Cognitive Psychology 124 (2021) 101357

Contents lists available at ScienceDirect

Cognitive Psychology

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

Evaluating recall error in preschoolers: Category expectations influence episodic memory for color☆

Kimele Persaud [a,*], Carla Macias [a,*], Pernille Hemmer [b], Elizabeth Bonawitz [a,c]

[a] Department of Psychology, Rutgers University, Newark, NJ, USA
[b] Department of Psychology, Rutgers University, New Brunswick, NJ, USA
[c] Graduate School of Education, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Episodic memory
Prior category knowledge
Color categories
Cognitive development
Probabilistic models
Relational memory

ABSTRACT

Despite limited memory capacity, children are exceptional learners. How might children engage in meaningful learning despite limited memory systems? Past research suggests that adults integrate category knowledge and noisy episodic traces to aid recall when episodic memory is noisy or incomplete (e.g. Hemmer & Steyvers, 2009a,b). We suspect children utilize a similar process but integrate category and episodic traces in recall to a different degree. Here we conduct two experiments to empirically assess children's color category knowledge (Study 1) and recall of target hue values (Study 2). In Study 1, although children's generated hue values appear to be noisier than adults, we found no significant difference between children and adult's generated color category means (prototypes), suggesting that preschool-aged children's color categories are well established. In Study 2, we found that children's (like adult's) free recall of target hue values regressed towards color category means. We implemented three probabilistic memory models: one that combines category knowledge and specific target information (Integrative), a category only (Noisy Prototype) model, and a target only (Noisy Target) model to computationally evaluate recall performance. Consistent with previous studies with older children (Duffy, Huttenlocher, & Crawford, 2006), quantitative fits of the models to aggregate group-level data provided strong support for the Integrative process. However, at the individual subject level, a greater proportion of preschoolers' recall was better fit by a Prototype only model. Our results provide evidence that the integration of category knowledge in episodic memory comes online early and strongly. Implications for how the greater reliance on category knowledge by preschoolers relative to adults might track with developmental shifts in relational episodic memory are discussed.

1. Introduction

Reconstructing events from memory is an important facet of cognition, given that it informs how we perceive, interact with, and reason about the world around us. As with all computational processes, human memory is limited in its capacity and resolution, raising questions of how the mind handles the reconstruction of events from memory. That is, how do we strategically encode information that

☆ These data were presented at the Annual Meeting of the Society for Mathematical Psychology and the Annual Meeting of the Cognitive Science Society.

* Corresponding authors.
E-mail address: kimele.persaud@rutgers.edu (K. Persaud).

https://doi.org/10.1016/j.cogpsych.2020.101357
Received 12 March 2020; Received in revised form 26 August 2020; Accepted 25 October 2020
Available online 10 November 2020

supports later use, while minimizing effort, error, and expansive storage? This question is doubly interesting for young children whose memory systems are still developing. Relative to adults, children have comparatively limited cognitive resources (Davidson, Amso, Anderson, & Diamond, 2006; Diamond, 2006; Keresztes, Ngo, Lindenberger, Werkle-Bergner, & Newcombe, 2018) and their ability to maintain information in memory becomes compromised when faced with increased cognitive load (e.g. increased inhibition demands). Thus, an important question of development is what process might be adopted by the minds of young learners to reduce uncertainty (i.e. noise or error) when reconstructing information from memory?

There is a considerable amount of research investigating reconstructive memory in adults, which might be informative for an investigation into the processes children use. For adults, the reconstructive process is predicated on our ability to integrate multiple streams of information to reduce the uncertainty in the information recalled (e.g. Hemmer & Steyvers, 2009a). As such, our knowledge and expectations about the world (i.e. semantic memory) and our episodic memories are intricately intertwined. For example, recalling what I ate for breakfast last Tuesday (i.e. episodic memory), might be influenced by my category knowledge of foods associated with breakfast and my expectations for my morning routines (i.e. semantic knowledge). An array of scientific work provides evidence that adults develop prior semantic knowledge and expectations that are well calibrated to the statistical regularities of the environment (e.g. Griffiths & Tenenbaum, 2006). In turn, adults use this knowledge to optimally perform a broad range of cognitive tasks including categorization (Huttenlocher, Hedges, & Vevea, 2000), reasoning (Oaksford & Chater, 1994), and generalization (Tenenbaum & Griffiths, 2001). This interaction further extends to the domain of memory (e.g., Bae, Olkkonen, Allred, & Flombaum, 2015; Allred, Bae, Olkkonen, & Flombaum, 2015; Huttenlocher, Hedges, & Duncan, 1991; Hemmer & Steyvers, 2009a,b), where prior category expectations can both inform whether certain information is worth storing with higher fidelity, as well as provide a means to "fill in" partially-stored information later.

While once thought of as a source of memory distortions (Brewer & Treyens, 1981), knowledge and expectations for a stimulus category can improve average recall (Huttenlocher et al., 1991, 2000). For example, Huttenlocher et al. (2000) found that people quickly develop expectations for the underlying categorical distribution of stimulus features and use this knowledge to fill in noisy and incomplete memories. They demonstrated that responses regressed toward the mean of the overall category, thereby, improving average recall. This relationship between prior knowledge and episodic memory can be captured within a simple Bayesian framework, which assumes that prior knowledge and expectations for the environment is optimally combined with noisy episodic content to produce recall of episodic experiences (Hemmer & Steyvers, 2009a,b; Huttenlocher et al., 2000; Persaud & Hemmer, 2014). Bayesian models have been instrumental to the understanding of the optimality of using category expectations when reconstructing events from memory.

Although it is clear that adults rely on category knowledge to inform recall, it is less clear the degree to which children rely on a similar process for reconstructed episodic memories. Interestingly memory at earlier stages of development (i.e. in children) prioritizes storing generalized semantic and category knowledge over fine grain episodic details (e.g. Keresztes et al., 2018). That is, children will often remember general facts about the world (e.g. The Nile is the longest river) or scripted information about events (e.g. what happens at birthday parties), but not many episodic details (e.g. where did I learn about the Nile River? Or what kind of cake did I have at the birthday party?). With age, children start to show better memory for more nuanced episodic information (Drummey & Newcombe, 2002). Of course, semantic knowledge is also still rapidly developing in early childhood, which might impact the strength of reliance on that knowledge during recall. In this way, the integration of semantic knowledge and episodic information for reconstructive memory might look markedly different between adults and younger children. By exploring this critical period of memory development, we can investigate important questions regarding the contribution pre-existing semantic knowledge plays in the reconstruction of episodic memories across development. Furthermore, we can discover whether developmental changes in episodic memory performance are marked by the ability to efficiently integrate category knowledge and noisy episodic representations.

To date, little research has specifically investigated these intertwined aspects of memory in early childhood. For example, Duffy, Huttenlocher, and Crawford (2006) used assumptions of a computational model known as the Category Adjustment Model (or CAM – Huttenlocher et al., 1991) to evaluate the contribution of category knowledge to memory in children. CAM assumes that if category knowledge is integrated in memory, recall would exhibit regression to the mean effects. The model also assumes that the noisier the episodic information, like memories in younger children, the stronger recall will regress to the mean. To evaluate these assumptions, Duffy et al. (2006) asked 5–7-year olds to study and then recall the sizes of artificial objects (e.g. pictorial fish). Category knowledge for the object sizes was learned through the task (i.e. based on the frequency with which objects were presented at select sizes).

They found that like adults, children's recall regressed toward the mean of the underlying category distribution. This means that on an individual trial, a child might not remember the exact studied size, so they use their learned category knowledge of the most frequently studied object sizes to help reconstruct the true size. They also found that memory in younger children exhibited steeper regression to the mean patterns, relative to older children. They concluded that children use category knowledge to estimate stimulus features from memory. Based on these results it appears that younger children, like adults, might be integrating category knowledge and episodic memory to reconstruct events. Although information from both sources might be noisy, the goal of the system might still be to combine them. If this is the case, an important task for research is to understand how these pieces of information are weighted, whether this varies by age, and most importantly, why this may be the case.

Given that children seemingly rely on different systems or have access to different kinds of information at different stages of development, there are a number of alternative processes that might explain their recall performance. For example, one alternative is that younger children are not integrating category knowledge and noisy episodic information, especially when the initial episodic memory traces have severely low fidelity. In early childhood, category knowledge is still developing and the ability to encode and retrieve episodic memories, even moreso (Keresztes et al., 2018). In younger children, episodic information might be so noisy and inexact that a better process to reduce uncertainty is to rely solely on semantic knowledge for reconstruction and not integrate episodic

memories. In other words, there may be a developmental shift in the use of the integrated process across memory development. In this case, reconstructive memory for some younger children might be better explained by a 'category only' model as opposed to an integrated model. Additionally, since early childhood prioritizes category information over fine grain episodic content, there might be a significantly greater reliance on the category information overall.

Moreover, a 'category only' model could also capture the regression pattern observed in the Duffy et al. (2006) experimental data. The younger children in their sample might not be integrating information, but the initial memory traces could have such low fidelity, that they are solely relying on category information. Estimates based solely on category information could also result in regression to the mean patterns. Strict reliance on category knowledge might be one factor that contributes to the general pattern of worse memory performance in children, relative to adults. Since competing models might capture the same behavioral patterns in recall, it is unclear which process children are using.

A second alternative is that children may lack clearly developed pre-existing knowledge for a given study domain, and thus may be reluctant to incorporate that knowledge when reconstructing study information. Instead, reconstruction might rely solely on noisy target episodic memories. For example, work by Hitch, Woodin, and Baker (1989) demonstrated that children make use of different coding processes based on age. More specifically, when tasked with recognizing objects, young children tend to rely on visual coding and older children rely on a combination of visual and phonological (labeling) coding. Therefore, it could be the case that for particular domains that are underdeveloped children are relying on their visual coding process (i.e., noisy target episodic trace) without considering prior knowledge.

Lastly, the process for reconstruction might not be uniformly adopted across all preschool-aged children (or in people of any age). Thus, a third alternative is there might be individual differences in the processes that explain performance in early development. While some children might rely on one process, such as the integrative process described above, another might rely solely on category knowledge. Even further, a third child might not have clear category knowledge in the study domain, and opt to not use category knowledge at all, and instead make estimates based on the noisy episodic information. In this way, the reconstructive process in younger children might be more nuanced than what was previously considered.

Evidence that younger children in general might be using a less mature or a different cognitive process to facilitate memory reconstruction comes from traditional studies of episodic memory in children. These studies of episodic memory test how well children are able to bind multiple features of episodic events in memory (Lloyd, Doydum, & Newcombe, 2009; Sluzenski, Newcombe, & Kovacs, 2006; Newcombe, Lloyd, & Ratliff, 2007; Ngo, Newcombe, & Olson, 2018). In these tasks, children are presented with a study event (e.g. a color paired with an object) and are asked whether they recognize each feature on its own and then together. The hallmark finding of these memory studies is a dramatic improvement in performance between 4 and 6 years of age.

An important caveat of these studies is that semantic category information is removed from the stimulus environment to avoid a confound of episodic memory with semantic knowledge (i.e. there is no a priori associations between the episodic information that needs to be bound). However, the removal of such useful information might interrupt the ability of younger children to employ the cognitive process of integrating category information, and thus further hinder performance. The poorer recall in younger children might result from an inability to integrate category knowledge for at least one feature while trying to recall the episodic events. Critically the influence and potential over-reliance on category information in children's reconstructive memory can be predicted and elucidated via computational approaches. However, to the best of our knowledge, previous work has not computationally evaluated this relationship in preschoolers, neither has assessed individual differences in memory reconstruction in children.

The focus of this work is to identify the developmental origins of a well-known finding in the memory literature, that individuals employ category knowledge to help recall information from memory. We also aim to explore whether differences in this integrative process might contribute to some of the differences in memory performance between young children and adults. Children's memory resources and capacities are still developing and for some age groups, have yet to reach adult-like levels, but this is a separate question from whether the mechanisms and processes that facilitate recall are also in place in development and to a similar degree. Differences in the integration of category knowledge between children and adults might account for some of the variability in memory performance between age groups that was once heavily attributed to differences in capacity constraints. To better understand the computational goals of the memory system, namely episodic and semantic memory, and why the system has adapted to behave this way, it is important to evaluate memory systems as they develop (see Yee, Jones, and McRae (2018) for discussion of using development across the lifespan to understand semantic memory).

In this study, we employ computational approaches to evaluate the impact of category knowledge on memory outcomes in early childhood to reveal the memory processes that best explain preschooler's recall performance. Probabilistic modeling provides the opportunity to acquire a more fine-tuned understanding of memory and learning across varying domains of knowledge. When implemented at the aggregate level of analysis (i.e. fit to aggregate data across participants), computational models can shed light on the general computational processes that underlie cognitive performance (Gopnik & Bonawitz, 2015), such as the reconstruction of events from memory. At the group- and individual-level of analysis, probabilistic modeling can reveal nuanced individual deviations in the reconstructive process. Here we seek to extend previous studies of memory in two substantive ways: (1) evaluating if the process of integrating knowledge and memory comes online earlier in memory development (e.g., during the preschool years) and (2) by testing alternative models that might explain younger learner's recall behavior at both the group and individual participant level. In what follows, we will propose, explore, and detail the predictions about memory that fall out of three different memory processes. Finally, we will present empirical work suggesting that both preschool-age children's and adults' color memory is best captured by a recall process that integrates prior knowledge and recall targets.

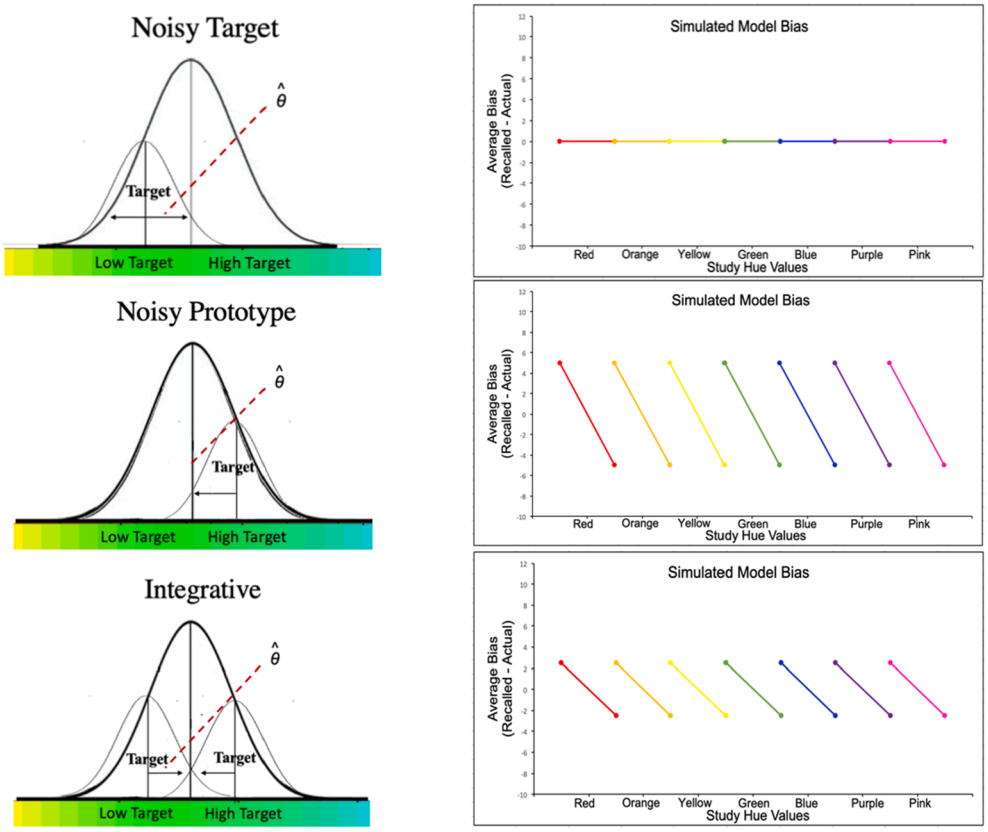

Fig. 1. Idealized predictions from each process. Θ is an estimate of a response under each strategy. Left Panels: The larger curves are distributions over the colors that belong to a given color category centered over the prototype of the category. The smaller curves are distributions centered over target values and denote the direction of where color values are likely to be recalled in response to the target as prescribed by the different models. Top left panel: prediction for response distribution under the Noisy Target process. Top right panel: idealized qualitative prediction of recall bias by category under the Noisy Target process. Middle left panel: prediction for response distribution under the Noisy Prototype process. Middle right panel: idealized qualitative prediction of recall bias by category under the Noisy Prototype process. Bottom left panel: prediction for response distribution under the Integrative process. Bottom right panel: idealized qualitative prediction of recall bias by category under the Integrative process.

1.1. Potential reconstructive memory processes in early childhood

In this section, we present a brief overview of three potential processes of retrieval from memory that young learners might use: The Noisy Target process, the Noisy Prototype process, and the Integrative process. We chose these three because they make clear predictions for how the developing memory system might behave. They also link to models that have been previously conceptualized in the adult literature to evaluate visual perception and working memory (e.g., Bae et al., 2015; Allred et al., 2015), categorical estimation (Huttenlocher et al., 2000), and episodic memory (Hemmer & Steyvers, 2009b) which are all relevant cognitive processes for the investigation of memory in young learners. These comparable models and representations have been employed to evaluate memory in adults, and thus make reasonable predictions for the type of information that might be reflected in young learners' recall of stimulus features.

1.1.1. Noisy target process

One possibility is that in early childhood, episodic content is simply noisy: children store what they can get and do their best with it (Zhang & Luck, 2008; Brady, Konkle, Gill, Oliva, & Alvarez, 2013). The Noisy Target process assumes that children store information in episodic memory as noisy traces of studied values (e.g., specific color hues). The Noisy target process is theoretically and conceptually similar to models that have been instrumental to our understanding of what happens to the fidelity of information over time and how fidelity changes between working and long-term memory in adults (e.g., Brady et al., 2013; Persaud & Hemmer, 2016). More specifically, this process is theoretically similar to the popular remember-guess paradigm of visual working memory (e.g., Zhang & Luck, 2008), where recall (or estimates) of a study value is thought to reflect continuous values and not category information. Relatedly, the Noisy Target process is conceptually similar to the 'fine-grain' component of the category model proffered by Huttenlocher et al. (1991) and the Metric only model (Bae et al., 2015; Allred et al., 2015), which assume that no category information is used to help recall studied features.

Under the Noisy Target process, the storage of studied values is not influenced by other information stored in memory, such as prior category knowledge and expectations for the token. In other words, young children might not rely on prior knowledge to fill in memory gaps, when episodic traces are noisy or incomplete. If this is the case, we should expect the noise (or error) in recall to be normally distributed around the true studied feature values, with no apparent bias toward a particular recall value (see Fig. 1, top left panel for illustration). If a regression line was fit to recall bias within a particular category, the slope of the regression line would be approximately zero (see Fig. 1, top right panel for illustration). This straight line or zero slope line does not imply that there is no noise in memory, but rather it would suggest that noise in memory is random or unsystematically distributed. The circumstances that might engender this process include situations where target information is stored with relatively high fidelity and the observer has very little difficulty or interference in retrieval of target values. Additionally, this process may also be revealed when associations to a prior category have not been established or have yet to develop over the course of the study. Perhaps children's responses, in particular, are likely to reflect this process if category knowledge is not well established or trusted.

1.1.2. Noisy prototype process

A second possibility is that younger learners' episodic recall process might be to categorize a study token as a prototype. In this case, memory in young children is simply a pointer to a prototype (in the same way that a verbal label might point to a category prototype (Donkin, Nosofsky, Gold & Shiffrin, 2014). That is, they have expectations for the statistical regularities of the environment, and new instances are simply stored as the most likely match to one of these learned regularities. The Noisy Prototype process assumes that information (e.g., a specific color value) is stored in episodic memory as categorical representations of studied features (e.g., mean of the category to which the color value belongs) and those representations are used as pointers to that type. Therefore, recall is not a noisy episodic trace of the studied value, but rather a categorical representation (or prototype) of the studied value (e.g., the mean of the category). The Noisy Prototype process is conceptually similar to the category level representation of stimulus features in categorization (Huttenlocher et al., 2000) where memories can be severely inexact and to increase accuracy, instead of encoding a fine grain feature, recall more closely reflects the category. Relatedly, it is reminiscent of the CATONLY (category only) model of visual working memory where study values are encoded in terms of the probabilistic category to which they belong and recall responses are thought to be simulated draws from a distribution over the category in which the study values belong (Bae et al., 2015; Allred et al., 2015).

Unlike the Noisy Target process, reliance on a prototype in this way will result in a strong apparent bias in recall where an observed hue value that is greater than the prototype will be systematically underestimated (because the prototype falls below the observation), and an observed value with a lower hue than the prototype will be systematically overestimated (because the category prototype falls above the observation). If a regression line was fit to recall error under this recall process, the slope of the line would be negative one (very steep) (see Fig. 1, middle panels for illustration), and there would not be a significant difference between the mean of the recalled hues and the category prototype mean.

The circumstance that might warrant a child adopting the Noisy Prototype process is when target information is stored with exceptionally low fidelity and/or the young observer has difficulty retrieving the initial target information. In this case, the observer relies on the prototype instead of the target information. For example, if a child observer passively witnessed an event ("Rachel ate a red apple") and was later asked about a feature of the event (e.g., the color of the apple), they might simply recall the most prototypical feature (category mean of red apples), especially if memory for the information is noisy and might not have garnered much attention during encoding. Given the limitations in memory for young learners, it is possible that their recall of study values more closely reflects category information as opposed to fine grain continuous feature values. Interestingly, Duffy et al. (2006) found that recall in younger school-aged children exhibited steeper regression to the mean patterns, relative to older children. This raises the possibility that younger children's recall performance might be best characterized by the Noisy Prototype process.

1.1.3. Integrative process

A third possibility is that young children, like adults, might optimally integrate noisy episodic content and their prior category knowledge (e.g., prototype representation of the category). That is, prior category knowledge and expectations for the statistical regularities of the environment are integrated with noisy episodic content to facilitate retrieval from memory. In this instance, the developing memory system makes an inference about a token in memory by attempting to resolve noise and uncertainty with the best stored categorical representation or prototype. This means that the more noise or uncertainty there is, the more the token will regress towards the expected category prototype (Hemmer & Steyvers, 2009a,b).

Under this process, prior category knowledge is used to fill in the gaps when episodic traces are noisy or incomplete. Thus, recall is a tradeoff between these two streams of information. When the memory representation for the target feature is strong, recall will closely resemble the studied feature value. Conversely, when the category representation is strong, and the memory representation is noisy, recall will more closely resemble the category representation. Unlike the Noisy prototype process in which new information will never outweigh the category expectation, an Integrative process affords reconciliation between learning and expectations from past experience. However, similar to the Noisy Prototype process, recall error would also reflect an apparent bias of over and under-estimation of values around the mean of the category (see Fig. 1, bottom panels for illustration). However, given that observers are integrating the episodic trace with the prototype representation, there would be less bias overall. If a regression line were fit to error, it would be negative, but significantly greater than negative one (less steep than the Noisy Prototype process).

The Integrative process can be instantiated with a simple rational model of memory (see Hemmer & Steyvers, 2009a,b; Persaud & Hemmer, 2014 for a similar approach). We provide details for this approach in Appendix A The integrative process used here was inspired by the color model used in Persaud & Hemmer, 2014, and is conceptually similar to Bae et al. (2015) and Allred et al. (2015)

CATMET: dual content model. The circumstance that may warrant the use of the Integrative process is when there is variability in the fidelity of the target information. Recall for target information that has high fidelity will closely reflect the target and information with low fidelity will closely reflect the category representation.

In this paper, we evaluate preschoolers' color knowledge and episodic memory for color. We then model the three processes that can potentially explain preschooler's recall performance. After, we compare preschoolers' knowledge and color recall to adults. We also explore individual differences in the fit of the three process models to preschoolers' and adult data. In the following section, we present the rationale for evaluating preschoolers' memory strategies in the domain of color.

1.2. Color knowledge

Color is an ideal domain for assessing prior expectations and memory in young learners for a number of reasons. For one, color is a domain in which humans have demonstrated that category knowledge plays a role in memory. For example, work exploring how environmental variations influence memory of color found that individuals from wildly different environments (e.g., Bolivia's indigenous Tsimané people and U.S. populations) form different expectations about color categories and use these expectations to aid recall. More specifically, despite differing color category expectations, performance from both Tsimané and US groups shows that recall regressed towards their respective color category means (Persaud, Hemmer, Kidd, & Piantadosi, 2017). This suggests that although color categories may vary across cultures, the cognitive system likely operates over them in the same way. Thus, color knowledge is an ideal domain to explore differences in memory across diverse groups and age populations.

Secondly, assessing prior expectations and memory for color in young learners is interesting because to some extent children know color, however, it might be less defined in comparison to adults. For example, from the time of birth, children (with typical vision) come into contact with and are surrounded by color. Thus, color cannot help but be experienced. There is also evidence to suggest that early on, children have established color category beliefs (Pitchford & Mullen, 2003). However, what remains unclear is how noisy children's color representations are and more importantly, if and how children use those potentially noisy representations to assist in recall. If it is the case that children's color category representations are less defined, this might differentially impact recall performance, relative to adults, in which case children's recall performance might be better explained by a different process.

Finally, color knowledge is a domain in which noise is straightforward to quantify and measure with respect to our proposed processes. In computing a bias score (recalled hue value minus studied hue value), we are able to not only quantify noise in memory but also its directionality (e.g., towards or away from the color category mean). This will allow us to capture whether learners are relying on color category knowledge to recall a specific hue value (e.g., regression towards the color category mean) and the extent to which the learner relies on category knowledge or episodic memory. Therefore, color provides an ideal domain to investigate competing models of memory in children and adults.

1.3. Approach: Empirically evaluating memory processes

The primary goal of this work is to explore how prior color category knowledge influences children's episodic memory for color. Specifically, we aim to experimentally quantify and compare color knowledge (e.g., color category means) in children and adults. Children and adults may have varying expectations for color that can, in turn, impact the processes used to aid recall. We evaluate color category knowledge by computing the mean of generated hue values for 7 color categories. Second, we investigate how prior knowledge (e.g., color category expectations) influences memory. Lastly, we evaluate three processes (the Noisy Target process, Noisy Prototype process, and Integrative process) that characterize how children retrieve information from memory. We compare this to adults. In what follows, we present the results of the color generation task.

2. Study 1: Color generation

The color generation task was used to determine the hue values that children and adults' associate with given color labels. In this task, participants were given specific color labels and were asked to 'generate' the hue value that best represented that label. We averaged the 'generated' hue values for each color category across participants[1] to obtain a mean hue value for both children and adults. We predicted a systematic agreement between subjects for the values corresponding to the labels centered on the core color categories. We take this as a measure of peoples' prior expectations for color.

[1] While more rigorous practices might access prior category expectations at the individual subject level for young children, taking multiple samples within subjects (e.g. asking the same child to produce multiple exemplars from the same color category) can lead to confusion for what the experimenter is asking of the child (e.g. Bonawitz, Shafto, Yu, Gonzalez, & Bridgers, 2020). Furthermore, there is evidence to suggest that children's color category knowledge by preschool age is developed (Pitchford & Mullen, 2003). Thus, we aggregate generated data and compare children's color expectations to that of adults.

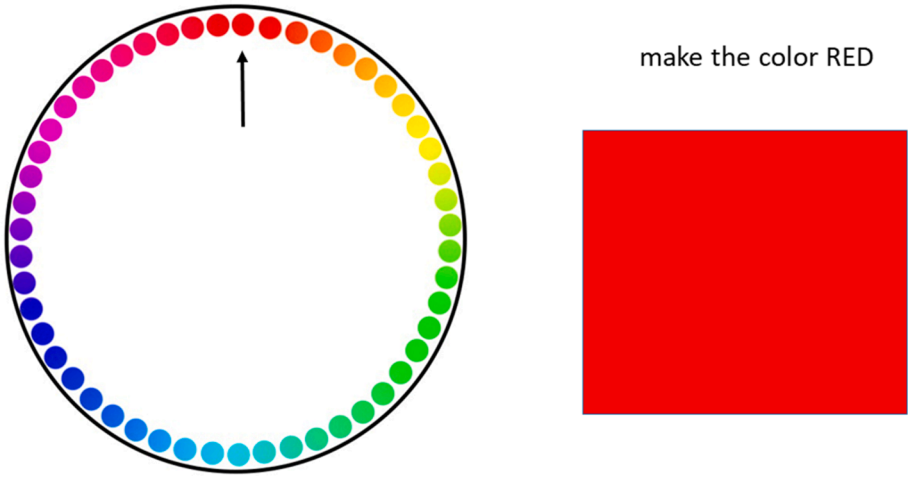

Fig. 2. A depiction of the display during the color generation study (Study 1a and 1b). All participants were presented with a partially occluded color wheel on the left of the screen and square on the right. Participants identified the best representative of a given color category by pointing to a color on the color wheel.

2.1. Study 1a: Preschooler color generation

2.1.1. Method

2.1.1.1. Participants. Thirty-five preschoolers (mean age: 59.78 months; range: 40.87–71.87 months) were recruited from local preschools in Newark, New Jersey. Participants were removed due to an experimental error (3) or failure to answer the check question (2) leaving a total of 30 children reported in the analyses. This study was approved by Rutgers University-Newark Institutional Review Board.

2.1.1.2. Materials. The stimuli consisted of 7 core color category labels (i.e., red, orange, yellow, green, blue, purple and pink), as well as "light blue" and "dark blue" labels serving as fillers. We chose these 7 color categories because these are primary color categories children learn and label in early childhood (Pitchford & Mullen, 2003). The color wheel mirrored that used in Persaud and Hemmer (2014), but was modified for use with preschoolers. In the Persaud and Hemmer study, adults were asked to use a mouse to select colors from a color wheel that was covered by a black mask. Here we used a sample of preschool age children whose motor skills are still developing making their accuracy and efficacy on point and click tasks less reliable when using a mouse (Hourcade, Bederson, Druin, & Guimbretiere, 2004). Instead of using a mouse to select colors, children were instructed to point to the color they remembered studying. Thus, to facilitate children's understanding of how to respond in the task, the color wheel was covered with a white mask that had roughly 50 (¼ inch) holes around its circumference (for a visual of the color wheel see Fig. 2). The colors within the holes were not perceptually different, but spacing was large enough so that it was clear which hole was being pointed to by the child. The mask was important for two reasons. First, it discretized the color wheel, which made it easier for children to identify and point to their recalled target color. Second, the white mask prevented the experimenter from misinterpreting the participants' recalled color. That is, having children point to a hole on the wheel when selecting colors minimized the ambiguity of their choice for the experimenter.

The color wheel sampled colors from the winHSL240 (hue, saturation, and luminance) color space. The colors varied in hue only from 0 to 239, while saturation and luminance were held constant at 100 and 50 units, respectively. Due to the mask covering the color wheel, the hue values that were visible to participants varied in increments of approximately 5 hue steps. Importantly, the holes on the wheel were placed such that the correct target colors could be selected, which would result in zero bias (i.e. no difference between the color studied and the color recalled). All stimuli were presented on a 15-inch Apple Macintosh display monitor with a vertical refresh rate of 60 Hz. The display was calibrated using X-Rite i1 Pro2 color calibration software.

2.1.1.3. Procedure. First, participants were verbally presented with a color label. Then they were instructed to generate the color value that best represented that color label by pointing to its ideal representative on a color wheel. One at a time, one of the 9 color labels appeared in the upper right corner of the computer screen in Georgia font. The color labels appeared in a random order and were read aloud by the experimenter. For example, children were asked, "Can you show me red? Use your finger and point to the color red". Each participant identified the ideal color category member by pointing to the "ideal" color on the color wheel. As the participant identified the ideal representative of a given color label, the experimenter moved a cursor over the color wheel to where the child was pointing, and selected the indicated color using a wireless mouse.

After the participant chose a color label representative, to verify that the chosen color was correct, the experimenter asked, "Is this the color you chose?". If the participant requested to modify their choice, the experimenter would prompt the participant to re-pick the

Table 1
Preschoolers' and Adults' Color Category Means and Standard Deviations.

	Child Mean (SD) hue units	Adult Mean (SD) hue units
Red	236.9 (4.56)	237.27 (2.94)
Orange	19.79 (7.29)	19.10 (4.47)
Yellow	39.89 (4.74)	39.06 (3.00)
Green	76.01 (11.49)	81.07 (6.69)
Blue	150.17 (17.98)	147.86 (9.78)
Purple	192.10 (11.75)	184.18 (5.14)
Pink	203.34 (5.77)	207.69 (7.26)

color that best represented the color label. Participants could generate the color they thought best corresponded to the given color label as many times as they wished. Once participants were satisfied with the color they generated, the experimenter moved on to the next trial by pressing the "spacebar". For each trial the color wheel randomly rotated 45 degrees. This random rotation controlled for a color location bias. Participants generated colors for 9 color labels once for a total of 9 trials. The task was self-paced and took on average 5 min to complete.

2.1.2. Results

To evaluate children's color categories, we computed the average chosen/generated hue value and standard deviation for 7 color categories: red, orange, yellow, green, blue, purple and pink.[2] We chose these 7 color categories because these are primary color categories children learn and label in early childhood. Table 1 shows children and adult color category means and standard deviations for these 7 color categories respectively.

2.2. Study 1b: Adult color generation

2.2.1. Method

2.2.1.1. Participants. Thirty-five undergraduate students at Rutgers University-Newark participated for course credit. All participants provided self-reports of normal color vision. This study was approved by Rutgers University-Newark Institutional Review Board.

2.2.1.2. Materials. The materials were identical to the stimuli used in Study 1a. Participants in Study 1b were presented with the same 7 core color categories, as well as "light blue" and "dark blue" labels. The color wheel mirrored that used in study 1a in that the color wheel was physically covered by a white mask and the color wheel sampled colors from the winHSL240 with a constant luminance and saturation of 50 and 100 units respectively. All stimuli were presented on a 15-inch Apple Macintosh display monitor with a vertical refresh rate of 60 Hz. The display was calibrated using the X-Rite i1 Pro2 color calibration software.

2.2.1.3. Procedure. Identical to the procedure used in Study 1a, adult participants were presented with 9 color labels, one at a time and were asked to generate the color hue value corresponding to that label by pointing to the ideal representative on a color wheel. Participants could modify their response as many times as they wished until they thought the generated color was the best representative for the color label. Once participants were satisfied with the generated color, the experimenter pressed the "spacebar" to continue to the next trial. Identical to Study 1a, for each trial the color wheel randomly rotated 45 degrees to control for a color location bias. Adult participants generated colors for 9 labels once, for a total of 9 trials, presented in random order. The task was self-paced and took on average 5 min to complete.

2.2.2. Results

To evaluate adult's color categories, we computed the average chosen/generated hue value and standard deviation for 7 color categories: red, orange, yellow, green, blue, purple and pink. Table 1 shows children and adult color category means and standard deviations for these 7 color categories respectively.

2.2.3. General results

To evaluate whether children's color categories differ from adults, we conducted a series of independent samples *t*-test to compare the adult and children generated hue values for each of the 7 color categories. Children's generated hue values were not significantly different from adults for all categories except for purple (bonferroni corrected with Bayes Factors[3] favoring the null, $p < .007$: red $t(63) = 0.38$, $p = .71$; $BF_{01} = 4.99$; orange $t(63) = -0.48$, $p = .63$; $BF_{01} = 5.07$; yellow $t(61) = -1.87$, $p = .07$; $BF_{01} = 1.23$; green $t(63) = 2.18$, $p = .03$; $BF_{01} = 0.35$; blue $t(51) = -0.61$, $p = .54$; $BF_{01} = 3.987$; purple $t(63) = -3.56$, $p = .0007$; $BF_{01} = 0.074$; pink $t(62) = 2.61$,

[2] We chose to only analyze data from the 7 main color categories and not the two filler colors (light blue and dark blue) as these were not used for Experiment 2 and we did not have data from subordinate labels for the other color categories.

[3] Bayes Factors were computed using the JASP Statistical Software (2020).

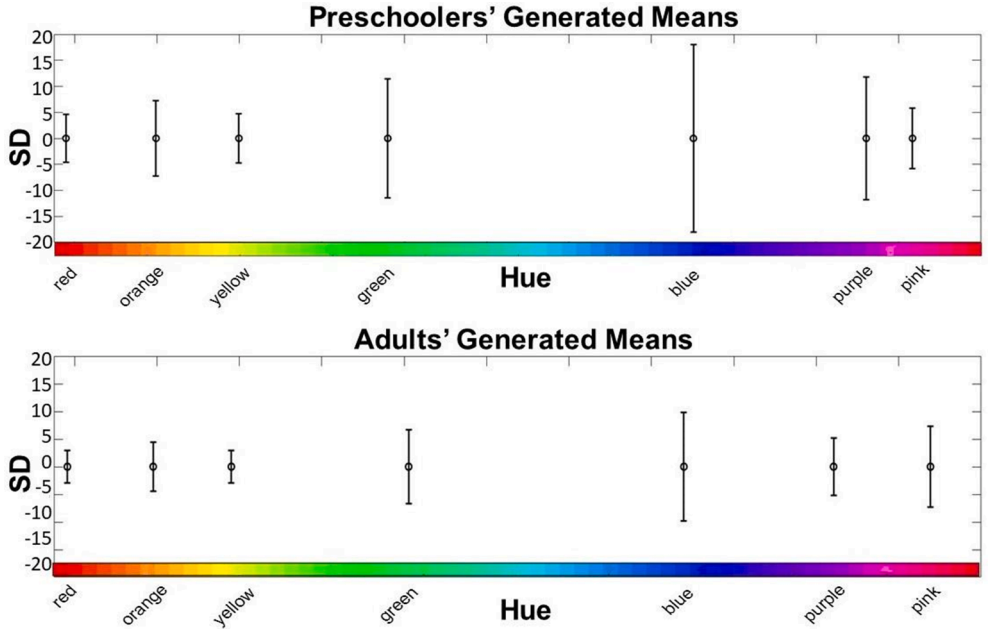

Fig. 3. ***Top panel:*** Children prior color category knowledge. Overall there are no significant differences between adult and children's prior color category knowledge except for the color category purple. The vertical lines depict the noise in the generated color values for each color category. Overall, there is more variance in children's color category knowledge in comparison to adults. ***Bottom panel:*** Adult prior color category knowledge. Each data point represents the mean hue value for each of the 7 color categories used in experiment 1. The color category for each data point is labeled by the color range on the x-axis. The lines depict the variance or noise found in each color category. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

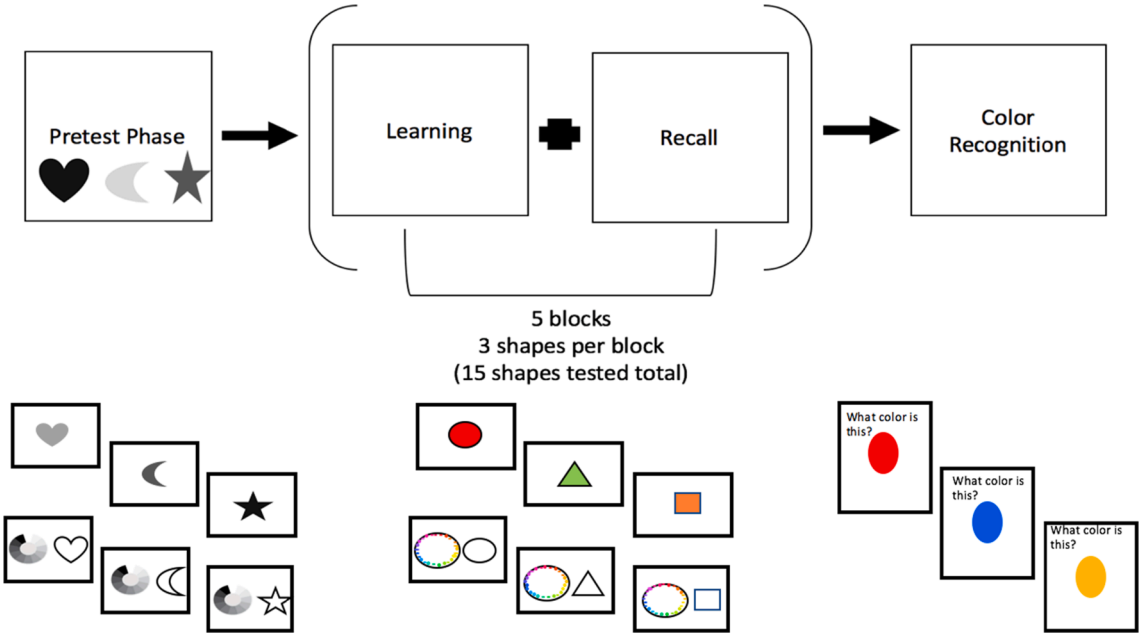

Fig. 4. (left to right) The procedural order for the preschoolers and adult color recall task experiment including example stimuli presented during each block. Pretest: Participants studied 3 grey-scale shapes (i.e., an almost white moon, a middle grey star, and an almost black heart). In a recall task, participants identified what shade of grey went with each shape. Learning + Recall: Immediately following the pretest participants studied and then recalled 15 unique color-shape paired items, divided across five blocks, such that each block contained three items. In the recall phase, the selected color filled the shape adjacent and to the right of the color wheel. Color recognition: Following the five experiment blocks, children were shown a color patch of each of the 15 tested colors, one at a time, and were asked to verbally label it.

$p = .01$; $BF_{01} = 0.83$).

A correlated samples *t*-test comparing the standard deviation of children and adult's generated color values for each of the 7 color categories revealed that there was a significant difference between the variance across the different categories between children and adults ($t(6) = -2.78, p = 0.03$). Although there was not much of a difference in the mean hue values adults and children generated for the color categories, children's generated values were more diffuse around the means. This might suggest that the boundary colors between the categories are still being fine-tuned in preschoolers. Also, the smaller deviation around the mean values generated by adults suggest a greater subjective agreement on the colors that belong within the categories, relative to children.

Taken together, these primary results suggest that children's color category expectations, at least for color prototypes (i.e. category means) are well established by the preschool years. Furthermore, they provide a basis to explore our primary question of interest—whether category knowledge influences memory for color (see Fig. 3 for a visual comparison of children and adults generated color category values).

3. Study 2: Color memory

The goal of the color memory study was to evaluate the impact of color category knowledge on episodic recall for preschoolers. To assess color memory, we used a cued recall task where children studied and recalled 15 unique color-shape pairs. Given the nature of the episodic memory task we used an experimental methodology that was sensitive to children' needs. More specifically, we considered the limitations in preschoolers' working memory capacity and feature binding abilities (e.g., binding a color with a shape), given that relational memory binding at this age is still maturing (Sluzenski et al., 2006). We piloted a similar task and found that children were able to remember 3 object-color pairs at a time. Additionally, we chose to present preschoolers with 3 object-color pairs at a time to avoid floor effects (see Hitch et al., 1989 for similar experimental approach). Fig. 4 provides a procedural schematic of the color memory task.

3.1. Study 2a: Preschooler color memory

3.1.1. Method

3.1.1.1. Participants. Thirty-nine preschoolers (mean age: 54.17 months; range 43.43–72.79 months) were recruited from local preschools in Newark, New Jersey. Each participant completed a pretest training task, learning and recall task, and a color labeling task. Children in the Color Memory experiment (Study 2) did not participate in the Color Generation experiment (Study 1). Six participants were dropped because (1) they did not speak English nor understand the procedure, (3) they failed the color labeling task by mislabeling 3 or more target colors, (1) they failed pretest criteria (see below), or (1) experimental error, leaving a total of 33 children. This study was approved by Rutgers University-Newark Institutional Review Board.

3.1.1.2. Materials. Participants were presented with 15 unique color-shape pairs. The 15 target colors were hue values that were ±1 SD deviation[4] of the average hue value for each of the 7 categories[5] tested in Study 1a. In this way, each color category had two target colors; a high and low hue value, relative to the mean of the category. The 15th color was a filler color (from a middle blue range). The colors varied in hue only, while saturation and luminance were held constant[6] at 100 and 50 units, respectively. The HSL values were then converted into RGB values for presentation in Matlab via a Matlab color conversion algorithm (Bychkovsky, 2020 – hsl2rgb and rgb2hsl conversion (https://www.mathworks.com/matlabcentral). All participants studied the same color-shape pairs presented in 5

[4] The rationale for using ±1S.D. was three fold. First, as with developmental studies, procedures must be truncated to keep children engaged, and did not have the flexibility to run 100 trials as would be necessary to get power estimates across a continuous range. Second, we needed color values that were far enough away from the mean to be perceptually distinguishable for young viewers. Third, by choosing larger standard deviations we ran the risk of ending up in a new category and we wanted to make sure that we were not pushing up against/into neighboring categories, which would occur with categories with narrower hue ranges such as yellow, pink, and purple. Finally, we chose standard deviation because by using a proportional value, we are in some sense equating the distance of the categories, allowing for more equitable comparisons across categories (though this was not the primary focus of the result).

[5] Preschooler and adult color category means did not significantly differ from one another therefore using the generated hue values and standard deviations from Study 1a to create the 15 target hue values used in study 2a and 2b is justifiable. We also employed the larger standard deviations from the children's results in Study 1 to ensure both child and adult participants would be able to discriminate between the target items.

[6] While we did calibrate the monitors, there is some possibility that luminance slightly varied across colors, despite holding it constant for the intended colors (Bae, Olkonnen, Allred, Wilson, & Flombaum, 2014), as we did not conduct post hoc measurements to test the luminance of each color. However, it is unlikely that luminance differences confounds or influences our results in an impactful way. First, the observed regression to the mean effect is a standard finding in the memory literature across stimuli domains (e.g. size, height, spatial location, etc.) and despite color memory studies varying in color space, monitor calibration, and post-hoc measurements of chromaticity and luminance, the regression pattern that results from prior category knowledge still holds (Bae et al., 2014; Persaud & Hemmer, 2014; Persaud et al., 2017). Also, it is not clear why potential shifts in luminance would lead to a systematic bias in the particular direction of the category hue mean, especially when regression was observed across all categories for both adults and children. Furthermore, our results do not hinge on the assumption that the precision of all colors are comparable. In fact, our models take into account the fact that certain color categories might have inherent differences in noise. Nevertheless, the impact of fluctuations in luminance on regression to the mean is an important empirical question that future work must explore.

pseudo randomized orders, such that participants in group 1 studied one order, participants in group 2 studied a different order, and so on. All stimuli were presented to participants on a 15-inch Apple Macintosh display monitor with a vertical refresh rate of 60 Hz. The display was calibrated using the X-Rite i1 Pro2 color calibration software.

3.1.1.3. Procedure

3.1.1.3.1. Pretest. Participants were first introduced to a pretest phase in which they were shown 3 grey-scale shapes (i.e., a moon that was almost white, a star in a middle grey range, and a heart that was almost black). In a fixed order (i.e., heart, moon and star) all 3 shapes were shown twice, one at a time for 3 s each without a delay. Participants were instructed to pay attention to the different shades of grey that accompanied each shape. Then, in a recall task, participants were prompted to identify what shade went with each shape. To minimize cognitive load and ensure that children understood the task, the shapes were shown in the same order during test as they were presented during study. The pretest phase served as a way to familiarize participants with the subsequent task procedure and to provide a measure to exclude children who had trouble understanding the procedure. Children who were unable to successfully complete the shape and grayscale matching task (e.g., pointing to more than one color) were dropped from analysis (1 child was dropped).

3.1.1.3.2. Test. The main experimental blocks immediately followed the pretest. Participants were prompted to study and then recall 15 unique color-shape paired items, divided across five blocks, such that each block contained three items. In each block there was a learning phase and a recall phase. In the learning phase, three color-shape items were shown one at a time. After the last item was shown, all three were shown again, one at a time, in the same fixed order. That is, each three color-shape paired items within a block was repeated a second time such that participants had two exposures to the study stimuli. Color-shape pairs were shown twice to ensure that they had time to attend to, encode and then bind the featural property (i.e., color) to the given item (i.e., shape). Each shape was shown for a duration of 3 s and accompanied by attention directing speech by the experimenter (i.e., labeling of the shape). There was no delay between the shape presentations. Neither was there a delay between the main experimental blocks.

Immediately following the learning phase children completed a cued recall task where they were shown black and white versions of each of the 3 color-shape items (see Fig. 4) one at a time and were asked to identify which color paired with the displayed shape. The order of the shape presentation in the learning and recall phase were identical. Participants identified the shape and color pairing by pointing to the color that was paired with the displayed shape on a color wheel. The wheel was masked with a cutout that allowed participants to select a hue from among 50 (¼ inch) holes along the circumference. As the participant identified the shape and color pairing by pointing to a color on the color wheel, the experimenter selected it using a wireless mouse. To verify that the chosen color was correct, the experimenter asked, "Is this the color you chose?" If the participant did not agree with the selected color, the experimenter prompted the participant to select the color that went with the presented shape until the child said it was accurate.

3.1.1.3.3. Color labeling task. Following the five experiment blocks, children completed a color labeling task. In the color labeling task participants were shown a color patch of each of the 15 tested colors, one at a time, and asked to verbally label it. For example, the experimenter would say, can you tell me what color this is? The color labeling task not only ensured that children were familiar with the 7 core color categories presented to them during the learning phase, but it provided insight into the potential miscategorizations (e.g., red mistaken for orange) that may have impacted memory outcomes. The labeling task was simply to ensure that this group of children, like those in Study 1, were knowledgeable of the color categories they studied (this was confirmed during analysis). The verbal responses were recorded via a video camera and notes taken by the experimenter.

3.1.2. Results

The goal of the main analyses was two-fold: first to determine if preschoolers' bias (or error) in recall was systematic (i.e. influenced by category knowledge), and to evaluate which process (Noisy Target, Noisy Prototype, or Integrative) best captured the pattern of bias. We first confirmed that the children in this sample had clear category knowledge based on performance in the labeling task (which occurred after the final memory trial). The labeling results revealed that children correctly labeled 86% of color trials, and incorrectly labeled 14% of color trials. Importantly, of those incorrectly labeled trials, 13% were labeled with a neighboring color category (e.g. labeled high yellow, green or labeled low pink, purple), and only 1% was completely mislabeled (e.g. orange color labeled blue). Thus, we conclude that the children in the memory task were knowledgeable of the color categories they studied.

Prior to our main analysis of the memory data, we excluded responses that fell outside of the mean of the neighboring color categories to the target, as this indicated a misremembering of shape-to-color match. This resulted in the exclusion of 40% of trial data[7] indicating that preschoolers had some difficulty with remembering all items; however, 60% success rate is also significantly greater than chance responding ($p < .0001$, by 2-tailed binomial test, where chance = 3/7) suggesting that although challenging, preschoolers were able to remember the approximate color for a substantial number of the shapes. This inclusion criteria allowed us to get a more accurate measure of response error (response hue - target hue) for remembered study values within each color category. We then fit linear regression models to the averaged bias scores for each category. The methodological tests were chosen to specifically address the questions of interest. As this is a rich data set the number of analyses could be substantial. Instead we conduct tests to specifically address the questions of interests.

[7] Given the amount of data that were excluded from the study, we have also analyzed these excluded trials as well as implemented the three models to characterize the process that gave rise to the dropped data. The analyses revealed that these data were extremely noisy and mirrored that of a guessing process. We include details of the modeling of the dropped trials in the supplemental materials.

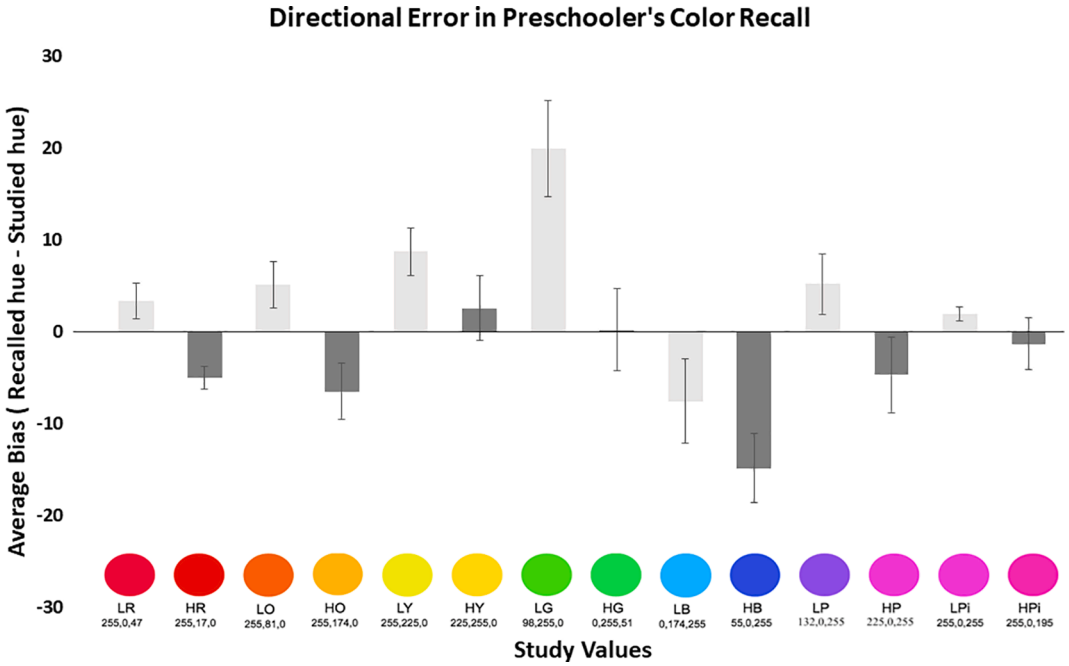

Fig. 5. Children's mean bias scores for all low and high target hue values. Below the respective mean bias score is a display of test materials (RGB values) used in Experiment 2a: Child recall. The L and H depict whether a color is categorized as low (L) or high (H) hue values. The abbreviations for each of the target values are as follows: low red (LR), high red (HR), low orange (LO), high orange (HO), low yellow (LO), high yellow (HY), low green (LG), high green (HG), low blue (LB), high blue (HB), low purple (LP), high purple (HP), low pink (LPI) and high pink (HPI). The numbers depicted below each color swatch is the RGB value that accompanies that target color. These colors may not reprint accurately. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Slopes and X-intercepts for Preschoolers and Adults by Color Category.

	Child Slopes	Child Mean (SD)	Adult Slopes	Adult Mean (SD)
Mean	−0.60		−0.46	
Red	−0.90	236.31	−0.92	232.87
Orange	−0.81	19.42	−0.22	23.32
Yellow	−0.38	48.19	−0.44	39.90
Green	−0.84	91.48	−0.47	81.15
Blue	−0.19	129.16	−0.40	142.39
Purple	−0.59	188.94	−0.39	193.19
Pink	−0.52	203.84	−0.40	202.65

Note. Although it is mathematically intuitive to interpret the y intercept, here we chose to display the x intercept as it provides meaningful information about where the inferred category means for each color set exist. Specifically, the point at which the line crosses the x axis (at y = 0) represents the point at which no regression to the category mean is observed. As expected, these cross-over points also align with the category means generated by participants in Experiment 1. We hold the preschooler and adult slopes constant (e.g., mean) because this allowed us to see the difference in intercepts across color categories (Hemmer & Steyvers, 2009a). Children and adult negative slopes suggest that for both groups recall is regressing towards the mean of the category.

3.1.2.1. Bias analysis. An initial indicator of the processes used to reconstruct events from memory can be evidenced by the recall error observed in the task. Recall that the Noisy target process predicts that error would be unsystematic across high and low study values, and therefore, not significantly different from zero. In contrast, the Noisy Prototype and Integrative processes predict that high study values will be underestimated (error less than zero) and low study values will be overestimated (error greater than zero). Thus, to determine if response error was systematic across color categories, we first evaluated whether bias for low hue values differed from zero and then if bias for high values differed from zero (Fig. 5). One-sample *t*-tests revealed that both bias for low values differed from zero ($t(142) = 19.76$, $p < .0001$, $d = 2.34$) and bias from high values differed from zero ($t(133) = 15.10$, $p < .0001$, $d = 1.84$). These differences in bias suggest that error to some degree was systematic. Taken together, this suggests that young learners are systematically using prior category knowledge to recall specific hue values.

3.1.2.2. Regression analysis. Two linear regression models were fit to the mean error scores for each of the 7 color categories in order

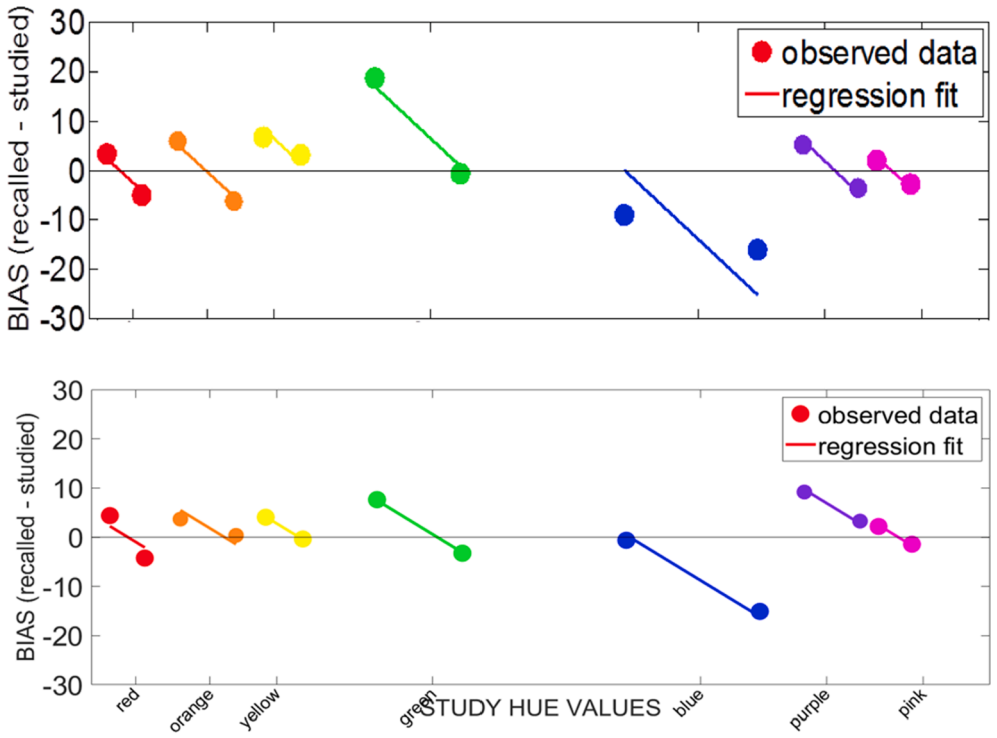

Fig. 6. Preschooler's recall bias by color category. Positive bias indicates over estimation and negative bias indicates underestimation. The black line indicates no bias. The data points are color coded with the hue for that color range and the corresponding labels are given on the x-axis. The lines give the regression fits for each preferred color label. The regression fit assumed a fixed slope (averaged across all 7 categories) and separate intercepts for each color category to assess differences in the intercepts across categories. This, in part, might explain the poorer fit to the child data (especially in the color category blue). **Bottom panel**: Adult recall bias by color category. Positive bias indicates over estimation and negative bias indicates underestimation. The black line indicates no bias. The data points are color coded with the hue for that color range and the corresponding labels are given on the x-axis. The lines give the regression fits for each preferred color label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to evaluate the degree to which participants' responses regressed toward the category mean and also to confirm that responses regressed to different prototypes for each category. Since the Noisy Target process predicts unsystematic error, the slope of regression lines fits to error for any category would not be significantly different from zero. In contrast, both the Noisy Prototype and Integrative models predict regression, thus the slopes would be different from zero with different intercepts for each color category. The first linear regression allowed the slopes to vary to evaluate the degree of regression in each category (see Table 2). The second regression fit assumed a fixed slope (averaged across all 7 categories) and separate intercepts for each color category (see Fig. 6 top panel). Fixing the slope in the second regression model allowed us to assess differences in the intercepts across categories. This analytical decision is identical to previous studies evaluating the effects of prior knowledge on reconstruction memory (Hemmer & Steyvers, 2009a,b). Category slopes, the fixed slope, and intercepts are reported in Table 2. A one-sample *t*-test revealed a significant difference between category slopes and 0 ($t(6) = -6.04$, $p < .0001$, $d = -2.28$), suggesting that participants were indeed regressing in recall. The negative slopes for each category further suggest that regression was directed toward category prototypes. A one-sample *t*-test also showed a significant difference in intercepts and 0 ($t(6) = 3.16$, $p = .02$, $d = 1.19$). The different intercepts for each of the color categories indicated regression toward a different mean hue value for each of the color categories.

3.1.2.3. Statistical evaluations of the three processes. Results from the bias analysis as well as the linear regression fits shed light on whether or not participants have adopted the three processes. To reiterate, if individuals were using the Noisy Target process, error would be unsystematic around the mean of the categories. However, the structured bias (i.e., regression pattern) was systematic across all categories, which qualitatively rules out the Noisy Target process as an explanation of the data. This leaves the Noisy Prototype and Integrative processes. According to the Noisy Prototype process, when individuals recall an observed target hue value, they rely solely on the mean hue value of the target hue's color category.

In other words, under the Noisy Prototype process individuals neglect to reference the studied hue value (episodic content) and use their prior category knowledge to estimate the studied value. To test the strict employment of this type of process in the aggregated data, we first calculated the difference between the chosen hue values and the prior means of the respective categories (learned from Study 1b) and then we compared those difference scores to zero. If preschoolers were solely using prototypes (i.e., category knowledge), there would be no difference between the calculated difference scores and zero. However, a one-sample *t*-test revealed that the

difference scores (participant responses for each hue - prototype mean of that hue) were significantly different from zero ($t(276) = -4.77, p < .0001, d = -0.29$), which suggests that in the aggregate, children were not solely relying on prototypes to recall hue values. An additional prediction of the Prototype process is that the slopes of the regression line fit to the data should not be significantly different from −1. However, a one-sample *t*-test revealed a significant difference between the slopes of the regression lines fit to the category data and −1 ($t(6) = 3.99, p < .01, d = 1.51$). Taken together these results provide strong evidence that in the aggregate, all preschoolers are not adopting the Noisy Prototype process.

3.1.2.4. Computational evaluations of three processes. We implemented each of the three models, as detailed in the Appendix A. To directly compare our three processes based on the aggregated data from children, we computed log-likelihood scores under each model.[8] The Integrative model better fit the data (log likelihood: −1169.36) than either the Noisy Target (−1183.79) or Noisy Prototype model (−1303.44). Comparing Bayes Factors of the two best fitting model types (Noisy Target and Integrative) revealed "very strong'' evidence in favor of the Integrative model ($BO = 1.85 \times 10^6$; Log Ratio = 0.99). These results suggest that a model assuming an integrative approach where preschoolers combined their knowledge of the color categories with noisy episodic traces best explains their aggregated recall data.

3.2. Study 2b: Adult memory for color

In the following study, we assessed free recall for color in adults in order to explore whether there are developmental shifts in the degree to which prior category knowledge influences reconstructive memory (i.e. comparison in performance between young children and adults). Identical to Study 2a, participants viewed 15 distinct shape-color pairings. In test trials, participants were asked to remember the colors of the studied shapes using a computerized color wheel to indicate responses (as in Study 1b). Participants were also asked to provide a color label for the studied shapes (as a methodological check). We assess free recall (and model the data) in adults to first confirm that category influences on memory are observable within this experimental paradigm, given that it's well known that adults use category knowledge to aid recall. Using adults as a comparison, we explore if the integration of information is a process that extends across developmental groups.

3.2.1. Method

3.2.1.1. Participants. Thirty-four Psychology undergraduate students at Rutgers University-Newark participated for course credit. All participants provided self-reports of normal color vision. These participants were not involved in Study 1b. This study was approved by Rutgers University-Newark Institutional Review Board.

3.2.1.2. Materials. Identical to the stimuli used in Study 2a, adult participants were presented with 15 distinct hue and shape pairs. The 15 target colors used in each shape were values that were ±1 SD deviation of the average color for each of the 7 color categories tested in Preschooler's Color Memory Study 1a). Participants studied each shape and color only once. Stimuli were presented on the same calibrated monitors used in Study 1b.

3.2.1.3. Procedure. The procedure in this study was identical to the procedure used in study 2a.

3.2.1.3.1. Pretest. Participants were first presented with a sequence of 3 shape and grayscale pairings (i.e., a moon that was almost white, a star in a middle grey range, and a heart that was almost black) twice in a fixed order and then, one by one, asked to recall which shade of grey went with each shape. This allowed participants to get acclimated to the testing procedure.

3.2.1.3.2. Test. Immediately following the pretest phase adult participants were introduced to the 15 shape and color pairings in blocks of 3 at a time. The shape and color pairings in each block were shown once and then again, immediately after in a fixed order. After the block study presentation, participants completed a recall task in which they identified the 3 color and shape pairings. Each participant completed 5 experimental blocks.

3.2.1.3.3. Color labeling task. Following the five experimental blocks adult participants verbally labeled all 15 target colors (including middle blue).

3.2.2. Results

To evaluate performance in the memory task, we first calculated recall bias as the difference between the recalled hue values and the studied values. Response values that fell outside of the mean of neighboring categories were excluded, as this indicated a misremembering of shape-to-color pairs. This resulted in the removal of 9.8% of memory trials. After, we fit a linear regression model to recall bias in each of the seven categories.

[8] The Noisy Prototype model assumes that the noise in memory comes from Study 1. Therefore, this model does not have a free parameter like the other two models. To address this, we implemented an alternate version of the prototype model where we still assume a distribution centered on the category means, but we search the space for the best noise value (sigma = 18; loglik = −1187.50). This iteration still loses to both the Noisy Target and Integrative models.

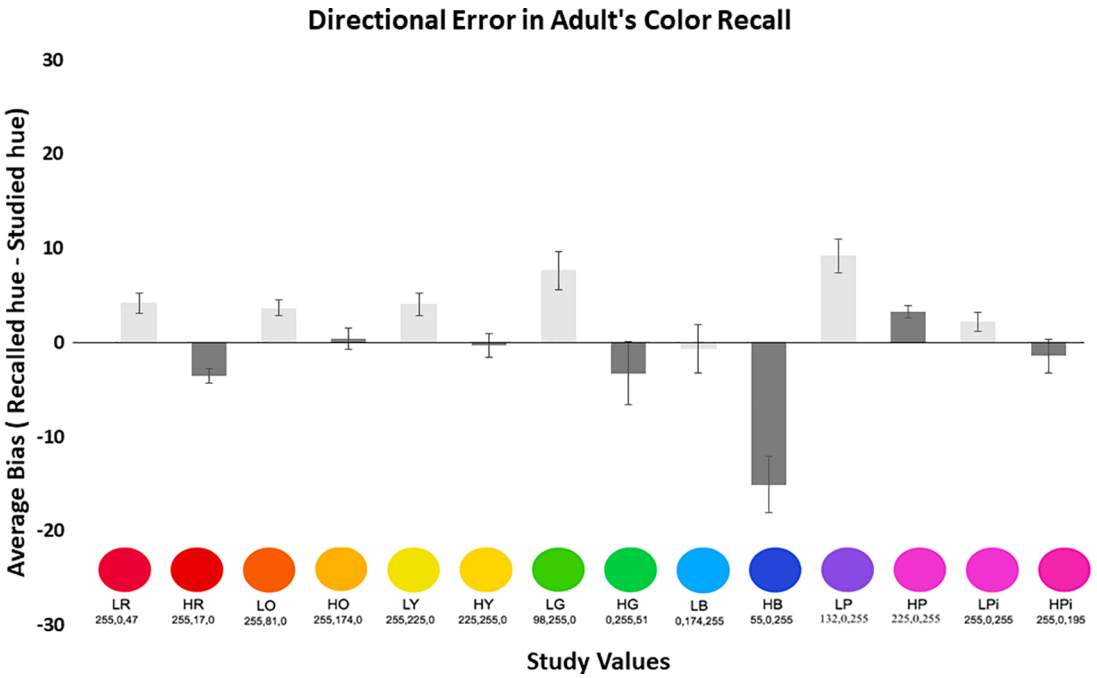

Fig. 7. Mean bias scores for all low and high target hue values. Below the respective mean bias score is a display of test materials (RGB values) used in Study 2b: Adult recall. The L and H depict whether a color is categorized as low (L) or high (H) hue values. The abbreviations for each of the target values are as follows: low red (LR), high red (HR), low orange (LO), high orange (HO), low yellow (LO), high yellow (HY), low green (LG), high green (HG), low blue (LB), high blue (HB), low purple (LP), high purple (HP), low pink (LPI) and high pink (HPI). The numbers depicted below each color swatch is the RGB value that accompanies that target color. These colors may not reprint accurately. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.2.1. Bias analysis. Similar to the analysis employed for the preschoolers' results, we tested whether bias scores differed from 0 and then if bias scores for high and low target values differed within each color category (see Fig. 7). One-sample *t*-tests revealed that bias scores for both low values ($t(215) = 6.54$), $p < .0001$, $d = 0.63$) and high values ($t(212) = -3.63$, $p < .0001$, $d = 0.35$) significantly differed from zero. These differences in bias suggest that error to some degree was systematic. A series of paired samples t-tests revealed a significant difference in bias scores between high study values and low study values for every color category (red $t(28) = 7.51$, $p < .0001$, $d = 1.6$; orange $t(24) = 2.66$, $p = .01$, $d = 0.61$; yellow $t(31) = 2.10$, $p = .04$., $d = 0.62$; green $t(27) = 4.04$, $p < .0001$, $d = 0.69$; blue $t(29) = 3.63$, $p = .001$. $d = 0.92$; purple $t(25) = 4.80$, $p < .0001$. $d = 0.81$; pink: $t(23) = 2.81$, $p = .01$., $d = 0.42$). Taken together, the bias analyses revealed some regression toward the category prototype for high and low study values. For six of the seven color categories, it appeared that participants overestimated values below the mean of each category and underestimated the values above the mean of each color hue range.

3.2.2.2. Regression analysis. Two linear regression models were fit to the mean bias scores for each of the 7 color categories. The first linear regression allowed the slopes to vary to evaluate the degree of regression in each category (see Table 2). The second regression fit assumed a fixed slope (averaged across all 7 categories) and separate intercepts for color category (see Fig. 6 bottom panel). Fixing the slope in the second regression model allowed us to assess differences in the intercepts across categories. This analytical decision is identical to previous studies evaluating the effects of prior knowledge on reconstruction memory (Hemmer & Steyvers, 2009a,b). Category slopes, the fixed slope, and intercepts are reported in Table 2.

A one-sample *t*-test revealed a significant difference between category slopes and 0 ($t(6) = -5.69$, $p = .0013$, $d = -2.15$), suggesting that participants were indeed regressing in recall. The negative slopes for each category further suggest that regression was directed toward category prototypes. A one-sample *t*-test also showed a significant difference in intercepts from 0 ($t(6) = 3.10$, $p = .02$, $d = 1.17$). The different intercepts for each of the color categories indicated regression toward different mean hue values for each of the color categories.

3.2.2.3. Statistical evaluations of the three processes. The overall finding that error was different from zero qualitatively or statistically rules out the Noisy Target model. To statistically evaluate the Noisy Prototype model, we computed the difference between response scores and the category mean of each corresponding category and evaluated whether those scores differed from 0. A one-sample *t*-test revealed a significant difference between the response – category difference scores collapsed across all 7 color categories and zero ($t(428) = -8.31$, $p < .0001$, $d = 0.57$). Additionally, a one-sample *t*-test revealed a significant difference between the slopes of the regression lines and −1 ($t(6) = 6.56$, $p < .0001$, $d = 2.48$). Taken together, these results suggest that adults, like young learners, were

not using the prototypes alone to recall hue values.

3.2.2.4. Computational evaluations of three processes. To quantitatively compare the Noisy Target, Noisy Prototype, and Integrative processes, we computed log-likelihood scores—calculating the probability of observing the participant responses under each model (See Appendix A for model analysis details). Note that for the Noisy Target and Integrative processes, the only free parameter is the degree of variance in the memory distributions. Thus, we searched the space to find the variance measure that maximizes the log likelihoods for each model.[9] The Integrative model fit the data better (log likelihood: −1593.37) than either the Noisy Target (−1660.00) or Noisy Prototype model (−2080.73). Comparing Bayes Factors based on the log-likelihoods of the two best models (Noisy Target and Integrative model) revealed very strong evidence in favor of the Integrative model ($BF = 8.65 \times 10^{28}$; Log Ratio = 0.96; see Kass & Raftery, 1995). Taken together with the regression analysis and quantitative comparisons of the models, the results suggest that overall, adult participants were using an approach that combines their noisy episodic content with prior knowledge of the color categories to reconstruct studied hue values.

Another way to evaluate whether the Integrative process is capturing some meaningful aspect of cognitive representation is to search for the best fitting parameters on the prior distributions, given participant recall in Study 2. Indeed, a search for the noise values that maximize the log likelihoods of the data fit to each color category revealed a strong correlation between the best fitting standard deviations on the prior distributions and the standard deviations observed from participants in the prior knowledge task presented in Study 1. This correlation pattern was observed for both the children and adults (Pearson: children $r = 0.89$, $p = 0.0067$; adults $r = 0.88$, $p = 0.0086$). This result provides additional support for the Integrative process, demonstrating that the prior distribution terms were not arbitrarily assigned, but rather meaningfully capture human categories.

3.2.3. Comparison of adult and child regression fits

Our main interest in differences between children and adult performance was in the use of model strategy. Initial evidence for a difference between the two groups would be evident in either differences in regression slopes, which could indicate a difference in the type of strategy employed (e.g. prototype model leads to steeper regression than integrative) or if groups were using vastly different category prototypes to regress toward. This might indicate a different use or knowledge of the color categories evaluated in the study. To test these possibilities, we conducted paired samples t-tests comparing children slopes to adult slopes and children intercepts to adult intercepts. The t-tests revealed no significant difference in slopes between the two developmental groups ($t(6) = -1.334$, p = .231, d = −0.504) and no significant difference in intercepts ($t(6) = -0.479$, $p < .649$, d = −0.181).

3.3. Individual differences

The results of the analysis and model fitting highlight an important role that category knowledge plays in episodic memory at early development (i.e., preschool age). In this section, we perform a critical in-depth analysis of children's recall data to tease apart underlying individual and group-related differences in the reconstructive process. Exploring individual and age-related differences is motivated by the finding that not only do children rely on category knowledge, but also that memory in preschoolers exhibited steeper regression to the mean patterns, relative to adults. It could be the case that at the individual subject level, children might differ in the best fitting model, such that those with steeper regression might be better fit by the Noisy Prototype model, while less steep regression might be better captured by an Integrative model. Moreover, the Noisy Target model might better fit children who do not show regression patterns and recall behavior is very noisy.

Furthermore, recall performance in children might not only differ at the individual subject level, but also at the individual trial level, especially if contextual strategies, such as spontaneously labeling study features, are employed to facilitate recall performance. For example, while running this study with the preschoolers, we observed that participants spontaneously labeled the colors, as they studied them and/or as they recalled them. For example, one preschooler (age = 4.64 years), stated, "Purple, purple, purple. I got this!", while studying a purple hue value. Counterintuitively, while labeling may boost the learner's ability to remember that an item was observed from a particular category, it may also lead to noisier storage of specific stimuli that deviate from category means, because the label provides a cheaper (albeit potentially less accurate) compression option than storing the details of the original. In this way, this individuating behavior of labeling might impact their reconstruction of events in memory at either the individual subject or trial level. Research suggests that labeling can influence recall of continuous color values, such that labeling results in information being lost gradually as opposed to suddenly (see, Donkin et al. (2015) for discussion on the role of labeling, sudden death, and gradual decay in memory). Similarly, reliance on visual and verbal information in memory shifts with development: younger children rely more heavily on visual information while older children rely on both (Hitch et al., 1989). It is theorized that while visual inputs access the visual component of memory, spoken inputs might access the phonological component of memory, resulting in a difference in behavior (Hitch et al., 1989). This provides further reason to suspect that there might be a difference in the best fitting models for children who spontaneously label colors or for specific trials where colors are labeled. Therefore, the goal of this follow-up analysis was to evaluate whether young children and adults, at the individual subject level, employ different processes to recall episodic events and whether the behavior of spontaneously labeling observed in preschoolers was better fit by a particular model.

[9] Like with preschooler's data, we implemented an alternate Noisy Prototype model where we searched the space for the best noise value (sigma = 12; loglik = −1690.53). This iteration of the model also loses to both the Noisy Target and Integrative models.

Table 3
Frequency of Model Fits for Preschoolers and Adults.

Model	Count (%)	
	Children	Adults
Integrative	11 (33.3%)	27 (79.41%)
Noisy Target	7 (21.2%)	3 (8.82%)
Noisy Prototype	15 (45.5%)	15 (45.5%)

Table 4
Frequency of Model Fits based on Labeled and Unlabeled Trials.

Model	Count (%)	
	Labeled	Un-labeled
Integrative	3 (0.10%)	9 (27.27%)
Noisy Target	4 (0.12%)	7 (21.21%)
Noisy Prototype	26 (78.78%)	17 (51.51%)

For this analysis, we fit the Noisy Target, Noisy Prototype, and Integrative models to each individual's data.[10] We then evaluated the log likelihood scores of the model fits to determine which account explained behavior for the greater proportion of children and adults. For the Noisy Target and the Integrative models, we assumed that the noise parameter was equal to the best fitting noise values learned from the model fits to the aggregated data for each age group (adults: Integrative = 4, Noisy Target = 12. Preschoolers: Integrative = 6, Noisy Target = 18). After, we assessed whether labeling behavior affected the proportion of children fit by each of the models.

3.3.1. *Individual differences model fitting results*

3.3.1.1. Adults. The integrative model was not only the best fitting model for the aggregated data, but it also captured the greatest proportion of data at the individual subject level (see Table 3). In other words, of the 34 adult participants, the greatest proportion was better fit by the Integrative model (n = 27/34), followed by the Noisy Prototype (n = 4/34), and then the Noisy Target (n = 3/34). This composition of model fits was consistent with previous research showing that adults integrate prior category knowledge and noisy episodic content when reconstructing events from memory (Hemmer & Steyvers, 2009a,b; Persaud & Hemmer, 2016).

3.3.1.2. Preschoolers. Although the Integrative model was the best fitting model at the aggregate data level, at the individual level a greater proportion of children were better fit by the Noisy Prototype model (n = 15/33), followed by the Integrative model (n = 11/33), and then the Noisy Target model (n = 7/33) (see Table 3).

3.3.1.3. Age group comparison. To evaluate whether the proportion of participants best fit by each of the three models was dependent upon age group, we used the Freeman-Halton extension of the Fisher's Exact test to compute the (two-tailed) probability of obtaining a distribution of values in a 2 (children vs adults) × 3 (Integrative vs Noisy Target vs Noisy Prototype) contingency table, given the number of observations in each cell. The results revealed that the observed proportion of best fitting models was dependent on age ($p < .005$), where the majority of adults were best fit by the Integrative model and the majority of children were best fit by the Prototype model.

3.4. *Modeling based on labeling strategy*

The results of the individual model fitting demonstrate that adults are more uniformly fit by the Integrative model, whereas children are more dispersed, with the greater majority are best fit by the Prototype model. Next, we explored a potential explanation for this finding by investigating the role of children's spontaneously labeling, a recall process that was borne out of the experimental task. More specifically, we evaluated trial level and group level differences in the model fits due to spontaneous labeling. At the trial level, of the 277 trials included in the analysis, 38% of them were labeled. At the group level, of the 33 child participants, 19 spontaneously produced at least one label. This appears to suggest that labeling was a consistently used strategy employed by children, especially on a trial-by-trial basis.

To better understand the effects of labeling at the trial level, we first split the data in two separate datasets, one containing only the labeled trials and the other containing only the unlabeled trials. We then fit the three models at the aggregate and individual subject level for both datasets (see Table 4). Unsurprisingly, the Integrative model provided the superior fit to both datasets, presumably

[10] For the individual differences analysis, we use group level priors rather than individual priors to assess our models. Although we find a great deal of consistency and low variance across categories, future work might evaluate the impact of subject level priors on recall performance in children.

because the model pays a lower cost for responses that, over the aggregate span between the observed target and category mean. However, at the individual subject level for the labeled-only trials dataset, we found that the majority of participants' labeled trials (26) were best fit by the Noisy Prototype model, with only 3 participants best fit by the Integrative model, and 4 by the Noisy Target model. In contrast, for the unlabeled trials dataset, the distribution was less skewed, where 9 participants were best fit by the Integrative model, 7 by the Noisy Target model, and 17 by the Noisy Prototype model. A Fisher's Exact test revealed a marginally significant difference ($p = .054$) in the distribution of best fitting models between the labeled trials and unlabeled trials, such that most participants' labeled trials were best described by the Prototype model, while the unlabeled trials were slightly more dispersed. Although this unplanned analysis should be interpreted with caution, it provides support for the idea that trial-by-trial strategy differences may impact the degree to which expectation from category priors influences memory. However, an alternative account for the finding that labeling leads to more reliance on the prototype is that some other common causal factor at the child level explains this result, rather than the trial-by-trial level strategy differences. That is, children who tend to label may also happen to be children who tend to be best explained by a prototype model, but these might be independently explained by a third factor[11].

Although we cannot definitively identify a causal role without an experimental intervention, analyzing this result at the child level provides support against this alternative account. That is, if labeling was a general strategy, we might expect that for children who can be classified as consistent labelers, all of their individual data (not just labeled trials) might be better fit by the Prototype model. In turn, the consistent labelers might be driving the finding of preschoolers best fit by the Prototype model. To evaluate this possibility, we first classified children into two groups: labelers and non-labelers. Labelers referred to preschoolers who provided labels (at either study, test, or both) on more than 50% of trials ($n = 10/33$) and non-labelers were all other children tested ($n = 23/33$). We chose to use this classification because spontaneously labeling on more than 50% of trials suggests a consistent strategy of the individual to assist in recall. The proportion of children best fit by the three models did not differ between labelers and non-labelers, Fisher's Exact test, $p = .50$, suggesting that individual differences between children are unlikely to account for the fact that different labeling strategies associated with different best fitting models at the trial level.

4. Discussion

In this paper, we investigate the interaction between prior category expectations and episodic memory in early childhood. We first quantified preschoolers' color category knowledge to understand the expectations they might bring to the task of remembering. After, we investigated preschoolers' episodic memory to assess the influence color expectations have on recall. Performance in both tasks was then compared to that of adults. We employed the use of probabilistic modeling techniques to further explore potential explanations of how error is represented in memory by fitting the behavioral data to three specific models: The Noisy Target, Noisy Prototype, and Integrative. We also evaluated individual differences in the fits of these models to each age group.

In Study 1, we assessed color knowledge in both children and adults and found that by age four, children's color categories appear to be similar to adults (i.e., no difference in category means) with the exception that children's aggregated category responses (and possibly category boundaries) are noisier as evidenced by larger standard deviations in the means. In Study 2, we assessed memory for color and found that recall of studied hue values, for children (like adults), regressed towards the color category means. This suggests that on average, when children are reconstructing events from memory, like adults, they rely on prior color category expectations to inform recall.

Although there were marked similarities between children and adults' recall processes, one additional comparison to note is the amount of "noise" between the target representation in the children and adult's Integrative processes. Indeed, when comparing the best sigmas to fit adult versus child responses, we found that the adults have a smaller sigma ($\sigma = 12$) than the children ($\sigma = 18$). The slope of the regression lines fit to recall similarly captured this—as wider targets (larger sigmas) lead to a stronger pull from the category knowledge. This might suggest a greater reliance on the category prototype for children relative to adults.

To further evaluate recall performance, we then explored whether differences in the fit of the three recall models varied at the individual subject level for both age groups. There is some work beginning to explore whether category biases vary at both the aggregate and individual subject levels. For example, Landy et al. (2017) assessed category influences on the recall of emotional facial expressions found that, at the aggregate level, adults regress toward the mean of the spectrum of facial expressions. However, at the individual level, recall was biased by 3 different emotional expression categories: one at the happy end of the spectrum (+1), one at the angry end of the spectrum (−1) and one that captured the entire range of the category. Thus, to explore potential differences in model fits at the individual subject level, we fit the three recall models to individual subject data for both age groups.

Unsurprisingly, we found that at the individual subject level, the greatest proportion of adults were best fit by the Integrative model (80%), followed by the Noisy Prototype model (11%), and then the Noisy Target model (8%). In stark contrast, the proportion of model fits to the children's data was more diffuse. A significantly greater majority of preschoolers were best fit by the Noisy Prototype model (45%), followed by the Integrative model (33%), and then the Noisy Target model (21%). Thus, not only were preschooler's data

[11] Another potential factor that may impact age differences in memory outcomes is a learner's motivation to encode information. Previous work exploring motivation and learning in young learners suggest that when paradigms are age appropriate and engaging, young children exhibit better memory outcomes in comparison to when they are not engaging (Ngo, Newcombe, & Olsen, 2019). However, this engagement does not always improve performance to adult levels, suggesting cognitive development is still a main factor driving differences between age groups. We controlled for this by making our task interactive. That is, children had to point to indicate their answers and periodically responded when asked if they were ready to move on.

noisier than adults (evidenced by the noise parameters), the process they employed to reconstruct events from memory was more variable across child participants. Furthermore, on individual trials, child participants adopted the recall strategy of spontaneously labeling the studied colors, which further promoted use of the Noisy Prototype process.

4.1. Relevance

The results of this study support and extend the findings of previous research showing the prior category knowledge influences recall (Duffy et al., 2006; Persaud & Hemmer, 2014). Although much is known about how category knowledge and episodic memories interact in adulthood (Allred et al., 2015; Bae et al., 2015; Hemmer & Steyvers, 2009a,b), far less is known about how this process is adopted in early memory development. Evidence formally evaluating the role of category knowledge in the reconstructive process of memory in development is thin (Duffy et al., 2006). Here we find computational support for this Integrative process at the aggregate level. We also extend the findings of previous work by computationally showing that at the individual subject level, children are much more variable in terms of the best fitting reconstructive processes (with greater reliance on category knowledge). Further we find that differences in memory strategies such as labeling affects the fit to these different memory processes.

There are a number of factors that might explain the finding that the Noisy Prototype model captured performance from more children than the Integrative model in terms of best fit at the individual level. First, early memory development is marked by an up prioritization of category information over nuanced episodic information. Also, neural mechanisms that support semantic category information are thought to mature before those supporting episodic memory (Keresztes et al., 2018). As a result, a majority of children may have encoded target information as a pointer to the category from which the target belongs, such as a category representative (i.e., the category mean) as opposed to encoding the exact color value studied. Such behavior would equate to encoding a red color value as a prototypical shade of red (e.g., the color of a red apple) as opposed to encoding the specific shade of red studied. In this way, children may be reliant on the information that is encoded and filtered through their more mature memory system.

Alternatively, it could be the case that the use of category knowledge happens at retrieval. After the initial testing phase, the original studied information could have become less accessible over time and instead of reproducing the noisy information, children reproduced a value closer to the category representative to reduce error and/or uncertainty. Whether the influence of category knowledge occurs at encoding, retrieval, or both is a question for future research. This lays the foundation for future work in memory assessing not only the importance of category knowledge in episodic memory in early childhood, but how this information might serve the strategic nature of reducing uncertainty in the reconstruction process across memory development.

This work connects to broader investigations of episodic memory development, particularly studies examining mechanisms that might impact age-related differences in recall. It is well known that relational episodic memory, on average, is better for young adults, than both younger children (Ngo et al., 2018) and older adults (Chalfonte & Johnson, 1996). However, the gap in episodic memory performance is attenuated when older individuals are tasked with making semantically relevant memory judgments (e.g. was a word studied with a semantically related or unrelated pair), but not for semantically unrelated judgments (e.g. word pair location – Jarjat, Ward, Hot, Portrat, & Loaiza, 2020). This might suggest that semantic category information plays a facilitatory role in the reconstruction and/or preservation of information in episodic memory for older adults. In our work, we show a similar importance or reliance on semantic category information for the purposes of recall in preschool-aged children. This has important implications for how episodic memory across development is understood and evaluated. For example, experimental paradigms that remove semantic category information from the stimulus environment are particularly taxing for groups that rely on this information to aid memory retrieval. Certainly, studying episodic memory with such controlled methods is important. However, doing so might obscure the natural or real-world reconstructive process of episodic memory for certain groups, namely those groups that rely heavily on semantic information. Thus, future work might further examine stimulus driven mechanisms that foster age-related differences in episodic memory performance across development.

Importantly, computational approaches will be useful in this endeavor because they can tease apart the relative contribution of different factors to recall. It should be noted that understanding the role of category knowledge in episodic memory is not just to observe an increase in memory accuracy to a young adult-like degree. It is well known that category knowledge can aid recall when episodic content is noisy and incomplete. Instead, the goal for investigating category knowledge in development is to assess whether differences in how the integrative recall process is used tracks with developmental shifts in episodic memory performance. Computational approaches are also important for specifying the computational problems the memory system seeks to solve across development, and for evaluating how the memory system makes use of information that is readily available to reduce uncertainty and memory load.

This work also connects to pre-existing literature on episodic memory binding in development. Here we consider whether our task of binding color to shape is the same as relational binding as defined in the literature. In traditional developmental relational binding tasks, the bound items might occupy clearly distinct representations in memory and the task for the participant is to bind those representations. For example, Sluzenski et al. (2006) tested relational binding using a task requiring participants to bind distinct objects with their accompanying backgrounds. In our task, however, there might not be a general consensus on whether color, a feature of the study object, occupies a separate representation in memory from the shape, and thus, whether this is truly a relational binding task. Based on the argument of distinct representations in memory between features and objects, we suggest that color to shape is a relational binding task.

First, previous research has classified binding of color-to-object as a relational binding task (e.g. Chalfonte & Johnson, 1996). Additionally, research on working memory capacity in infants suggests that in infancy, memory for objects comes online earlier in development and is dissociable from memory for the features of objects (Kibbe & Leslie, 2008; Kibbe & Leslie, 2011). While infants can

remember the presence of an object, memory for features of the object are still developing. Moreover, recent work assessing memory properties in adults have found that objects and their features are not always stored as a unitized entity such that memory for one feature of an object does not correlate with memory for other features of the same object (Utochkin & Brady, 2020). Thus, if features and objects are stored separately in memory, linking those features to objects is a relational binding process.

More broadly, this work further connects to research on how category information may support learning more generally. For example, past work suggests that children employ prior beliefs to reason about causal relationships (e.g., Schulz, Bonawitz, & Griffiths, 2007), to guide attention (Kidd, Piantadosi, & Aslin, 2012), exploration (e.g. Bonawitz, van Schijndel, Friel, & Schulz, 2012), and to facilitate word learning (Borovsky, Ellis, Evans, & Elman, 2015). In this way, learning depends on the ability to integrate prior knowledge and experience with new event information. Our work potentially informs this literature by suggesting that the use of pre-experimental prior category knowledge extends to other cognitive domains, namely episodic memory, suggesting that the integration of relevant prior representations with incoming information is a general cognitive mechanism.

In this paper, we show that children have clear category expectations and that they use this information to help reconstruct events from memory. This leads to a more interesting and perhaps contentious debate regarding the nature of categories and category learning. Indeed, there is ample evidence to suggest that children develop category expectations early in life, such as structured spatial categories (Huttenlocher, Newcombe, & Sandberg, 1994), and that they use this information to form spatial judgments about objects in the environment. It is beyond the scope of this paper to address how early these expectations develop (as we focus on preschoolers for methodological practicality), though our results suggest that Bayesian frameworks may provide one starting point for characterizing this acquisition.

4.2. Future work

Our results raise many interesting avenues for future research. First, the uninstructed strategy of verbal rehearsal played a particularly important role in recall for children. It appears to have reduced uncertainty in children's recall, but also resulted in a greater regression to the category mean. As a result, future work might explore whether this process of spontaneous verbal labeling is a universal recall strategy across different stimulus domains and across age groups, and whether its strategic adoption is associated with other aspects of cognitive development. Because the goal of this paper was to evaluate recall in the youngest population possible that allowed for adult-matched procedures, there was not much variability the childhood ages sampled in the memory study, limiting power for this additional analysis. However, we examined spontaneous labeling and development with an analysis of a median split of the data. This revealed a greater production of spontaneous labels and reliance on category means by the older preschoolers in our sample (see Appendix B for details), suggesting that the older preschoolers were adopting this mnemonic strategy. This result is also consistent with previous evidence showing that older children begin to exhibit memory errors when phonological information (i.e., an object label) is ambiguous but younger children do not (Hitch et al., 1989). This would suggest that at different stages of development, children are relying on and/or have access to varying coding processes (visual or verbal) and this may impact the kinds of errors we see in their recall. Lastly, future work regarding labeling might also examine whether the degree of reliance on category means is still observed when spontaneous labeling is not permitted (as is observed in articulatory suppression tasks), helping to illuminate the impact of cognitive strategies on default memory processes.

Open questions also remain surrounding the mechanisms that drive the reliance on prior knowledge to inform memory gaps. Despite the clear influence of expectations on memory for color across cultures and development, not all categories of knowledge are as robust early on in life. Furthermore, substantial research suggests that cognitive processing of color information is, by nature, categorical (e.g. Berlin & Kay, 1969; Persaud et al., 2017; Pitchford & Mullen, 2003). Therefore, color is a domain that lends itself quite well to the process of integrating category knowledge, given that categorical information is easily retrieved through verbal labeling. However, category information for other stimulus domains might not be as readily accessible. Although in this study children have demonstrated the ability to integrate prior color category knowledge when recalling a specific hue value and rely on this knowledge to a greater degree than adults, this may not be the case for other domains of knowledge. However, there is some related work demonstrating that children pool together information from multiple streams to make cognitive judgments in other domains. For example, Nardini, Bales, and Mareschal (2016) showed that when children combine visual and auditory information in a speeded spatial task, they have less variable response times in comparison to when either type of information is presented alone. Additionally, although model predictions comparing response time distributions indicate that children at all ages (4–12-years old) were integrating information, overall speed and efficiency of sensory integration improved with age. Further exploration of cue integration in children (7–10-years old) suggests that the ability to integrate cues and reduce noise is contingent on whether the learner has sufficient representations and biases in their perceptual estimates that are relevant to the task (Negen et al., 2019). This work suggests that a learner's ability to integrate information may depend upon the nature of their representations and the stage of their cognitive development, but that integration persists in other domains, nonetheless.

Indeed biases in memory for feature values across different perceptual spaces (e.g. spatial location, height, size) have also been demonstrated. For example, Huttenlocher and colleagues showed that people are biased toward subjective spatial categories (e.g. four quadrants in a circle) when estimating the locations of dots in a circle. These spatial categories can improve average accuracy, particularly for noisy represented study locations. Presumably, different experiences (e.g., quantity and quality) will impact recall in the same scenario in different ways for different individuals. Indeed, the degree to which an individual relies on prior knowledge to fill in memory gaps will depend on both the fidelity of memory and the strength of domain knowledge. A learner with poor fidelity and strong category expectations might be expected to show strong drift towards an expected category prototype, whereas a learner with higher fidelity and weaker expectations may show very little drift. Future work may examine how these competing factors may interact

with one another.

Relatedly, there might be interesting differences in the degree of reliance on prior knowledge for categorical domains that are less perceptual and sensory driven in nature (e.g. word lists). There is some evidence to suggest that individuals employ information that is akin to category information in the form of semantic associations when recalling information from memory. For example, adult participants were biased to recall words from a study word list that were never present on the list but were strongly associated with the semantic category of the words on the list (Roediger & McDermott, 1995). We predict that children would integrate in a similar fashion provided that they have access to knowledge of the category. However, whether or not using categorical information within these less perceptual domains is a useful process for reducing uncertainty for children is an empirical question.

Lastly, we also observed that four-year-old preschoolers are already using self-generated encoding and retrieval cues to facilitate later recall. That is, children were not instructed to label colors as they were presented in the task. Instead many children spontaneously verbally rehearsed colors to aid recall. Thus, it also becomes interesting to ask whether young learners are employing a metamemory computation or evaluation of what information might be easily retrieved and how a facilitatory mechanism, like spontaneously labeling might aid the retrieval process. Future research is needed to investigate how the memory system operates in other domains, and how recall outcomes vary based on the degree to which domain knowledge is well established/developing and robust.

Despite the specificity of this current work (i.e., memory for color knowledge), the findings presented in this paper have theoretical implications for the strategic nature of the developing memory system. Our findings suggest that, at least within the domain of color, when preschoolers are somewhat able to access noisy memory traces, they adopt an adult-like reconstructive process that may be optimal for learning. When memory is less clear, they still strategically use information readily available to them in the form of category knowledge. This might be one potential explanation as to why children are robust learners despite having a limited memory capacity.

While it might be true that children have noisier memory in comparison to adults, we see that they are integrating prior knowledge and episodic trace to reconstruct memory, suggesting that this mechanism comes online early in memory development. Interestingly, our results might further suggest that it is not just a maturation of structures that supports episodic memory that is changing in development, but also a fine-tuning of the optimality by which category knowledge and episodic memory trades off in reconstruction. Importantly, there are well-defined models of memory in the adult literature that can be used to explore memory in young learners. In combination with existing adult memory models, computational modeling techniques are a great tool to address some of these questions.

4.3. Conclusions

Reconstructing events from episodic memory is an essential process of everyday cognition. An important question is what processes might be adopted by the minds of young children to reduce uncertainty in their memory representations and accomplish this important cognitive task. To address this question, we assessed recall in preschoolers and adults and implemented probabilistic modeling techniques to understand the processes that generated noise in memory. We explored three processes—the Noisy Target process which captures random noise, Noisy Prototype process which captures prototypical noise, and an Integrative process which captured category biased episodic trace noise—and investigated which process best captured the recall data. Both statistical and computational analyses found that for adults at both the aggregate and individual subject level, responses were best captured by the Integrative strategy. However, reconstruction for preschoolers was more nuanced suggesting a greater reliance on category information and variation in this reliance across individual child participants.

This work sheds light on our current understanding of episodic memory, specifically relational memory in development. While we found shifts in recall performance across development, overall integration of category knowledge was found in both children and adults. The fact that this process is demonstrated in both early development and adulthood compounds the evidence that the employment of category knowledge is an important underlying mechanism of episodic memory. Furthermore, our approach has uncovered changes in recall performance across development that reflect a change in the degree of integration of category knowledge, providing an important next step in more precisely characterizing what is changing in memory development and a potential model of why.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received support from the National Science Foundation Graduate Research Fellowship under Grant Number NSF DGE 0937373 (KP), National Science Foundation SBE Postdoctoral Fellowship NSF-SMA 1911656 (KP), National Institutes of Health, IMSD Minority Biomedical Research Support Program under grant number 2R25GM096161-07 (CM), National Science Foundation CAREER Grant Number 1453276 (PH), NSF SES-1627971 (EB), and the Jacobs Foundation (EB). These funding sources were not involved in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit this article for publication. The child color memory data was presented independently at the 2017 Cognitive Development Society (CDS) conference as a talk titled, "Expectations about color categories inform preschoolers recall" in Portland Oregon.

Appendix A

Evaluation of the three processes: The Noisy target process, Noisy Prototype process, and Integrative process

There are a number of hypotheses that might explain the recall process of young learners as they engage in the color memory task. Here we explore three: The Noisy Target process, Noisy Prototype process, and Integrative process. The Noisy Target process predicts that young learners will noisily recall target hue values without incorporating their knowledge of the category mean hue value. This recall behavior would result in no systematic patterns born out in the data (no regression). In contrast, the Noisy Prototype process predicts that young learners will rely solely on their knowledge of the mean hue value of each category to facilitate recall. This behavior would reflect a strong, systematic bias toward the category means (steep regression at slope of −1) in the recall data. The Integrative process, consistent with Bayesian strategies of memory, predicts that young learners will integrate their noisy episodic traces of the target hue value with their knowledge of the mean of the color categories to produce recall. Similar to the Noisy Prototype process, the Integrative process also predicts a systematic bias towards the color category mean, but to a lesser degree due to the integration of the noisy memory traces.

To evaluate each process, we computed the probability densities of all responses given certain sets of assumptions for how noisy memory traces and category knowledge interact. Each process makes unique assumptions for how noise is represented in memory and the role that category knowledge plays in the recall process. Thus, the probability densities tell us how likely a response is given the unique process assumptions for how the response was generated. Importantly, the differences in assumptions between processes are reflected in how the probability distributions are parameterized. Each was implemented separately for the adults and the children. For the adults, the models were fit to all 34 participants and for the children, they were fit to all 33 participants.

The Noisy Target process assumes that the only information a participant uses when trying to recall a color is the original target color that was studied. Because memory is often imperfect, the original representation might be jittered by noise. Thus, for this process we asked how likely responses were if they were drawn from a distribution that is centered on the original target value, offset by noise. To address this question, we first calculated the difference between recall responses and the corresponding studied hue values. Normalizing the data in this way resulted in values that were positive and negative. This normalization allowed us to then compute the probability of each response from a single normal distribution where the parameter values were fixed[12] (i.e., centered over 0 with some noise). To determine the noise value, we iterated over 100 different noise values, ranging from 0 to 200 in steps of 2, in search of the value that maximized the log likelihood of responses under this process. This search through the space of possible noise values revealed that the best fitting values were 12 for adults and 18 for children. In other words, if participants were drawing responses solely from the target distribution, the amount of deviation from the target (i.e., the noise) that best captured responses allowed for standard deviations of 12 and 18 for adults and children, respectively. Note that the best fitting noise value for children was greater than adults. Intuitively, this suggests that although the two groups might be using the same process to make responses, children's responses were slightly noisier than adults.

After computing the probability densities of the observed responses under this process, we then calculated the log likelihood of those probabilities. For process comparison, we summed those log likelihood values to get a single value to reflect the fit of the Noisy Target process. Given that the assumed memory noise for adults and children were different, we evaluated the Noisy Target process for adults and children separately.

Noisy Prototype process

The Noisy Prototype process assumes that the only information a participant uses when trying to recall a color is the category information from which the studied color belonged. Thus, for this process we asked how likely responses if they were drawn from a distribution centered on the mean of the color category, offset by noise. Recall that these category means were learned from Study 1a and 1b.

To evaluate the Noisy Prototype process, we first normalized the distribution of participant responses by calculating the difference between recall responses and the prior category means. We then computed the probability densities of each response value from a distribution centered on the corresponding category mean with the category noise. After, we computed the log likelihood of the observed set of responses under this process. As with the Noisy Target process, we evaluated the Noisy Prototype process separately for children and adults.

Integrative process

The concept of the Integrative process was to combine the distribution from the target model with the distribution from the

[12] We could have also evaluated the probability of responses without normalization (i.e., without subtracting out the studied value). To do this, we would have to change the mean of the distribution for each studied value. This is unnecessary given that the probabilities would be exactly the same under both distributions. For example, the probability of the response 100 to the study value 88 given a normal distribution with a mean of 88 and standard deviation of 12 is exactly the same as the probability of 12 (i.e., response - target) given a normal distribution with a mean of 0 and standard deviation of 12.

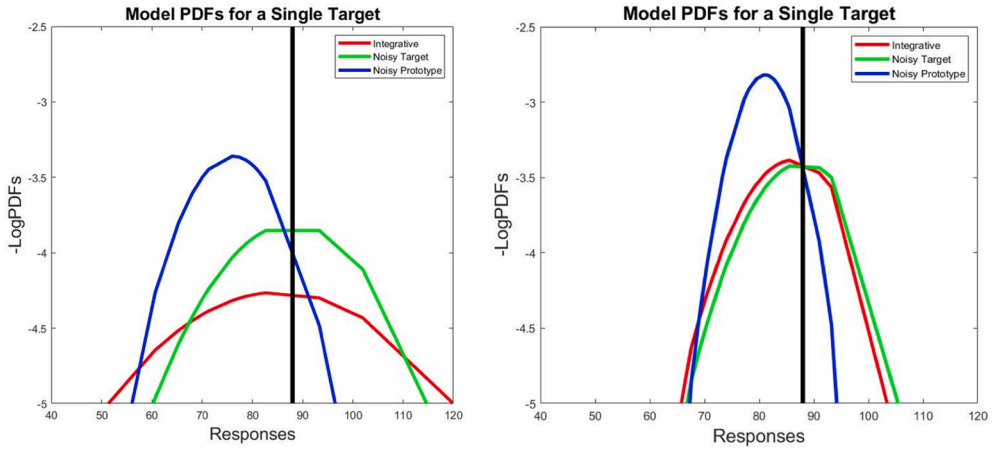

Fig. A1. Example pdfs from simulating aggregated responses from the three processes for a green target value. Left column: fits to the child data for the hue value 88. Right column: fits to the adult data for the hue value 88. In both graphs, the solid black line marks the location of the studied target value. Note that in both graphs, the majority of the mass of the Noisy Prototype (blue line) is shifted to the left of the studied value. This is because the study values are the high color targets within the category. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Prototype process to produce an integrated distribution to evaluate responses. In other words, the Integrative process assumes that participants use information from both the original study value as well as category information when trying to recall a color from memory. The simple rational model assumes that noisy data in the mind is optimally combined with prior knowledge about the environment. Bayes' rule gives a principled account of how to combine noisy memory representations with prior expectations to calculate the posterior probability,

$$p(\Theta|y) \propto p(y|\Theta)p(\Theta) \tag{A1}$$

The posterior probability $p(\Theta|y)$ describes how likely feature values Θ are given noisy memory traces y and prior expectations for the feature $p(\Theta)$. Thus, implementing the Integrative strategy was a two-step process. For a single response value, we first computed the mean and standard deviation of a posterior distribution integrating a target distribution (centered on the study value with memory noise) and the category distribution (centered on the category mean with category noise). Because both the prior and likelihood distributions are Gaussian (which are self-conjugate), the mean and variance follow relatively straightforwardly. The mean of the integrated distribution is given by,

$$\frac{1}{\frac{1}{\sigma_t^2}+\frac{n}{\sigma_c^2}}\left(\frac{t}{\sigma_t^2}+\frac{\mu_c}{\sigma_c^2}\right) \tag{A2}$$

And the noise of the integrated distribution is given by,

$$\frac{1}{\frac{1}{\sigma_t^2}+\frac{n}{\sigma_c^2}} \tag{A3}$$

where σ_t refers to the memory noise on the target distribution, σ_c refers to the noise on the category distribution (learned from Study 1), refers to the studied target value, and refers to the mean of the category distribution to which the target value belongs. For instance, if an adult participant studied the hue value 88, $\sigma_c = 6.69$, $t = 88$, and $\mu_c = 81.07$.

The memory noise parameter of the target distribution (σ_t) was treated the same as the aforementioned target strategy in that we iterated over 100 different values on the range of 0–200 to determine the noisy value that maximized the likelihood. We followed this maximization procedure for adults and children and the best noise values were 4 and 6 respectively. (For context, the hue range spans between 0 and 239 where a typical range for a single hue category might be about 20 "steps". Thus, the noise terms span approximately a quarter of the range of a typical category's hue values.)

Once the parameter values of the posterior distribution were computed, we then calculated the probability densities and log likelihoods of the responses given the posterior distribution. To reiterate the results, analysis of the log-likelihood values for each process provided strong evidence in favor of the Integrative strategy for both adults and children. See Fig. A1 for probability densities from each model fit to responses from a single studied value

Table 5
Frequency of Model Fits based by Median Split Child Age.

Model	Count (%)	
	Young	Older
Integrative	6 (37.50%)	5 (29.41%)
Noisy Target	6 (37.50%)	1 (5.88%)
Noisy Prototype	4 (25.00%)	11 (64.71%)

Appendix B

Within preschooler's age-related comparisons

Motivated by the Duffy et al. (2000) finding of a steeper regression in younger, relative to older children, we sought to evaluate age-related differences in modeling fitting using the data from our current sample of preschoolers. To evaluate group differences, we performed a median split to classify children as younger and older learners and then compared the proportion of younger and older children described by each model. Of the 33 participants in the study, 16 were classified as young and 17 were classified as older. The median age of the total sample was 53mos. (sd = 6mos.).

The median ages for younger and older children were 49mos. (sd = 2 mos.) and 56 mos. (sd = 5 mos.), respectively. We also sought to evaluate group differences due to spontaneous labeling that was borne out of the experimental task. Of the 16 children classified as younger, 7 produced at least one label and of the 17 older children, 12 produced at least one label. This further suggests that labeling was a consistent strategy employed by children in this task. To evaluate age related differences in the best fitting noise value, we implemented the Integrative model and for each participant, we searched over the space of possible noise values for the value that maximized the likelihood for each participant's data.

To evaluate whether the proportion of children best fit by each of the three models was dependent upon age, we used the Freeman-Halton extension of the Fisher's Exact test to compute the (two-tailed) probability of obtaining a distribution of values in a 2 (young vs older) × 3 (Integrative vs Noisy Target vs Noisy Prototype) contingency table, given the number of observations in each cell (Table 5). The results revealed that the observed proportion of best fitting models was dependent on age ($p = .031$). In other words, there was a significant difference in the distribution of best fitting models between the age groups. Younger preschoolers were evenly split in the number fit by the Integrative ($n = 6$) and Noisy Target ($n = 6$) models, followed closely by the Noisy Prototype model ($n = 4$). Interestingly, however, older children had a different composition. A much larger proportion of older children were better fit by the Noisy Prototype model ($n = 11$), followed by the Integrative model ($n = 5$), and almost not at all described by the Noisy Target model ($n = 1$).

Age and best fitting noise parameter

We searched for the best fitting noise value at the individual subject level to test for age related differences within the preschool population. The goal was to assess whether a difference in the amount of noise between age groups could explain why young and older children were better fit by different models. In conflict with our prediction, there was a weak non-significant negative correlation between age and best fitting noise value ($r = -0.17$, $p = .35$). This suggests that a difference in the best fitting model between age groups was not a result of a difference in the amount of noise in the data.

Lastly, based on the finding of a difference in model fit between labeled and unlabeled trials, we re-examined the role of labeling on age. We had originally classified whole individuals as either labelers or non-labelers and found no significant difference by age. Instead, we calculated the proportion of labeled trials provided by younger and older children, to test whether as a group, older children were more likely to provide labels during testing. A Fisher's Exact Probability Test revealed a significant difference in the proportions of labeled and unlabeled trials contributed by each age group ($p = .002$). A larger proportion of labeled trials were generated by older (66%) compared to younger children (34%).

Appendix C. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2020.101357.

References

Allred, S., Bae, G. Y., Olkkonen, M., & Flombaum, J. (2015). A new model for the contents of visual working memory. *Journal of Vision, 15*(12), 83.
Bae, G., Olkonnen, M., Allred, S., Wilson, C., & Flombaum, J. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision, 14*, 1–23.
Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144*, 744–763.
Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkley, CA: University of California Press.

Bonawitz, E., Shafto, P., Yu, Y., Gonzalez, A., & Bridgers, S. (2020). Children change their answers in response to neutral follow-up questions by a knowledgeable asker. *Cognitive Science, 44*, e12811.
Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology, 64*(4), 215–234.
Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. (2015). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. *Developmental Science, 19*, 918–932.
Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24*, 981–990.
Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*, 207–230.
Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition, 24*(4), 403–416.
Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*, 2037–2078.
Diamond, A. (2006). The early development of executive functions. *Lifespan Cognition: Mechanisms of Change, 210*, 70–95.
Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review, 22*, 170–178.
Drummey, A. B., & Newcombe, N. S. (2002). Developmental changes in source memory. *Developmental Science, 5*, 502–513.
Duffy, S., Huttenlocher, J., & Crawford, L. E. (2006). Children use categories to maximize accuracy in estimation. *Developmental Science, 9*, 597–603.
Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*, 75–86.
Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767–773.
Hemmer, P., & Steyvers, M. (2009a). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review, 16*, 80–87.
Hemmer, P., & Steyvers, M. (2009b). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1*, 189–202.
Hitch, G. J., Woodin, M. E., & Baker, S. (1989). Visual and phonological components of working -, memory in children. *Memory & Cognition*, 175–185.
Hourcade, J. P., Bederson, B. B., Druin, A., & Guimbretière, F. (2004). Differences in pointing task performance between preschool children and adults using mice. *ACM Transactions on Computer-Human Interaction (TOCHI), 11*(4), 357–386.
Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352–376.
Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology, 129*, 220–241.
Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive Psychology, 27*, 115–147.
Jarjat, G., Ward, G., Hot, P., Portrat, S., & Loaiza, V. M. (2020). Distinguishing the impact of age on semantic and nonsemantic associations in episodic memory. *Journal of Gerontology: Series B*. gbaa010.
JASP Team (2020). JASP (Version 0.13.1) [Computer software].
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.
Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2018). Hippocampal maturation drives memory from generalization to specificity. *Trends in Cognitive Sciences, 22*, 676–686.
Kibbe, M. M., & Leslie, A. M. (2008). Evidence for separate development of working memory capacity for objects and for features in infants. *Visual Cognition, 16*(1), 114–117.
Kibbe, M. M., & Leslie, A. M. (2011). What do infants remember when they forget? Location and identity in 6-month-olds' memory for objects. *Psychological Science, 22*(12), 1500–1505.
Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE, 7*(5), e36399.
Landy, D., Crawford, L. E., & Corbin, J. (2017). A hierarchical Bayesian model of individual differences in memory for emotional expressions. *Proceedings of the 39th annual conference of the cognitive science society*. London, EU: Cognitive Science Society.
Lloyd, M. E., Doydum, A. O., & Newcombe, N. S. (2009). Memory binding in early childhood: Evidence for a retrieval deficit. *Child Development, 80*, 1321–1328.
Nardini, M., Bales, J., & Mareschal, D. (2016). Integration of audio-visual information for spatial decisions in children and adults. *Developmental Science, 19*, 803–816.
Negen, J., Chere, B., Bird, L. A., Taylor, E., Roome, H. E., Keenaghan, S., … Nardini, M. (2019). Sensory cue combination in children under 10 years of age. *Cognition, 193*, 1040–1114.
Newcombe, N. S., Lloyd, M. E., & Ratliff, K. R. (2007). Development of episodic and autobiographical memory: A cognitive neuroscience perspective. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 35, pp. 37–85). San Diego, CA: Elsevier.
Ngo, C. T., Newcombe, N. S., & Olson, I. R. (2018). The ontogeny of relational memory and pattern separation. *Developmental Science, 21*(2), e12556.
Ngo, C. T., Newcombe, N. S., & Olson, I. R. (2019). Gain-loss framing enhances mnemonic discrimination in preschoolers. *Child Development, 90*(5), 1569–1578.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.
Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1162–1167).
Persaud, K., & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology, 88*, 1–21.
Persaud, K., Hemmer, P., Kidd, C., & Piantadosi, S. (2017). Seeing colors: Cultural and environmental influences on episodic memory. *I-Perception, 8*, 2041669517750161.
Persaud, K., McMahan, B., Alikhani, M., Pei, K., Hemmer, P., & Stone, M. (2017). In *When is likely unlikely: Investigating the continuum of linguistic vagueness* (pp. 2876–2881). London, EU: Cognitive Science Society.
Pitchford, N. J., & Mullen, K. T. (2003). The development of conceptual colour categories on pre-school children: Influence of perceptual categorization. *Visual Cognition, 10*, 51–77.
Roediger, H. L. I. I. I., & McDermott, K. B. (1995). Creating False Memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.
Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology, 43*, 1124–1139.
Sluzenski, J., Newcombe, N., & Kovacs, S. L. (2006). Binding, relational memory, and recall of naturalistic events: A developmental perspective. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*, 89–100.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629–641.
Utochkin, I. S., & Brady, T. F. (2020). Independent storage of different features of real-world objects in long-term memory. *Journal of Experimental Psychology: General, 149*, 530–549.
Bychkovsky, V. (2020). hsl2rgb and rgb2hsl conversion (https://www.mathworks.com/matlabcentral/fileexchange/20292-hsl2rgb- and-rgb2hsl-conversion), MATLAB Central File Exchange; Image Processing Toolbox, MATLAB, The MathWorks, Inc., Natick, MA.
Yee, E., Jones, M. N., & McRae, K. (2018). Semantic memory. In J. T. Wixted, & S. L. Thompson-Schill (Eds.), Stevens Handbook of experimental Psychology and cognitive neuroscience, (4th ed., Vol. 3: Language and Thought, pp. 319–356). New York: Wiley.
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–235.