998). The development of spatial location coding: Place learning and dead reckoning in the second and third years. *Cognitive Development*, *13*(2), 185-200.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of cognitive neuroscience*, *12*(6), 1013-1023.

Overman, W. H., Pate, B. J., Moore, K., & Peuster, A. (1996). Ontogeny of place learning in children as measured in the radial arm maze, Morris search task, and open field task. *Behavioral neuroscience*, *110*(6), 1205.

Piaget, J., & Inhelder, B. (1956). The child's conception of space. *New York: Humanities Pr*.

Scott, D. (1979). On optimal and Data-Based Histograms. *Biometrika. 66* (3), 605–610.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1-25.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive psychology*, *43*(4), 274-310.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Sims, N., & Gentner, D. (2008, January). Spatial Language and Landmark Use: Can 3-, 4-, and 5-year-olds find the Middle?. In *Proceedings of the Cognitive Science Society* (Vol. 30, No. 30).

Spelke, E., Lee, S. A., & Izard, V. (2010). Beyond core knowledge: Natural geometry. *Cognitive Science*, *34*(5), 863-884.

Stürzl, W., Cheung, A., Cheng, K., & Zeil, J. (2008). The information content of panoramic images I: The rotational errors and the similarity of views in rectangular experimental arenas. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(1), 1.

Tommasi, L., & Giuliano, A. (2014). Evidence of a relational spatial strategy in learning the centre of enclosures in human children (Homo sapiens). *Behavioural processes*, *106*, 172-179.

Uttal, D. H., Sandstrom, L. B., & Newcombe, N. S. (2006). One hidden object, two spatial codes: Young children's use of relational and vector coding. *Journal of Cognition and Development*, *7*(4), 503-525.

Wills, T. J., & Cacucci, F. (2014). The development of the hippocampal neural representation of space. *Current opinion in neurobiology*, *24*, 111-119.

Xu, Y., Regier, T., & Newcombe, N. S. (2017). An adaptive cue combination model of human spatial reorientation. *Cognition*, *163*, 56-66.

Appendix 1: Details and Code for the Bayesian Model Analysis

The WinBUGS code (Lunn, Thomas, Best & Spiegelhalter, 2000) is presented in Box S1. Each target was assigned a number 1-10 and the data were pre-processed so that responses just indicated which target they pointed nearest. Five chains were run with 2,000 burn-in samples and 20,000 recorded samples. The guessing probability weights were each given a Gamma(2,1) prior and then divided by their sum to calculate the guessing probability, which means that they all have a mode at 10% and some positive probability density all the way from 0% to 100%. The precision parameters were also all given a Gamma(2,1) prior, meaning that the prior mean response distance is 1m from the target when using the memory process. Informal experiments did not suggest that there is much sensitivity to these priors in relative terms – changing the priors will shift all of the posterior distributions in absolute terms, but the posteriors remain in the same order.

Box S1. WinBUGS code for the modelling analysis.

```
model{
memoryRate ~ dbeta(1,1)
for (i in 1:10){
    guessWhichTmp[i] ~ dgamma(2,1)
    guessWhich[i] <- guessWhichTmp[i] / sum(guessWhichTmp[1:10])
    precision[i] ~ dgamma(2,1)
    for (j in 1:10){
        memoryChooseTMP[i,j] <- exp(-Distance[i,j] * precision[j])
        memoryChoose[i,j] <- memoryChooseTMP[i,j] / sum(memoryChooseTMP[i,1:10])
        p[i,j] <- memoryRate * memoryChoose[i,j] + (1-memoryRate) * guessWhich[j]
    }
    responses[i,1:10] ~ dmulti(p[i,1:10], N[i])
}
}
```

We also wanted to see if there were any obvious biases in how this model predicts data versus the actual data. We took a probit transform of the actual response rates minus a probit transform of the predicted response rates (Table S1) and looked for patterns. We did not find any in particular. One might, for example, expect that reflections over the center

point of the local landmark would be more common than our model (which provides no special increase in probability for that) would predict. That would be the pairings of A-E, C-D, D-J, and F-I. However, there were actually less responses with target A at E than expected, and similarly for C at D, D at C, D at J, J at D, F at I, and I at F. One might also expect that reflections across the single line of symmetry of the unified scene would be underpredicted by our model. That would be A-G, B-F, C-J, and E-I. However, there were less responses than predicted for F at B, E at I, and I at E. The largest inaccuracies were the overprediction of correct target selection, so it's possible that the model could be improved slightly by changing the function that relates distance to predicted accuracy to be lower at the closest actual target and to slope more gradually. However, it doesn't appear that any major issues have arisen that would place significant doubt on the large trends in the results.

Table S1. Probit of actual response confusion matrix minus probit of model posterior predictive.

	A	B	C	D	E	F	G	H	I	J
A	-0.89	0.11	0.06	-0.39	-0.04	0.15	0.06	-0.30	-0.11	0.00
B	-0.19	-0.37	-0.10	-0.43	-0.06	-0.38	-0.14	-0.09	-0.07	0.09
C	-0.11	0.05	-0.52	-0.58	-0.11	-0.19	0.13	-0.04	-0.37	0.19
D	-0.29	-0.53	-0.39	0.19	-0.27	-0.40	-0.17	-0.14	-0.25	-0.38
E	0.13	-0.31	-0.29	-0.04	-0.37	-0.08	-0.53	-0.33	-0.24	-0.31
F	0.08	0.04	0.02	-0.44	-0.63	-0.39	-0.22	-0.09	-0.25	-0.06
G	0.15	-0.07	0.11	-0.45	-0.15	0.11	-0.95	-0.01	0.11	0.04
H	-0.12	-0.21	-0.32	-0.01	0.01	-0.30	0.08	-0.27	-0.05	-0.28
I	-0.19	-0.44	-0.29	-0.03	-0.42	-0.38	0.02	-0.07	-0.37	-0.23
J	-0.08	-0.14	0.11	-0.25	0.13	0.02	-0.12	-0.19	-0.12	-0.65

Appendix 2: Percent Correct

We also looked at a discrete measure of performance: how often the response was nearer to the actual target than any of the other target locations. This yields a "proportion correct" score (note that the alternative target locations are used for scoring but participants are never in practice making a choice among the 10 locations). This measure could be less powerful but it also has some potential advantages. The main analysis of distance error is underpinned by the assumption that performance depends on the precision of spatial recall. However, when a location is completely forgotten, or misplaced due to incorrect use of a landmark and/or an "egocentric" error, the amount of distance error is not meaningful. Table S2 presents an ANOVA over proportion correct with largely-agreeing results: a main effect of age group, target, and rotation amount, plus a landmark type x block interaction.

The differences are the non-significant effect of block and a significant age group x target interaction – the main thing driving this is that the younger children had more difficulty distinguishing the 'in the middle' target from the other targets along the interior of the landmarks. There is still no significant effect of landmark type. Figure S1 shows the mean proportion correct in the same arrangement as Figure 2 in the main text, with the note that higher scores are better in terms of percent correct and the other figure displays error, where higher scores are worse.

Table S2. ANOVA over percent correct.

	df	*SS*	*F*	*η2*	*p*
Age Group	2	23.197	56.930	0.041	<.001
Landmark Type	1	0.013	0.064	0.000	0.800
Block	1	0.604	2.965	0.001	0.085
Target	9	45.246	24.676	0.081	<.001
Rotation	7	16.829	11.801	0.030	<.001
Age Group*Landmark Type	2	0.106	0.260	0.000	0.771
Age Group*Block	2	0.099	0.243	0.000	0.784
Age Group*Target	18	6.162	1.680	0.011	0.036
Age Group*Rotation	14	4.155	1.457	0.007	0.119
Landmark Type*Block	1	2.640	12.959	0.005	<.001
Landmark Type*Target	9	1.345	0.733	0.002	0.679
Landmark Type*Rotation	7	2.410	1.690	0.004	0.107
Block*Target	9	0.780	0.425	0.001	0.922
Block*Rotation	7	1.636	1.147	0.003	0.330
Target*Rotation	63	15.430	1.202	0.028	0.135
Error	2151	438.236			
Total	2303	560.822			

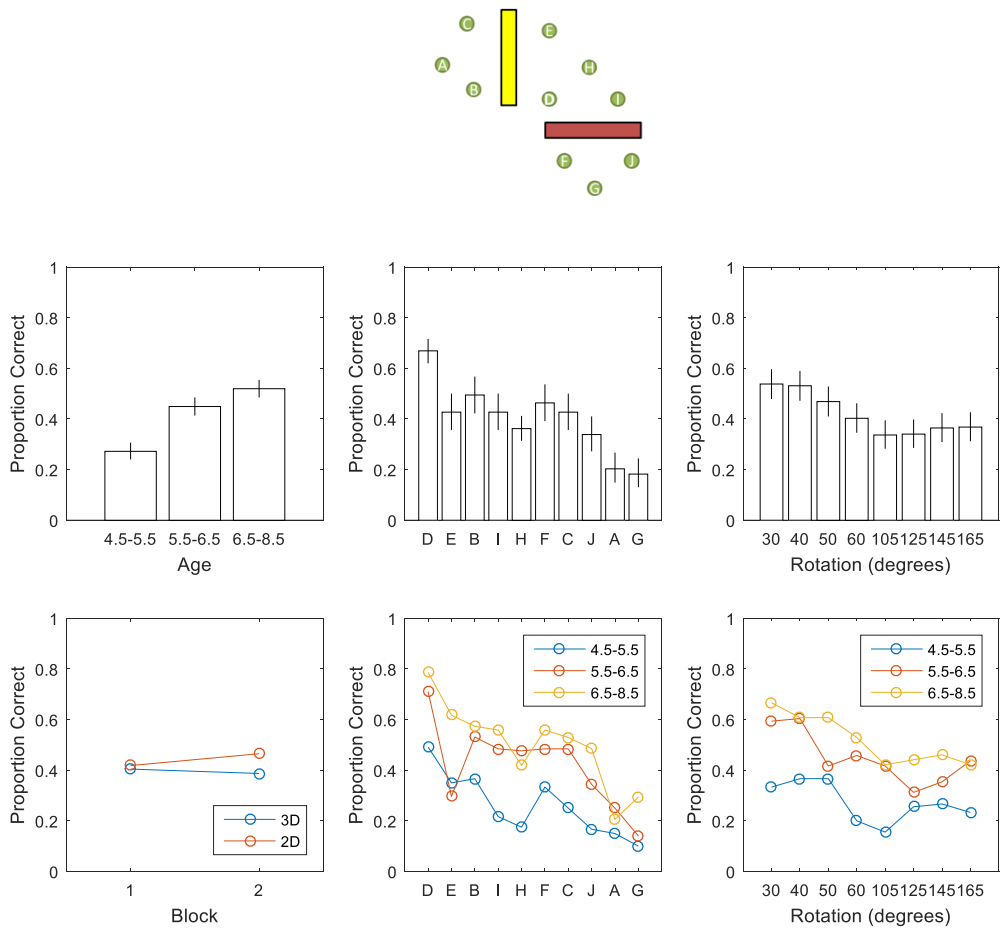

Figure S1. Results in terms of proportion correct.

Appendix 3: Primary ANOVA with Subject as a Variable

Our choice of a primary analysis without any repeated-measures component may seem unusual to some readers, so we also present an analysis where we have entered subject as a predictor variable. The results are very similar, with a main effect of block, target, rotation, and subject. Of course, it is not possible to assess an age group effect with this analysis method, nor a landmark by block interaction, since each subject is in only one age group and only experienced one order of testing. There are two additional significant outcomes here, an age group x target interaction and a target x rotation interaction. The first is driven by younger participants having larger average log-error for the targets farthest from the landmarks. The second is driven by longer rotations having a larger adverse effect on the same targets. On balance, these results support our main arguments in the text.

Table S3. ANOVA including subject as a variable.

	df	*SS*	*F*	*η2*	*p*
Landmark Type	1	0.912	1.507	0.000	0.220
Block	1	5.312	8.773	0.003	0.003
Target	9	134.988	24.771	0.071	<.001
Rotation	7	95.279	22.479	0.050	<.001
Subject	46	272.906	9.798	0.144	<.001
Age Group*Landmark Type	2	0.591	0.488	0.000	0.614
Age Group*Block	2	3.083	2.546	0.002	0.079
Age Group*Target	18	19.125	1.755	0.010	0.025
Age Group*Rotation	14	7.028	0.829	0.004	0.638
Landmark Type*Block	n/a	0.000	n/a	n/a	n/a
Landmark Type*Target	9	4.362	0.800	0.002	0.616
Landmark Type*Rotation	7	4.900	1.156	0.003	0.325
Block*Target	9	4.159	0.763	0.002	0.651
Block*Rotation	7	3.195	0.754	0.002	0.626
Target*Rotation	63	50.879	1.334	0.027	0.042
Error	2107	1275.781			
Total	2303	1899.778			

Appendix 4: Bootstrapping Analysis

To better understand the potential effects of the 2D versus 3D landmarks, we also looked at overlapping heatmaps of the responses. This could give us to some insight into what participants were doing differently in the two cases (i.e. if there was some notable pattern of places they responded in the 3D landmark case but not the 2D). We separated the data into 20 different lists: one for each target with 2D landmarks, then again for 3D landmarks. (In other words, we collapsed over age, rotation, and testing block.) We smoothed the data in each list with a kernel density estimation method, using Scott's rule of thumb (Scott, 1979) for the bandwidth, at a resolution of 1 cm^2 per pixel for each target. We then looked at their differences across landmark types but within targets. This gives us a map of the relative local density of the responses in the two conditions. We then needed a way to separate these differences into ones that are statistically significant versus those that could plausibly just be noise. To test this, we developed a bootstrapping method.

To illustrate how this works, imagine a scenario where you would use a t-test. There are two groups of data with 64 observations each, with a mean difference of 2 and each with a standard deviation of 1.5. The t-distribution is useful here but the test assumes that both are normally distributed. If both are highly non-normal, there is another method available. If the data from the two sets are pooled together and randomly sampled to form two groups over and over, the mean difference can be calculated for each remixing. After many remixes, the mean differences will have a certain apparent distribution. This distribution, which may or may not resemble a t-distribution, will have extremes at both edges. The actual mean difference can be placed onto this distribution and it is possible to infer if that mean difference is extreme or not when the two groups actually have the same underlying distribution. In other words, you can estimate the likelihood of seeing the observed mean difference if the group labels are actually irrelevant and the true mean difference is zero. This

method is generally not as powerful but it is not laden with any assumptions about the distribution of the data. This is important here because there is no clear assumption about how differences in local density should be distributed.

We randomly sampled the two sets (within targets but across 2D vs 3D) and re-computed the smoothed heatmaps, searched each for its largest difference, and repeated 1000 times. This gives us an empirical bootstrapping distribution of the kind of peak differences that are likely to occur by chance in these maps. This simulation lets us estimate a cutoff difference that is unlikely to occur in more than 5% of random arrangements (across 2D versus 3D) for each target. We saw that the heatmaps of the actual data only exceed this cutoff in three places (bright white spots in Figure S2), all in favor of higher 2D density and all near the target. This suggests, again, that there were no advantages for the 3D condition. Moreover, it does not suggest any particular area was strongly distracting the 3D responses away from the target, so it’s possible that they were not systematically incorrect but just simply more diffuse.

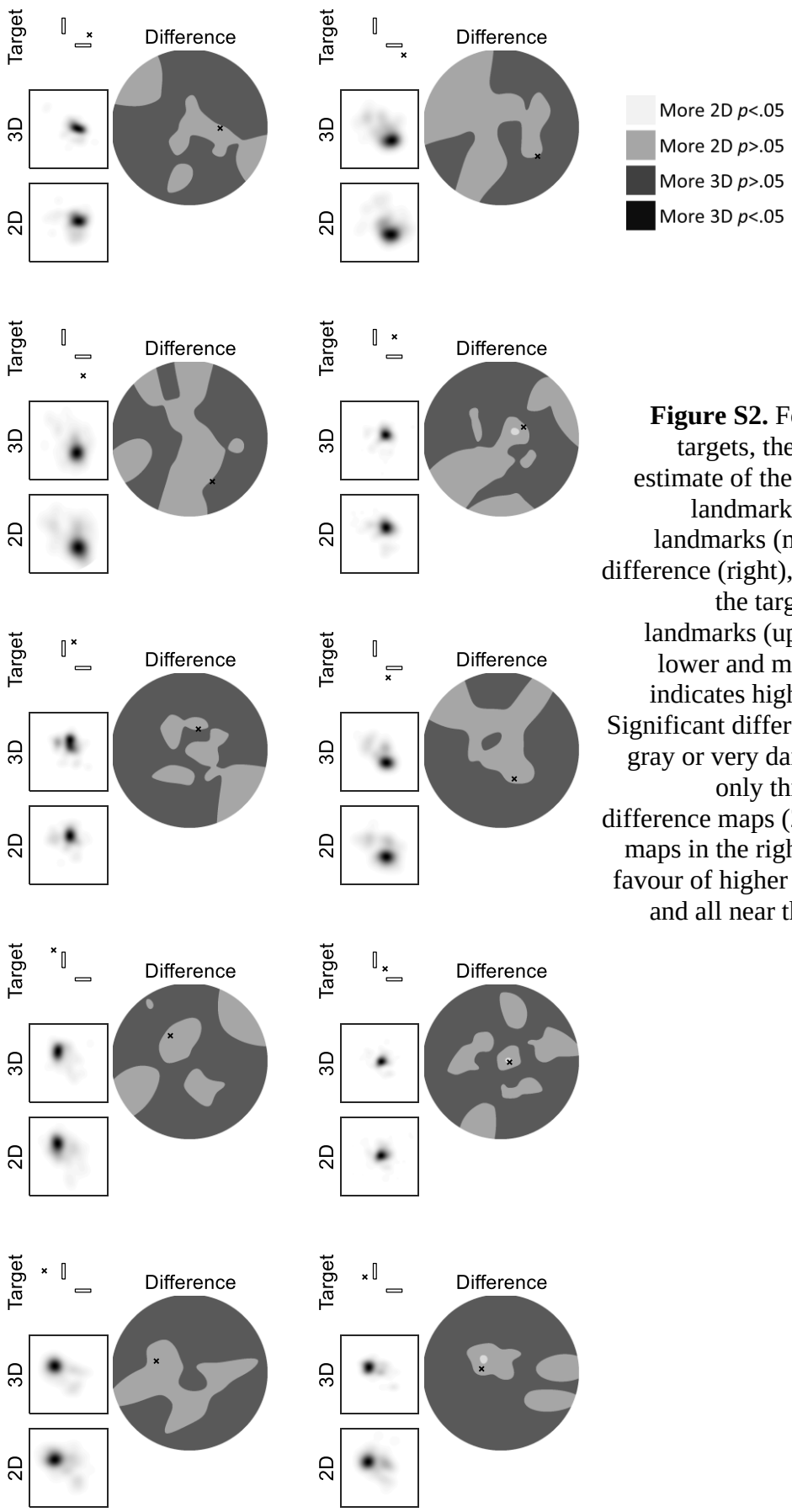

Figure S2. For each of the 10 targets, there is a smoothed estimate of the density with 2D landmarks(lower left), 3D landmarks (middle left), their difference (right), and a display of the target relative to the landmarks (upper left). On the lower and middle left, darker indicates higher local density. Significant differences (very light gray or very dark gray) occur in only three places on the difference maps (2nd, 4th, and final maps in the right column), all in favour of higher local 2D density and all near the actual targets.

Appendix 5: Bayesian Version of the Primary ANOVA

The Bayesian approach to data analysis is becoming more popular and more trusted in Psychology, so we also present a version of our primary analysis in the Bayesian tradition. These figures were calculated in JASP 0.8.1.1 (https://jasp-stats.org/). In the right-most column, figures above 3 represent strong evidence that the effect exists and figures below -3 represent strong evidence that they are not present. The results largely agree with the main analysis, with a main effect of rotation, target, age group, and block, plus a landmark type x block interaction. There is also a main effect of landmark type, suggesting that 2D performance was overall better than 3D performance. Results are inconclusive for a Block x Age Group interaction, and the rest of the effects are positively denied.

Table S4. Bayesian analysis results.

Effects	P(incl)	P(incl|data)	Log(BF $_{Inclusion}$)
Age Group	0.922	1	32.32
Target Number	0.922	1	32.32
Rotation	0.922	1	32.32
Block	0.922	1	11.941
Landmark Type x Block	0.428	1	11.803
Landmark Type	0.922	1	9.687
Age Group x Block	0.428	0.09	-2.027
Age Group x Landmark Type	0.428	0.015	-3.906
Age Group x Target Number	0.428	0.003	-5.408
Landmark Type x Rotation	0.428	0.002	-5.881
Target Number x Rotation	0.428	0.002	-5.943
Block x Target Number	0.428	6.765e -4	-7.009
Landmark Type x Target Number	0.428	3.434e -4	-7.687
Block x Rotation	0.428	3.319e -4	-7.721
Age Group x Rotation	0.428	5.576e -5	-9.506

The appearance of the landmark type effect, as a sidenote, an instance of an interesting but little-discussed advantage that Bayesian analysis can have. The Bayes factor over inclusion is calculated by estimating the posterior probability of all models that can be formed by combining the different main and interaction effects, then looking at the ratio between the prior probability of all models that have each effect to the total posterior probability of selecting a model that has the effect present. One of the outcomes of this is that it automatically 'weights' the importance of other effects for analysis. The final posterior calculation included few models that feature interaction effects other than the block x landmark type one. An ANOVA that includes just the main effects and the block x landmark type interaction also finds the same landmark type main effect, $F(1,2282) = 6.7094$, $p = .0097$. Removing the additional interactions has this effect because (1) it increase the error degrees of freedom and (2) parts of the sum of squares for the removed interactions are re-assigned to the main effect. The Bayesian ANOVA does this kind of trimming online, automatically, without the loss of the ability to draw conclusions about the additional interactions. In other words, the overall block effect could be a genuine one that only the Bayesian analysis could detect.

Appendix 6: Distance to Target

The planned analysis, presented in the main text, does not include the distance between the recall point and the correct target location as a covariate. This was planned as such because that distance is somewhat conflated with the target location; targets that are near the middle of the array are constrained in how far they can be away from the recall point when compared to targets on the exterior. However, the two can be quantified independently on each trial, it seems plausible that having to point farther away would lead to higher error, and some readers may be interested in any effects it may have. We ran an additional unplanned 3 (age group) x 2 (landmark type) x 2 (testing block) x 10 (target) x 8 (rotation) x continuous (distance from target to recall point) ANOVA. The results are in Table S4.

Under this analysis, all main effects are significant as well as the landmark type x testing block, age group x target, and target x distance to view interactions. Comparing this to the main analysis, the biggest change is that 2D landmarks led to significantly smaller errors than 3D ones. As with other alternative analyses, there was also an age group x target interaction, where the youngest children had greater errors when faced with the most difficult targets (those furthest away from the landmarks). However, note that this is a move from p = .054 to p = .046, so it does move the p-value over the significance threshold but only by changing it by less than 1%. Finally there was also an interaction between target and distance to view, with the target directly in the middle and the two furthest targets showing the least sensitivity to the distance to view (Figure S3).

Table S5. ANOVA including distance to the target.

	df	*SS*	*F*	η^2	*p*
Age Group	2	13.61	10.54	0.007	<.001
Landmark Type	1	2.63	4.08	0.001	0.044
Block	1	3.36	5.20	0.002	0.023
Target	9	28.66	4.93	0.015	<.001
Rotation	7	10.63	2.35	0.006	0.021
Distance to View	1	11.03	17.10	0.006	<.001
Age Group * Landmark Type	2	0.34	0.27	0.000	0.766
Age Group * Block	2	2.36	1.83	0.001	0.161
Age Group * Target	18	18.93	1.63	0.010	0.046
Age Group * Rotation	14	6.18	0.68	0.003	0.792
Age Group * Distance to View	2	0.66	0.51	0.000	0.600
Landmark Type * Block	1	16.41	25.43	0.009	<.001
Landmark Type * Target	9	2.27	0.39	0.001	0.940
Landmark Type * Rotation	7	5.15	1.14	0.003	0.335
Landmark Type * Distance to View	1	2.12	3.29	0.001	0.070
Block * Target	9	4.35	0.75	0.002	0.665
Block * Rotation	7	3.39	0.75	0.002	0.629
Block * Distance to View	1	1.30	2.01	0.001	0.157
Target * Rotation	63	47.10	1.16	0.025	0.187
Target * Distance to View	9	12.61	2.17	0.007	0.021
Rotation * Distance to View	7	1.20	0.27	0.001	0.967
Error	2130	1374.66			
Total	2303	1899.78			

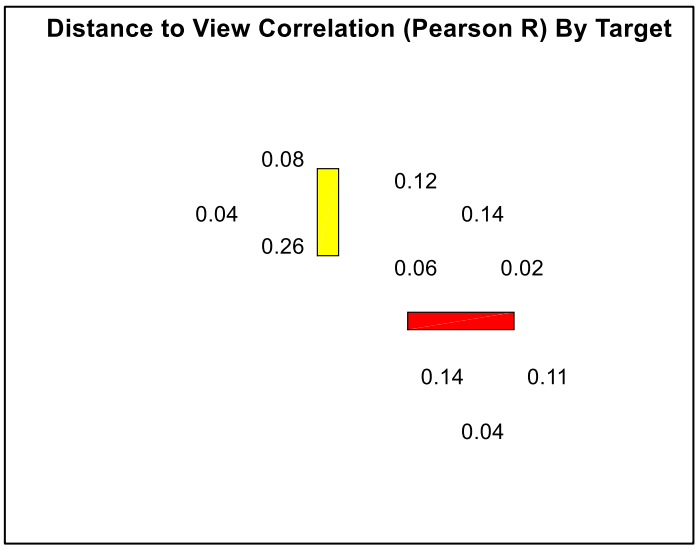

Figure S3. Correlation between the distance to view and the log-error broken down by target location.

Appendix 7: Modelling Results in Separate Age Groups

The modelling results in the main text used the full age range from 4.5 to 8.5 years. This leaves open the question of if or how the results might change across smaller age ranges. In this appendix we give the results obtained when running the model on the three age groups: 4.5-5.5 years, 5.5-6.5 years, and 6.5-8.5 years (see Figures S4-6).

As one would expect, the memory rate trends upwards as age increases. This means that in the youngest group, guessing parameters have smaller credible intervals and precision parameters have larger credible intervals, and vice versa for the older group. It is also true that the overall size of credible intervals has generally increased, since fewer data are available for each analysis. However, beyond this, results look broadly similar. For all three ages, the target ‘in the middle’ has a relatively high guessing parameter and a precision that is not especially high. If there is any noteworthy change at all, it is the fact that precision was lower for the fourth target down in Figure S4 for the youngest children, indicating that the target that was equidistant from the two landmarks but not directly in the middle was more difficult for the youngest children.

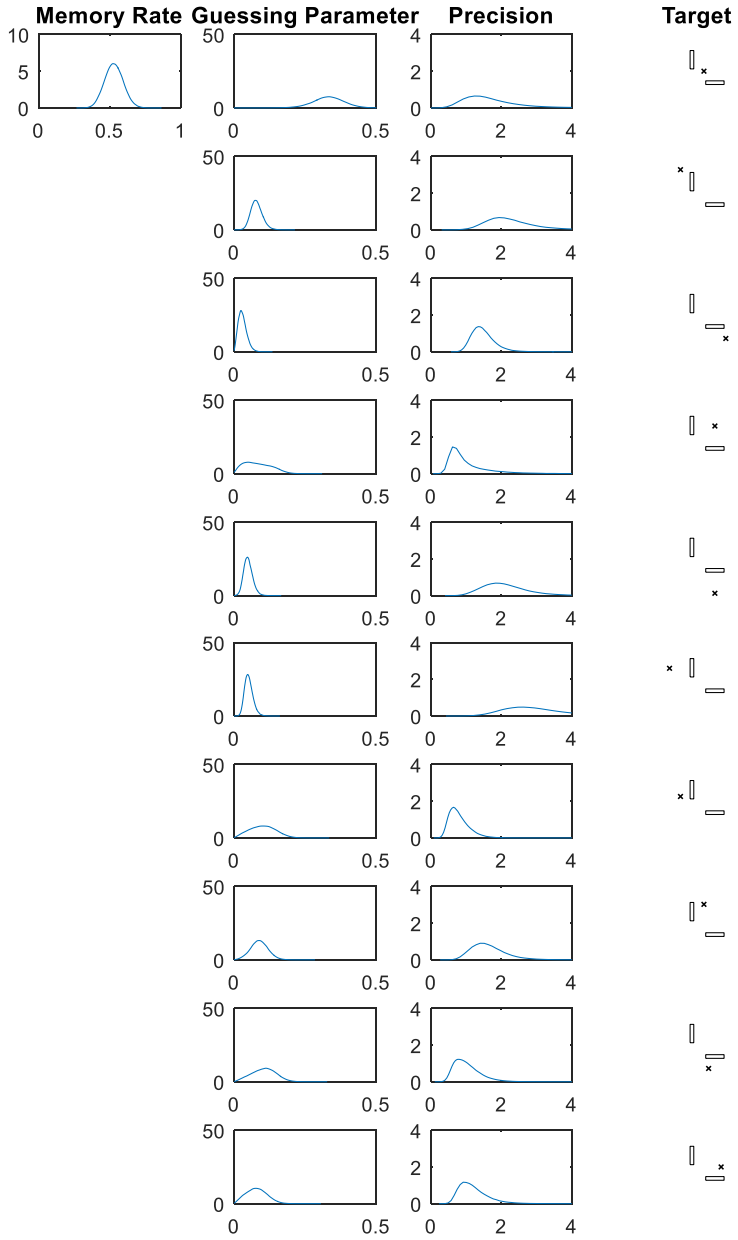

Figure S4. Model results for participants aged 4.5-5.5 years.

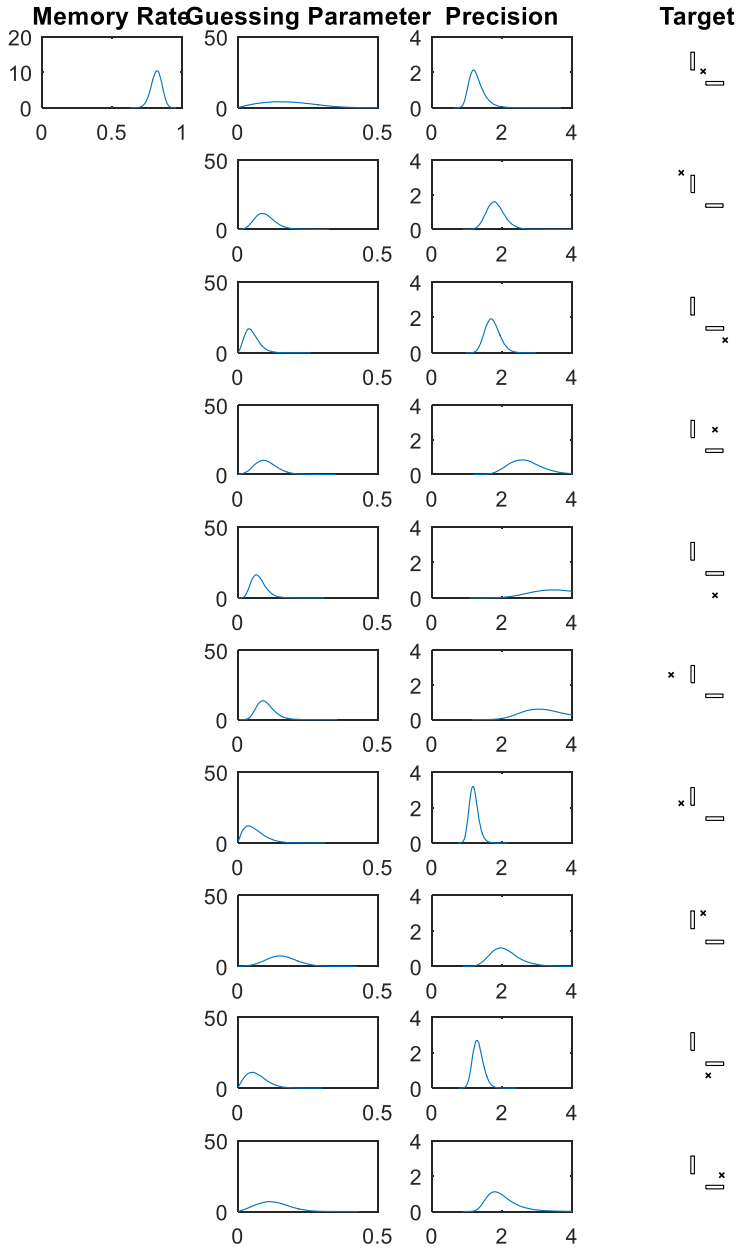

Figure S5. Model results for participants 5.5-6.5 years old.

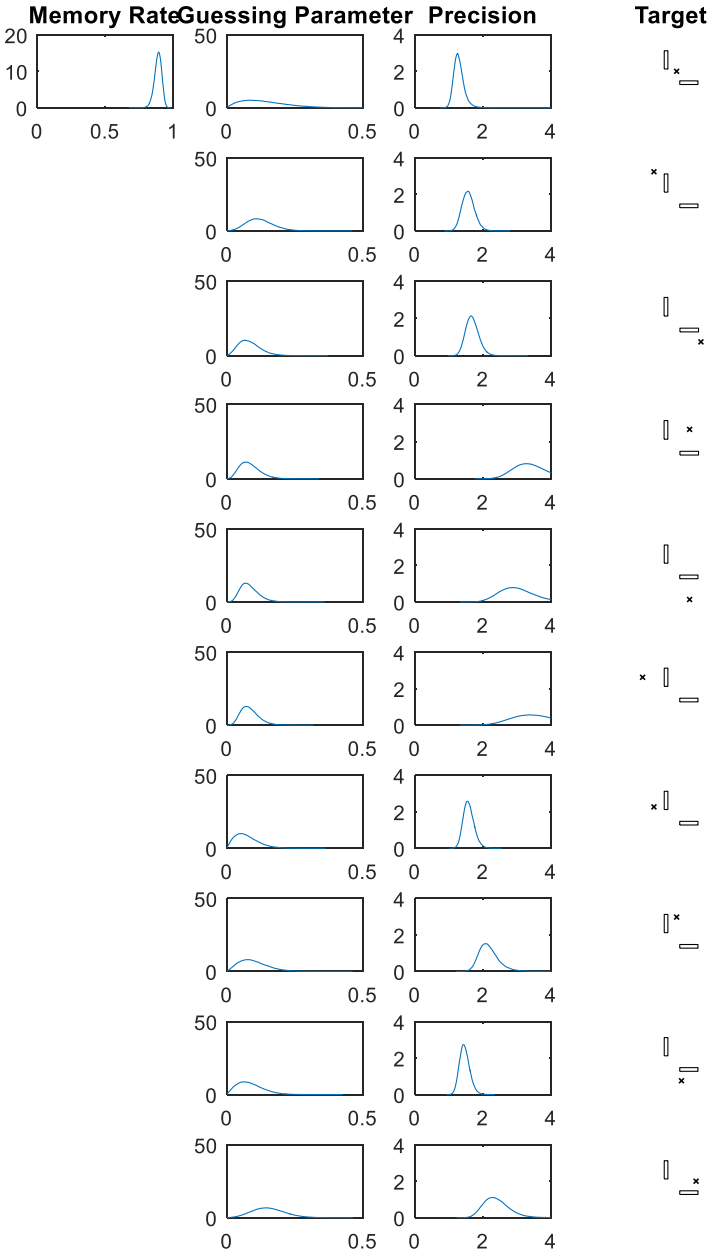

Figure S6. Model results for participants aged 6.5-8.5 years old.