Contingency of semantic generalization on episodic specificity: Variations across development

Chi T. Ngo*[1], Susan L. Benear*[2], Haroon Popal[2], Ingrid R. Olson[2], & Nora S. Newcombe[2]

[1] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

[2] Department of Psychology, Temple University

* Equal contributions

Correspondence concerning this article should be addressed to the lead contact: Chi T. Ngo.

Address: Lentzeallee 94, 14195 Berlin, Germany. Email: ngo@mpib-berlin.mpg.de

Summary

Popular computational models of memory have posited that the formation of new semantic knowledge relies on generalization from memories of specific but related episodes, at least when it occurs rapidly. This view predicts a contingency between new generalizations and episodic memory. However, very young children readily accumulate semantic knowledge at a time when their episodic memory capacities are fragile. This phenomenon challenges the notion that semantic knowledge acquisition and rapid generalization are necessarily gated by episodic memory. Here, we tested whether generalization depends on memory for individual episodes in children from 3 to 8 years of age and contrasted their performance with adults. We found that the interdependence of generalization and episodic memory changed across development. Young adults' generalization success was contingent on their memories for an item linked to its episodic context. In contrast, generalization by young children was contingent on memories of the specific identity of items and the availability of the conceptual common ground linking related episodes. This age-related contrast favors models of memory that can account for the relations between rapid generalization and episodic memory in immature systems.

Keywords: episodic memory, generalization, memory development, context binding, pattern separation, memory specificity

1. Introduction

People accumulate general knowledge about the world to guide exploration and support novel inferences, i.e., they acquire semantic memory. They also form memories of specific past events, i.e., episodic memory. Semantic memory emphasizes *generalization*, while episodic memory preserves the specificity of individual episodes through *binding* processes that link together multiple elements of an event and *pattern separation* processes that distinguish similar experiences. Many theories of memory suggest that multiple memory systems play complementary roles in supporting different mnemonic goals. However, the nature and extent of their inter-independence is unclear.

1.1. Multiple memory systems

Semantic memory is a dynamic collection of general knowledge disconnected from the place and time of initial learning. It is useful for generating predictions about appropriate actions in novel situations (Ghosh & Gilboa, 2014; Tulving, 1972). For instance, I might offer my lab mate tea after realizing that she has opted for tea over coffee previously. In contrast, episodic memory is characterized by remembering rich, personal experiences with high specificity. For instance, I can recall the time I accidentally knocked over my lab mate's owl-shaped mug filled with black tea in a lab meeting. Semantic and episodic memory are traditionally considered to be dissociable but highly intertwined memory systems (Schacter and Tulving, 1994; Tulving, 1972). However, there are opposing ideas about the nature of this dependence. Some memory models suggest that one route to semantic memory acquisition is that newly learned information is encoded initially as hippocampal-dependent episodic memory and then, either through repetition or gist extraction, becomes semantic (Complementary Learning Systems: McClelland, McNaughton, & O'Reilly, 1995; Standard Consolidation Theory: Winocur, Moscovitch, &

Bontempi, 2010). These views also posit another route to semantic memory acquisition—a *slow* learning mechanism—that bypasses the hippocampus and is supported primarily by the neocortex (McClelland et al., 1995). However, because generalization can occur rapidly, more recent models such as REMERGE (Recurrency and episodic memory results in generalization; Kumaran & McClelland, 2012) suggest the importance of a hippocampal big-loop recurrence that captures the higher-order relations among related episodes.

From this point of view, the fast emergence of novel semantic memory necessarily goes through episodic memory gating. In support of this view, superior generalization performance in adults is associated with better memory for the specific episodes that support the inferences (Banino, Koster, Hassabis, & Kumaran, 2016; Tompary, Zhou, & Davachi, 2020). In computational simulations, the behavioral findings were best accounted for by a mechanism in which generalization occurs at the point of retrieval, through the combination of related episodes on the fly. Further, patients with amnesia due to medial temporal lobe damage exhibit episodic memory impairments, and these impairments also hinder their performance on tasks that tap semantic memory (Greenberg, Keane, Ryan, & Verfaellie, 2009). In sum, these findings suggest a degree of contingency between episodic memory capacities and rapid generalization.

In contrast, the serial-parallel-independent (SPI) model posits that semantic memory encoding can be independent of episodic memory, but episodic memory encoding is dependent on semantic memory (Tulving, 1995; Tulving & Markowitsch, 1998). This tenet is based in part on the observation that episodic memory is a later-developing system, born out of semantic memory. Fitting well with this view are neuropsychological findings that developmental amnesia due to early-in-life hippocampal damage is associated with deficits in episodic memory, whereas semantic memory acquisition is relatively preserved (Vargha-Khadem et al., 1997), even though new semantic memories are acquired at a lower rate (Bindschaedler, Peter-Favre, Maeder, Hirsbrunner, & Clarke, 2011), and require a greater number of repetitions compared to healthy controls (Gardiner, Brandt, Baddeley, Vargha-Khadem, & Mishkin, 2008).

Separate lines of memory development

A central element in the debate on whether fast generalization requires episodic memory is the observation that semantic and episodic memory have asynchronous developmental profiles. Extracting commonalities across contexts is adaptive, as available semantic knowledge enables appropriate novel inferences based on existing conceptual structures (Keresztes, Ngo, Lindenberger, Werkle-Bergner, & Newcombe, 2018; Ramsaran, Schlichting, & Frankland, 2019) and lies at the heart of vocabulary acquisition (Clark, 2001). Infants and toddlers begin to amass generalizable knowledge about objects (Booth & Waxman, 2002) and familiar events (Hudson, Fivush, & Kuebli, 1992), providing evidence for substantial semantic memory ability very early in life. Likewise, generalization behaviors are apparent early in development: Infants and toddlers can detect recurring patterns across multiple experiences around 8 months of age (Saffran, Aslin, & Newport, 1996). They can also quickly generalize about object properties around 10 months of age (Baldwin, Markman, & Melartin, 1993). Young children between one and two years of age are capable of generalizing simple sequences through deferred imitation (e.g., a three-step sequence of making a party hat; Bauer & Dow, 1994; Lukowski, Wiebe, & Bauer, 2009). The capacity to generate knowledge through cross-episode integration on the fly is present early on in development but increases over the preschool and early school years (Bauer & San Souci, 2010; Varga, Stewart & Bauer, 2013). Taken together, some forms of generalization, including the rapid form, appear to be present early on in life.

Evidence of semantic memory acquisition and generalization behaviors in infants and toddlers appears mismatched with their relatively fragile capacities to remember the specifics of past events within their rich spatiotemporal contexts (Tulving, 1972). Evidence for what-where memories is not seen until the end of the second year of life (Newcombe, Balcomb, Ferrara, Hansen, & Koski, 2014). Relational memory, or context binding, becomes much more robust over the subsequent years, and is quite good by age 6 or 7 (e.g., Ngo, Newcombe, & Olson, 2018; Sluzenski, Newcombe, & Kovacs, 2006), with continuing improvements well into late childhood (Ghetti & Bunge, 2012).

In addition to binding processes, mnemonic discrimination between similar items (e.g., Canada, Ngo, Newcombe, Geng, & Riggins, 2019; Keresztes et al., 2017) or between complex associations learned in similar contexts (Ngo, Lin, Newcombe, & Olson, 2019) is another facet of episodic memory that shows protracted development. Four-year-old children were more likely than 6-year-old children to confuse a perceptually similar exemplars as something identical to what they previously saw. Although 6-year-olds did not show such a tendency, their discrimination level was not above chance, as it was in young adults (Ngo et al., 2018). Mnemonic discrimination has been thought to rely on pattern separation, a hippocampally-dependent neurocomputation that reduces the overlap between similar inputs (Norman & O'Reilly, 2003). Improvements in processes of relational binding and pattern separation that support highly specific episodic memories are thought to underlie the critical transition from fragile to robust episodic memory capacities in childhood (Newcombe, Benear, Ngo, & Olson, in press; Riggins, Canada, & Botdorf, 2020).

1.2. Current study

These developmental patterns show that some forms of generalization are present in the face of frail episodic memory capacities in early childhood. However, past research has primarily studied semantic and episodic memory processes separately and has focused on different developmental windows with entirely different paradigms, creating critical blind spots in our understanding of their co-development. Importantly, charting capacities in acquiring semantic knowledge and retaining the episodic details across development would have crucial implications for theorizing about the dependence between the semantic and episodic memory systems. Leveraging an age window in which aspects of episodic memory undergo substantial age-related changes, we targeted two questions: (i) is rapid generalization contingent on remembering the specifics of the past?, and (ii) if so, does the contingency differ across development?

To this end, we created an experimental paradigm that assesses generalization and episodic memory specificity using the same set of experiences but different tasks. Children and adults learned about various cartoon characters. Each character went to various places and found various objects for their "collections". These objects were semantically related, and all contexts were semantically congruent with the category of the objects. We defined *generalization* as the ability to detect and accumulate recurring features across related experiences such that it can apply to novel situations. In this case, we examined the ability to make a novel inference about a character based on the semantic overlap among the objects seen with each character. To examine episodic memory, we tested memories for specific item-context pairings (i.e., context binding). We also tested two kinds of mnemonic discrimination, which have not always been clearly distinguished in prior work. There has generally been a focus on the perceptual details of component items in an episode, e.g., a red or a green apple. We call tests of this kind "item perceptual specificity". We also included tests examining the identity of the items, e.g., an apple or a pear. We call tests of this kind "item conceptual specificity". This design allowed for a direct test of *contingency* between generalization and various kinds of memories. We focused on the developmental window between age 3 and 8 years old to cover a crucial period of memory development (Newcombe et al., in press).

Important to the question of a generalization-specificity contingency is the treatment of episodic memory as a multifaceted construct. That is, to better characterize episodic memory capacities, we considered both context binding and two kinds of pattern separation processes that support different aspects of an episodic memory. Context binding may be especially relevant to episodic memory as it creates the spatiotemporal structure of a specific episode, whereas pattern separation is important for reducing interference when retrieving specific items or item-context associations in the presence of other similar memories. Compared to context binding, memory for individual items (individual objects, backgrounds, or facts) is thought to develop earlier (e.g., Riggins, 2014; Sluzenski et al., 2006). However, these studies did not specifically tax pattern separation processes that support memory specificity, as the study and test lists consisted of dissimilar items. On the other hand, studies that have specifically aimed to examine pattern separation development have predominantly used individual objects (Canada et al. 2019, Kerezstes et al., 2017; Ngo et al., 2018). However, as noted above, these studies did not distinguish between interference at the conceptual versus perceptual features of a given objects, as lure items are almost always perceptually similar exemplars (e.g., reviewed in Liu, Gould, Coulson, Ward, & Howard, 2016). To better understand which aspects of episodic memory that may contribute to generalization, we tested item memory specificity for conceptual and perceptual features separately. In sum, we operationalized episodic memory specificity as memory for the context in which an event occurred (i.e., context binding) and memory for the specific details of the conceptual and perceptual features of an item (item conceptual specificity and item perceptual specificity).

2. Methods

2.1. Participants

A total of 32 younger children (15 females; 17 males; M_{month} = 57.63 ± 7.33, range = 36-70) and 38 older children (25 females; 13 males; 86.24 ± 8.46, range = 72-101) recruited from the Philadelphia and the surrounding suburbs participated in the study. All recruited children were free of color blindness and psychological, neurological, and developmental disorders as reported by a parent. Informed consent was obtained from the child's parent. Six additional children participated but were not included in the analyses due to incomplete procedure (*n*=3) or failure to understand the task procedure based on a screening procedure (*n*=3; 2 3-year-olds and 1 4-year-old; see section SI 1.2). The young adult sample consisted of 29 undergraduate students (18 females; 11 males; M_{age} = 20.07 ± 1.65; range =18–24) from Temple University. Young adults gave informed consent and reported having normal or corrected-to-normal vision. All children were given a small toy for their participation, except for those tested virtually (see section 2.4). All young adults were given partial course credit. This experiment was approved by the Temple University Institutional Review Board committee.

2.2. Overall Procedure

The procedure was identical for children and young adults. In addition to the memory task, children were administered the verbal portion of the Kaufman Brief Intelligence Test, second edition (KBIT-2), whereas young adults were given the American National Adult Reading Test

(AMNART; Grober & Sliwinski, 1991), as measures of general verbal skills. One child was not administered the KBIT due to fatigue.

2.2.1. Memory Task

Materials. Cartoon images of 20 unpopular and androgynous characters, 80 scenes, and 180 black-and-white, line-drawn objects were selected from the Google Image search engine. Unpopular and androgynous characters were used to reduce the probability of children having pre-existing semantic knowledge—including gender stereotypes—about the characters. Twenty categories of semantically congruent objects and scenes were chosen based on their probable familiarity to young children (e.g., musical, cooking, and medical instruments). Each character was arbitrarily assigned to a category (e.g., Luntik was assigned to musical instruments). Each character was placed in four different scenes to create four encoding trial images for that character. All four scene images paired with a given character were semantically congruent with the character's assigned category (e.g., Luntik was placed in four perceptually-distinct performance halls; see Figure 1D). The 180 objects were chosen such that there were nine distinct objects for every category (e.g., the musical instrument category consisted of a guitar, a piano, a drum, a trumpet, etc.). Every line-drawn object was manually painted with three distinct colors using Photoshop, which resulted in a total set of 720 object images from the original set of 180 objects. An additional three characters, nine backgrounds, and 17 objects were selected from Google Images to use in the training phase and as an example trial (see SI). These additional stimuli were semantically unrelated to those used for the study and test phases of the experiment.

Procedure. All participants were tested individually. The experiment was divided into two encoding-test blocks with nonoverlapping stimuli between the two blocks. Each block consisted of a character familiarization, an encoding phase, and a test phase. The test phase consisted of four tasks described below (see Figure 1A).

Character introduction. All participants were first told that they would be introduced to some new friends. We presented images of each character sequentially and in a randomized order. On each trial, the name of the character was presented on the top of the screen (e.g., “This is Luntik”), and the experimenter read aloud their names (e.g., “This is Luntik,” “This is Doraemon,” etc.). There were 10 characters per block (see Figure 1B).

Encoding. Participants were told that each friend was making a collection of their favorite things, and that each friend would go to different places to find things to add to their collection. Participants were informed that they should pay attention to see what each of their friends like. The encoding phase consisted of 40 trials, where each consisted of an image of a character in a context presented on the left side of the screen and an object presented on the right side of the screen (5s, 0.5s ITI). Every character appeared in four encoding trials, each time in a different context and paired with a different object. Critically, the context and objects paired with a given character were semantically related. For instance, Luntik—a character assigned to the musical instrument category—was seen in different performance halls and collected objects such as a drum, a guitar, a horn, and an accordion. The order of the encoding trials was randomized across participants, with the only restriction being that the same character would not appear in more than two consecutive trials (see Figure 1C).

Test. The test phase immediately followed the encoding phase and consisted of four self-paced three-alterative-forced-choice tasks. The tasks included: (1) generalization, (2) context binding, (3) item conceptual specificity, and (4) item perceptual specificity. These were administered in a fixed order across participants (see Figure 1E). Each task consisted of 10 trials (one trial per character) presented in a randomized order across participants.

Generalization. Every test trial showed a character at the top of the screen and three objects at the bottom of the screen. Participants were asked to choose one object that this friend would add to their collection. All three objects were novel items that did not appear in the encoding phase. One object was the target—the correct item that belonged to the category assigned to that character. The other two objects were lures—objects that belonged to different semantic categories assigned to two other characters. Target selection would indicate that participants successfully made a novel inference based on the related episodes associated with a given character.

Context Binding. Every test trial showed an image of a character in one of that character's four encoding contexts at the top of the screen and three objects at the bottom. Participants were asked to choose the object that that friend had found in that particular place. All three objects were seen with the character at encoding. One object was the target—the correct item that was seen with the character in that particular context. The other two objects were lures—objects that were seen with the character, but that were paired with that character in different contexts. Target selection would indicate that participants remembered the specific object-context co-occurrence.

The *item conceptual specificity* and *item perceptual specificity* tasks were linked, such that the item perceptual specificity trial immediately followed the item conceptual specificity trial for each character. Every *item conceptual specificity* test trial showed a character at the top of the screen and three line-drawn objects at the bottom of the screen and were asked, "Which one of these three things did this friend find for their collection earlier?" All three objects belonged to the same category assigned to the character (e.g., musical instruments for the Luntik trial). One object was the target—the correct item that had appeared at encoding. The other two objects were lures—

objects that belonged to the semantic category assigned to the character, but that never appeared at encoding. In the presence of conceptually similar lures, target selection would indicate that participants remembered the objects' identities with high specificity. Critically, all three objects were presented as the color-stripped line drawn versions because we subsequently tested participants' memories for the perceptual details of the objects.

If the participants correctly selected a target, the phrase, "That's right!" would appear on the screen for 2s and was read aloud by the experimenter. On trials in which the participant correctly selected the conceptual target, the *item perceptual specificity* test trial for the same character immediately followed. If the participant incorrectly selected one of the lures, corrective feedback was provided by a green circle surrounding the target, and the experimenter said, "You actually saw *this* object earlier." Then the item perceptual specificity test trial for that character followed. The rationale for providing feedback on the conceptual trials was to ensure that we would have an equal number of valid test trials on the item perceptual specificity task. Once participants advanced to the item perceptual specificity trial, they were shown the same character from the preceding item conceptual specificity trial, with three object images presented at the bottom of the screen. One object was a target—the identical object to the one that appeared at encoding. The other two objects were lures—similar exemplars of the target that differed in color. In the presence of perceptually similar exemplars, target selection would indicate that participants remembered the objects' perceptual attributes with specificity.

All nine objects in each category were randomly assigned as encoding items (4 objects), generalization target (1 object), generalization lures for other categories (2 objects), or item conceptual specificity lures (2 objects) across participants. All objects were fully counterbalanced such that they never appeared twice in the test phase. The procedure was repeated twice but with entirely different sets of categories, characters, contexts, and objects. This resulted in a total of 20 characters and categories, 80 encoding trials, and 20 test trials per task (80 test trials) in total. The order of the two encoding-test blocks was counterbalanced across participants. The task lasted approximately 35 minutes.

2.3. Verbal Skills

All children were administered the Kaufman's Brief Intelligence Test, 2nd edition (KBIT-2: Kaufman & Kaufman, 1990) to assess general verbal intelligence. Children were instructed to point to one of six images simultaneously shown on a page that was the best match for a word or phrase (e.g., "which of these lives in a forest?" — a picture of a deer), and to respond with a one-word answer to verbal riddles (e.g., "what can only be seen at night and twinkles in the sky?"— "star", "moon"). The test, with increasing levels of difficulty in each section, was terminated when a child provided incorrect responses in four consecutive trials.

2.4. Virtual testing procedure

Among the 70 children who participated in the study, 12 were administered the memory task and verbal skills task virtually via Zoom due to the COVID-19 pandemic. For the virtual testing format, we instructed participants' parents to set up either a desktop or laptop at children's eye level and test their internet connection. The experimenter shared their own screen with the participant such that the participant would view the screen in the same manner as participants who were tested in person. At test, when participants made memory judgments by pointing to one of the options in the 3AFC test, participants' parents were instructed to say, "left", "middle", or "right" to indicate to the experimenter which option the child had selected. Parents were specifically instructed to not name objects that appeared in the experiment, and to refer only to their relative position on screen.

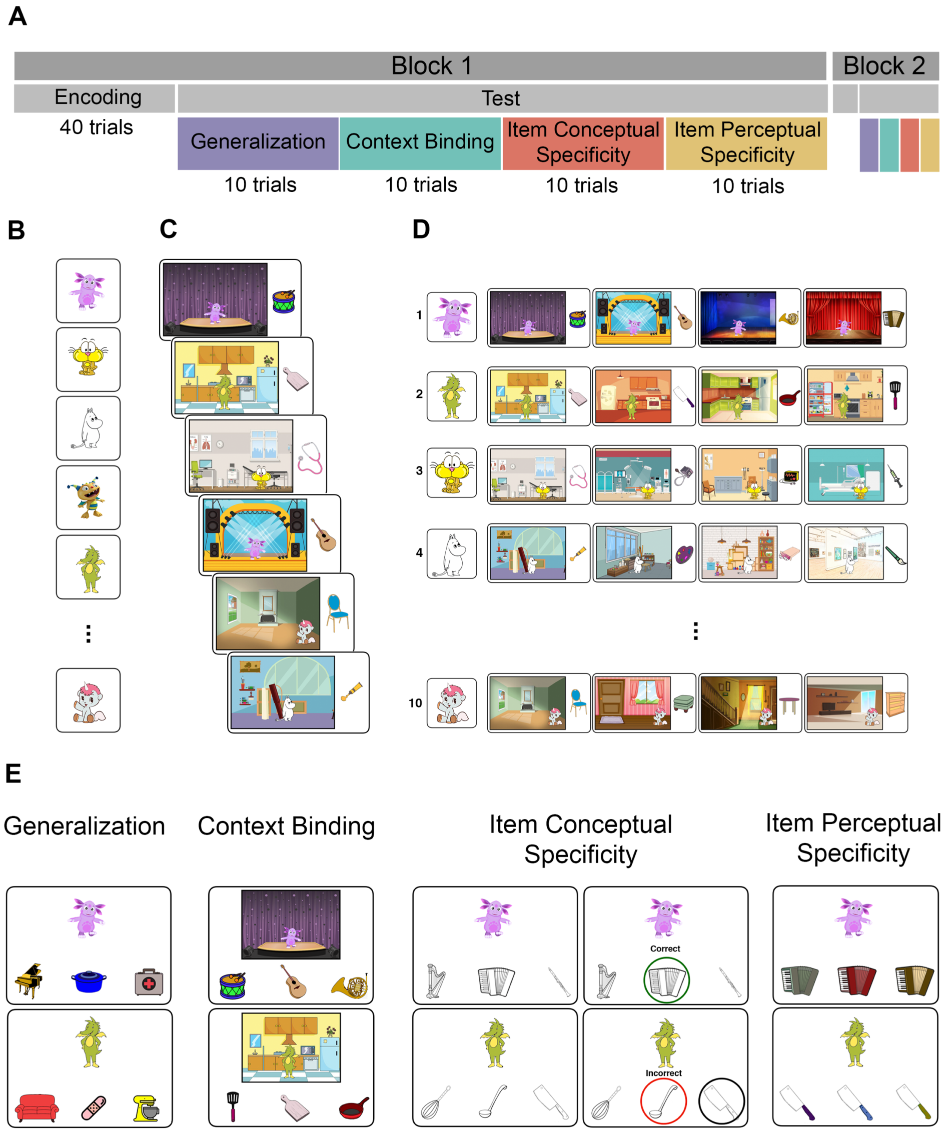

Figure 1. A schematic depiction of the memory task, including the overall experimental procedure (A), the character introduction phase (B), the encoding phase (C), the character-category assignment (D), and the test phase (E).

3. Results

Overall accuracy, collapsed across four tasks, did not differ between Blocks 1 and 2, *t*(98)= -1.49, *p*= .14, or between males and females, *t*(97)= -0.03, *p*= .98. Verbal skills as measured by KBIT and AMNART for did not correlate to any of the task performances in younger children (all *ps*>.31), older children (all *ps*> .09), or adults (all *ps* > .24).

3.1. Age-performance relation in children

First, we tested whether performance on each task differed by age in children by conducting Pearson correlations between age (in months) and task performance. We defined accuracy as the proportion of trials in which the targets were selected in 3 alternative forced choice (AFC) tasks. We found that among children, age was positively correlated with performance on the generalization task, *r(*68)= 0.46, *p*< .001, item conceptual specificity task, *r(*68)= 0.48, *p*< .001, item perceptual specificity task, *r(*68)= 0.35, *p*= .003, and showed a trend towards significance with performance on the context binding task, *r(*68)= 0.21, *p*=.09.

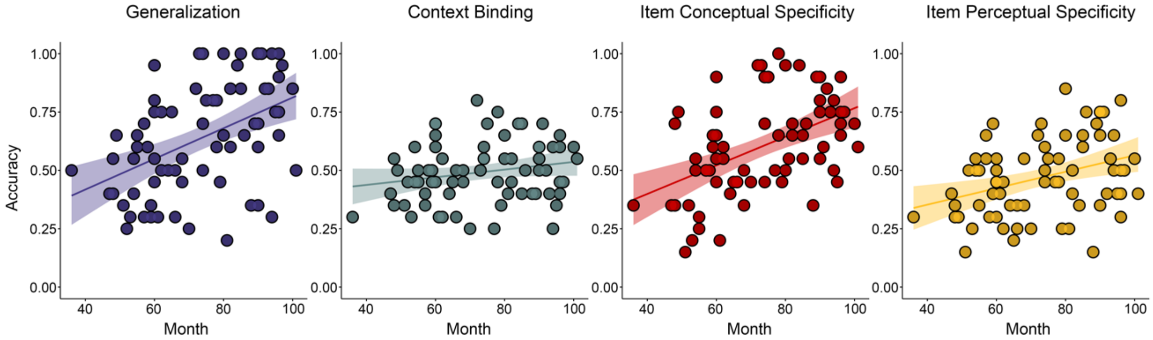

Figure 2. Scatterplots of accuracy (y-axes) and age in months (x-axes) for each task in children.

3.2. Age-related differences among age groups

To test whether memory performance varied with age through early childhood into young adulthood, we separated our child sample into younger children (aged, 3-5), and older children (aged 6-8). We conducted a mixed 3 (age groups) x 4 (tasks) ANOVA and found a main effect of age, *F*(2, 96)= 23.16, *p*< .001*,* partial η^2 = 0.33, a main effect of task, *F*(3, 288)= 57.21, *p*< .001*,* partial η^2= 0.37, and a significant interaction, *F*(6, 288)= 7.07, *p*< .001*,* partial η^2 = 0.13. Tukey post-hoc tests showed similar age patterns for the generalization, item conceptual specificity, and item perceptual specificity tasks. For generalization, older children and young adults did not differ from each other, *p*= .27, but both groups were better than younger children at generalization, *ps*< .001. For item conceptual specificity, older children and young adults performed similarly, *p*= 1.00, and both groups were better able to remember the item identities of the learned objects compared to younger children, *ps*< .001. For item perceptual specificity, again older children and young adults did not differ, *p*= 1.00, but they were better at remembering the objects' perceptual attributes compared to their younger counterparts, *ps*< .04. Surprisingly, there were no age-related differences in context binding performance among the three age groups, *ps*> .82. It is also important to note that even the youngest group of children performed above chance level (0.33) on all four tasks, *ps*< .02.

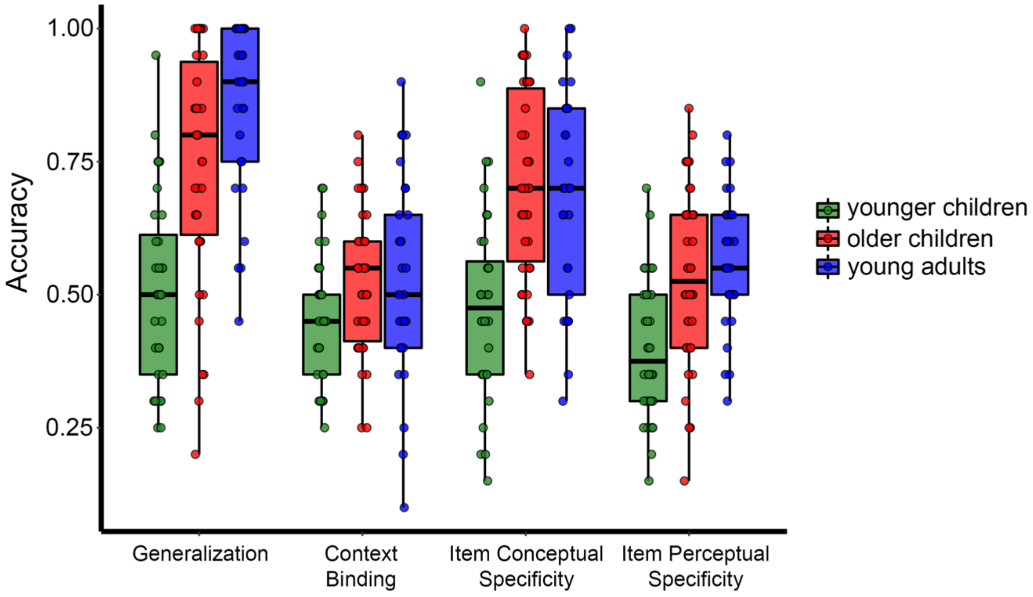

Figure 3. Distributions of participants' accuracy on each task for younger children (aged 3-5), older children (aged 6-8), and young adults.

3.3. Within-category semantic similarity

The generalization task in our paradigm was operationalized as the ability to make novel inferences based on character-category mapping through related episodes. Thus, we reasoned that semantic similarity among items within a category should be associated with generalization accuracy for a given category, such that learning items that are closer in semantic space should promote generalization success. To test this idea, we used Global Vectors for Word Representation (GloVe; Pennington, Socher, & Manning, 2014) to estimate the semantic similarity between the items in our stimulus set. GloVe is a vector space model that can be “trained” on a particular corpus of words by building a co-occurrence matrix and predicting the total co-occurrences between a target word and a context word. The premise of this approach is that the co-occurrence statistics between two words from large bodies of texts should reflect their semantic relationship. We used a pre-trained word vector on 42 billion tokens of web data (Common Crawl) which contains 1.9 million vocabularies to estimate the semantic similarity between every pair of items within a category in our stimulus set. To yield the semantic similarity score, we calculated a cosine similarity score ranging from -1 to 1 between every pair of words, with greater values denoting higher similarity between two words (see Figure 4A; further description of within-category semantic clustering, see SI 2.1). Given that for each category, we randomly assigned four items that appeared at encoding, and one generalization target at test, we computed a subject-specific semantic similarity score among these five items per category by averaging across 10 semantic similarity scores (10 pairwise among five items) for each participant (see Figure 4B).

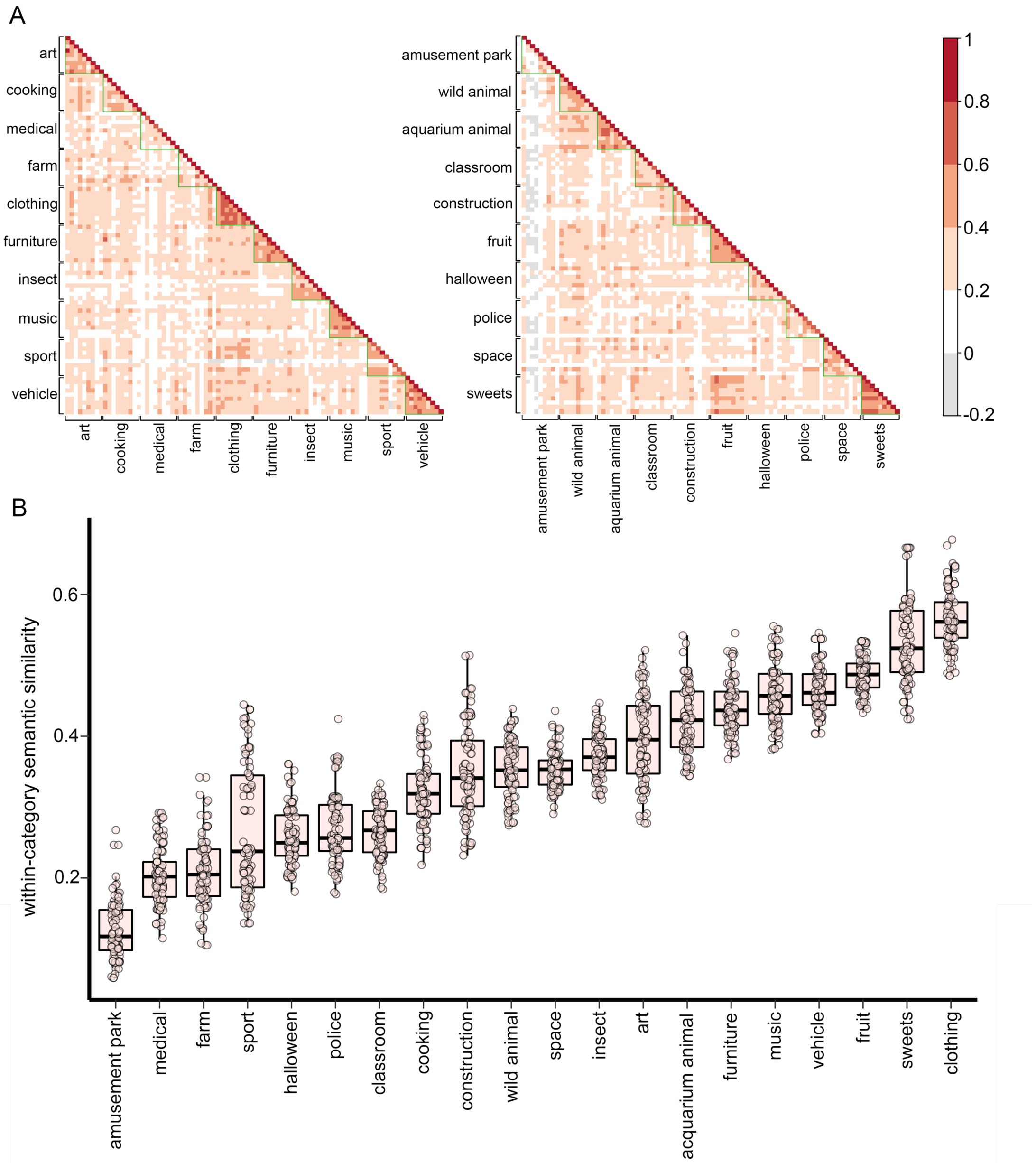

Figure 4. A matrix of all pairwise similarity scores calculated from Global Vectors for Word Representation (GloVe) for the 90-item stimuli set in blocks A (A, Left) and B (A, Right). Every cell represents an inter-item semantic similarity score, with darker colors representing higher scores. A distribution of participant-specific semantic similarity scores, defined as the mean of the four encoding items and a generalization target for each category (B). Higher similarity scores indicate that the items within a given category were closer to one another in semantic space (more similar). Every dot represents a participant.

3.4. Generalization-Episodic Specificity Contingency

A primary question of this research is whether generalization depends on episodic specificity and whether the contribution of episodic memory specificity to generalization ability depends on age. Further, the generalization task requires memory for the character-category mapping in order to successfully make inferences about each character's "collection," i.e. to generalize about that character's category. Thus, we reasoned that the degree of semantic relatedness among the various items seen with each character across different episodes may promote generalization success. We asked what aspects of episodic memory specificity would predict generalization success and whether within-category semantic similarity would promote generalization, for three age groups separately. We conducted a generalized linear mixed effects model with context binding accuracy, item conceptual specificity accuracy, item perceptual specificity accuracy, and semantic similarity as fixed effects, and participant and category as random effects, to predict generalization success on a trial-by-trial basis. Given that each participant contributed to multiple memory tasks, we modeled the non-independent binary outcome of generalization (successful or unsuccessful) response conditional on the attributes of each participant and each category by adding them to the models as random effects.

For younger children, item conceptual specificity, β= 0.44, *SE*= 0.18, *z*= 2.48, *p*= .01, and semantic similarity, β= 1.71, *SE*= 0.84, *z*= 2.05, *p*= .04, significantly predicted generalization success. Context binding accuracy, β= -0.09, *SE*= 0.17, *z*= -0.51, *p*= .61, was not associated with generalization success, but there was a trend towards significance for item perceptual specificity accuracy, β= -0.32, *SE*= 0.18, *z*= -1.80, *p*= .07, in the direction that inaccurate item perceptual specificity was coupled with higher probability of generalization success (see Figure 5A). These results suggest that remembering the specific item identities and a greater degree of semantic relatedness within a category led to a greater likelihood of generalization success in younger children. Importantly, a reduced model that included only semantic similarity and item conceptual specificity as fixed effects (AIC= 862.11, BIC= 884.42) did not significantly differ from the full model (AIC= 862.59, BIC= 893.82), $\chi^2(2)= 3.52$, $p= .17$, suggesting that adding context binding and item perceptual specificity accuracy into the model did not improve model fit for the data.

For older children, neither context binding accuracy, $\beta= -0.13$, $SE= 0.20$, $z= -0.67$, $p= .50$, item conceptual specificity accuracy, $\beta= 0.10$, $SE= 0.22$, $z= 0.45$, $p= .66$, item perceptual specificity accuracy, $\beta= 0.18$, $SE= 0.20$, $z= 0.88$, $p= .38$, nor semantic similarity, $\beta= 0.85$, $SE= 0.95$, $z= 0.89$, $p= .37$, were significantly associated with generalization success (see Figure 5B).

For young adults, context binding accuracy was significantly associated with generalization success, $\beta= .66$, $SE= 0.27$, $z= 2.43$, $p= .01$, whereas item conceptual specificity, $\beta = .29$, $SE= 0.27$, $z= 1.06$, $p= .29$, item perceptual specificity, $\beta= 0.03$, $SE= 0.26$, $z= 0.13$, $p= .90$, and semantic similarity, $\beta= 1.48$, $SE= 1.17$, $z= 1.27$, $p= .20$, were not (see Figure 5, bottom). A reduced model that only included context binding accuracy (AIC= 470.55, BIC= 488.01) did not differ from the full model (AIC= 473.95, BIC= 504.49), $\chi^2(3)= 2.61$, $p= .46$, in predicting the generalization success probability in young adults (see Figure 5C).

To test the stability of these findings with regards to semantic similarity, we also used Latent Semantic Analysis (LSA) to estimate the semantic similarity scores and repeated the analyses. The findings were consistent with those reported above (see SI 2.2).

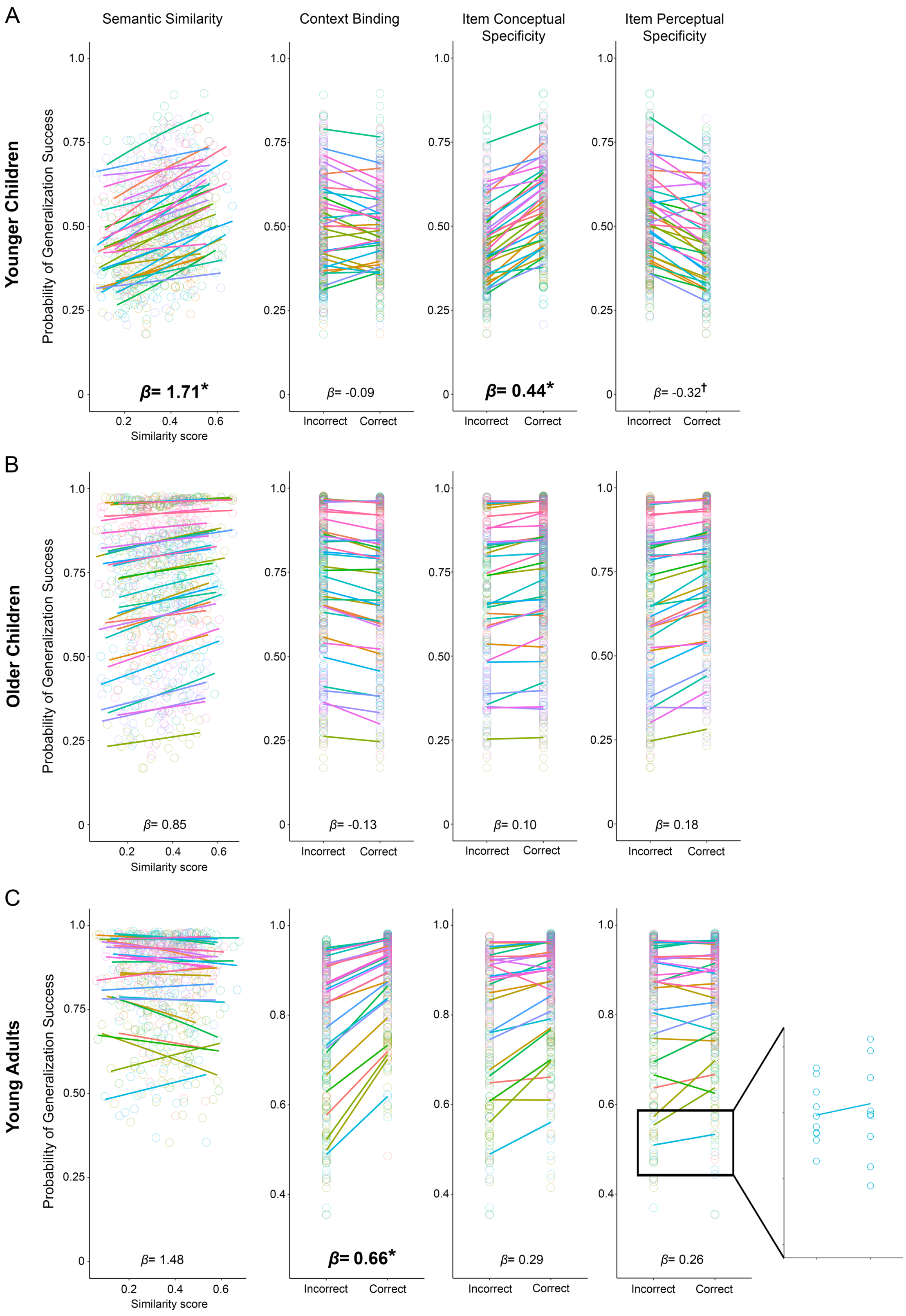
A
Semantic Similarity
Context Binding
Item Conceptual Specificity
Item Perceptual Specificity
Younger Children
Probability of Generalization Success
1.0
0.75
0.50
0.25
0
β= 1.71*
β= -0.09
β= 0.44*
β= -0.32†
0.2
0.4
0.6
Similarity score
Incorrect
Correct
B
Older Children
β= 0.85
β= -0.13
β= 0.10
β= 0.18
C
Young Adults
0.8
0.6
0.4
β= 1.48
β= 0.66*
β= 0.29
β= 0.26

Figure 5. Distribution of the estimated probability of generalization success (*y*-axes) by each fixed effect from the generalized linear mixed effects model in younger children (A), older children (B), and adults (C) (see Results 3.4.). For semantic similarity, participant-specific semantic similarity scores are plotted on the x-axes. For context binding, item conceptual specificity, and item perceptual specificity, correct and incorrect trials are plotted on the *x*-axes. Each colored line represents an individual participant; each dot denotes an individual trial. The black box (bottom panel, right) aims to illustrate a given participant's set of trials and the mean estimated probability of generalization success for that participant predicted by a given task (intended for schematic visualization purposes only). Significance notation: † $p< .08$, * $p< .05$.

4. Discussion

We tested generalization and various kinds of episodic memory specificity in tandem in children aged 3-8 and in young adults. The younger children generalized less than their older counterparts and remembered fewer perceptual and conceptual details about items than their older counterparts. Crucially, generalization by young children was contingent on memories of the specific item identity and the conceptual common ground that linked together related episodes, whereas generalization by young adults was contingent on memory for specific item-context linkage.

4.1. Age patterns in generalization and episodic memory specificity

First, generalization performance increases from early to middle childhood. This age effect on generalization aligns with some previous findings that show a protracted improvement on other paradigms that tap the extraction of statistical regularities among items within a continuous stream of visual stimuli (Arciuli & Simpson, 2011; Pudhiyidath, Roome, Coughlin, Nguyen, & Preston, 2019; Schlichting, Guarino, Schapiro, Turk-Browne, & Preston, 2017; Shufaniya & Arnon, 2018, but see Finn, Kharitonova, Holtby, & Sheridan, 2018).

We also found robust age effects in remembering the specific conceptual and perceptual attributes of objects: younger children were less able to remember the identity and the perceptual details associated with the objects compared to older children and adults. The result that item conceptual specificity improves with age adds important clarification to the literature. First, item memory has shown little age-related differences from early to middle childhood (Lloyd, Doydum, & Newcombe, 2009; Sluzenski et al., 2006), but perhaps this age pattern applies to situations in which conceptual interference among items is low. Second, studies that examined false memory for conceptually related lures in the Deese-Roediger-McDermott paradigm have shown that false memory increases with age (Brainerd, Reyna, & Forrest, 2002; Carneiro, Albuquerque, & Fernandez, 2009; but see Ghetti, Qin, & Goodman, 2002). Here, in this paradigm, we found the opposite pattern: conceptual specificity sharpens with age, although there are notable differences between these paradigms (e.g., differences in list study sizes, recall vs. recognition).

The age patterns on the item perceptual specificity task are consistent with previous work showing that, with age, children's ability to remember perceptual details of objects with high granularity improves (ages 4-8: Canada et al., 2019; Ngo, Newcombe, & Olson, 2019). Past research on pattern separation has primarily used stimuli of similar object exemplars (e.g. similar rubber ducks; reviewed in Liu et al., 2016), invoking interference between overlapping items spanning both the conceptual and perceptual dimensions. However, parsing the different sources of item-level memory specificity is important for charting memory development.

Together, our results suggest that crucial developments in generalization ability and episodic memory capacities span the transition from early to middle childhood. On the other hand, the findings regarding age and context binding were weaker. The relation between age and context binding in children only showed a trend towards significance, and there were no significant age group differences among younger children, older children, and young adults. Previous studies have consistently reported age-related improvements in context binding or relational binding in general throughout early and middle childhood (Lee et al., 2020; Riggins, 2014; Sluzenski et al. 2006). We speculate that the nonsignificant age effects in our paradigm are possibly due to the unusually higher number of item-context associations that appeared at encoding, together with the strong semantic congruency within a set of item-context pairs associated with each character, leading to a particularly challenging task. Another reason could be that participants learned about a given character in an interleaved fashion. Interleaved learning is thought to be beneficial for generalization by increasing between-category discriminability, whereas blocked learning may improve specificity and learning of details (Birnbaum, Kornell, & Bjork, & Bjork, 2013; Kornell & Bjork, 2008). Perhaps the interleaved learning trials in our design dampened adults' memory for specific item-context pairs to a greater extent. Further, children were exposed to all four types of questions prior to encoding in the 'screening' procedure, whereas young adults were not, which may have dampened the age-related differences in context binding. These factors may have amplified the level of interference and blunted young adults' performance on context binding via altering their encoding strategies.

4.2. Generalization-specificity contingency

Crucially, we showed that the contingency between episodic memory specificity and generalization is not homogenous across development. With our design, different aspects of episodic memory were targeted, so we were able to test *which* aspects of a past episode showed contingency with generalization. The different generalization-specificity contingency patterns across development reveal that different aspects of episodic memories may be used to make generalization judgments, likely depending on the neurodevelopmental status of the participants.

For younger children, the probability of generalization success was positively associated with memory for objects' conceptual specificity and the degree of within-category semantic relatedness. These findings suggest that early on in life, the conceptual common ground that links together related episodes is important for generalization success, suggesting a role for pre-existing semantic memory in facilitating generalization performance. Further, accurate item conceptual specificity was associated with greater likelihood of generalization success. The direction of this contingency has important implications on interpreting *how* generalization was supported in younger children. One possibility is that generalization arises from abstraction—a process by which memories for the specific instances are lost, but the emergent averaged representation can support generalization across episodes (discussed in Altmann, 2017). Our findings did not show a generalization-specificity trade off: the preservation of item conceptual specificity yielded higher probabilities of generalization success. These results suggest that in early childhood, children were able to integrate the overlapping elements across episodes– not abstraction – that allows rapid generalization.

Unlike young children, young adults' generalization success was tied to remembering the idiosyncratic item-context bound representations. Memory for the specific what-where relational structure is one key signature of episodic memory capacity (Johnson, Hashtrodi, & Lindsay, 1993; Tulving, 1972). These findings are consistent with the notion that rapid generalization may indeed rely on retrieving specific instances, as posited by previous work in young adults (Banino et al., 2016; Mack, Love, & Preston, 2018). In older children, none of the variables significantly predicted generalization success. It is likely that the sources of generalization are diffuse such that any given variable's predictability is not individually robust. Therefore, older children might be considered "intermediate" between younger children and young adults.

4.3. Multiple routes to generalization

Based on our findings, we suggest that there may be multiple routes to acquire what the literature refers to as gist or schemas, or what we call generalized memories. Importantly, different routes dominant at different points in development. In a mature system, rapid generalization could occur on-the-fly through integrating and formatting a network of related experiences. Here we showed that the preservation of rich contextual memories plays a role in generalization in young adults, as expected. However, in an immature system, we did not see this link. Instead, there is a reliance on specific instances at the level of conceptual specificity of individual items, and on the conceptual link among these items to promote generalization success. It is possible that without a full constellation of robust episodic memory capacities, younger children rely on the aspects of a specific episode that they are able to encode and retain, along with the support of overall semantic structures that tie together the related episodes. In contrast, successful generalization in young adults is contingent on the successful retrieval of the idiosyncratic contextual aspects of past episodes. These ideas fit well with previous findings showing that memory for individual items develops much earlier than item-context or item-item relational memory (Riggins, 2014; Sluzenski et al., 2006). Even in our youngest group of children, there is an aspect of episodic memory specificity at the item level that is tied to generalization success, namely conceptual specificity, suggesting a certain kind of contingency between semantic memory acquisition and episodic memory. This finding aligns with the notion that there is a degree of inter-independence between semantic memory acquisition and episodic memory (SPI; Tulving, 1995; Tulving & Markowitsch, 1998).

Neural bases of generalization and episodic memory capacities

The hippocampus and prefrontal areas participate in episodic memory specificity (Preston & Eichenbaum, 2013). Several models have posited that the dentate gyrus and CA3 subfields of the hippocampus are especially involved in coding individuated memories (Norman & O'Reilly, 2003; Schapiro, Turk-Browne, Botvinick, & Norman, 2017). Aligned with these ideas, development of the late-maturing subfields including the dentate gyrus and CA3 is associated with relational binding (Riggins et al., 2018) and pattern separation processes (Canada et al., 2019; Keresztes et al., 2017). Age differences in the structure (Sowell, Delis, Stiles, & Jernigan, 2001) and functional recruitment (Selmeczy, Fandakova, Grimm, Bunge, & Ghetti, 2019) of the prefrontal cortex have also been linked to memory improvements in late childhood and adolescence.

In human adults, the basal ganglia have been shown to be involved in various forms of statistical learning (Karuza et al., 2013; Poldrack et al., 2001; Turk-Browne, Scholl, Chun, & Johnson, 2009). More recent research has shown that the hippocampus also contributes to the integration of related events to form new generalizable memories using various paradigms including acquired equivalence (Shohamy & Wagner, 2008), concept learning (Bowman & Zeithamova, 2018; Kumaran, Summerfield, Hassabis, & Maguire, 2009), and associative inference (Zeithamova, Dominick, & Preston, 2012). In associative inference, the hippocampus contributes to generalization by interacting with the ventromedial prefrontal cortex to integrate episodes with shared elements (Schlichting, Mumford, & Preston, 2015). Evidence is mixed as to whether hippocampal dysfunction disrupts statistical learning and generalization. Some studies have found that hippocampal damage is linked to a decrease in statistical learning proficiency in humans (e.g., Covington, Brown-Schmidt, & Duff, 2018) and in generalization abilities in rodents (Montgomery et al., 2016). However, other studies have shown no impairment (reviewed in Ashby & Rosedahl, 2017). From the developmental literature, gray matter volumes in the hippocampus and prefrontal structures are correlated with statistical learning in children aged 5-8 (Finn et al., 2018), and specifically with the hippocampal head in children aged 6-14 (Schlichting et al., 2017). It is likely that the development of the hippocampus and its connections to the mPFC subserve the behavioral gains in generalization and episodic memory specificity in infancy and childhood.

An important question to be tested is whether generalization relies on different neural substrates at different stages of neural development. Some investigators have suggested that infants may rely on the early-developing monosynaptic pathway linking the entorhinal cortex to CA1 to perform fast generalization (Gómez & Edgin, 2016; Schapiro et al., 2017). The notion of uneven maturational rates between intrahippocampal pathways and improvements on tasks tapping into fast generalization well beyond middle childhood may indeed suggest that generalization relies on different mechanisms in infancy, in later stages of childhood development, and in adulthood (Newcombe et al., in press). Specifically, some forms of statistical learning could rely on the monosynaptic pathway early on in life, whereas inference-based generalization may further recruit the whole hippocampal circuitry and its coordination with the prefrontal cortex later in life.

9.6. Limitations

One limit of employing a cross-sectional design is that it prevents us from understanding the developmental *changes* of generalization and specificity. Charting the potential lead-lag between the two memory functions would further elucidate the dependency between semantic memory acquisition and episodic memory specificity.

One other limitation of this work is the utility of real-world objects and existing pre-categories, which could have introduced age-related differences in categorical knowledge. However, the utility of real-world objects enabled our design to assess the role of semantic clustering in generalization. Nonetheless, future research should investigate whether extracting statistical regularities between pseudo objects versus real-world objects share strong behavioral co-variance in children.

9.7. Conclusions

Our study reveals how the intricate interaction between two fundamental capacities of human memory may dynamically unfold over the course of development. Critically, our findings substantiate the notion that there may be multiple routes to inference-based generalization. That is, different aspects of episodic memories and pre-existing conceptual knowledge support novel semantic memory acquisition in children versus in adults. This developmental phenomenon has important implications for contemporary models of memory.

References

Altmann, G. T. (2017). Abstraction and generalization in statistical learning: Implications for the relationships between semantic types and episodic tokens. *Philosophical Transactions Royal Society B, 372*: 20160060.

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental science, 14*(3), 464-473.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review, 124,* 472-482.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64*(3), 711-728.

Banino, A., Koster, R., Hassabis, & Kumaran (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports, 6*: 31330.

Bauer, P. J., & Dow, G. A. A. (1994). Episodic memory in 16-and 20-month-old children: Specifics are generalized, but not forgotten. *Developmental Psychology, 30,* 403 – 417.

Bauer, P. J., San Souci, P. (2010). Going beyond the facts: young children extend knowledge by integrating episodes. *Journal of Experimental Child Psychology, 107*(4), 452-465.

Bindschaedler, C., Peter-Favre, C., Maeder, P., Hirsbrunner, T. & Clarke, S. (2011). Growing up with bilateral hippocampal atrophy: From childhood to teenager. *Cortex, 47*(8), 931-944.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive reasoning: The roles of discrimination and retrieval. *Memory & Cognition, 41,* 392-402.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology, 38*(6), 948-957.

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience, 38*(10), 2650-2614.

Brainerd, C. J., Reyna, V., & Forrest, T. J. (2002). Are young children susceptible to the false-memory illusion? *Child Development, 73*(5), 1363-1377.

Canada, K. L., Ngo, C. T., Newcombe, N. S., Geng, F., & Riggins, T. (2019). It's all in the details: Relations between young children's developing pattern separation abilities and hippocampal subfields. *Cerebral Corte*x, 29(8), 3427-3433.

Carneiro, P., Albuquerque, P., & Fernandez, A. (2009). Opposite developmental trends for false recognition of basic and superordinate names. *Memory, 17*(4), 411-427.

Clark, E. (2001). Emergent categories in first language acquisition. In M. Bowerman & S. Levinson (Eds.), *Language Acquisition and Conceptual Development* (Language Culture and Cognition, pp. 379-405). Cambridge: Cambridge University Press.

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The necessity of the hippocampus for statistical learning. *Journal of Cognitive Neuroscience, 30*(5), 680-697.

Finn, A., Kharitonova, M., Holtby, N., & Sheridan, M. A. (2018). Prefrontal and hippocampal structure predict statistical learning ability in early childhood. *Journal of Cognitive Neuroscience, 31(*1), 126-137.

Gardiner, J. M., Brandt, K. R., Baddeley, A. D., Vargha-Khadem, F., & Mishkin, M. (2008). Charting the acquisition of semantic knowledge in a case of developmental amnesia. *Neuropsychologia, 46*(11), 2865-2868.

Ghetti, S., & Bunge, S. A. (2012). Neural changes underlying the development of episodic memory during middle childhood. *Developmental Cognitive Neuroscience, 2*(4), 381-395.

Ghetti, S., Qin, J., & Goodman, G. S. (2002). False memories in children and adults: Age, distinctiveness, and subjective experience. *Developmental Psychology, 38*(5), 705-718.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia,* 104-114.

Gómez, R. L., & Edgin, J. O. (2016). The extended trajectory of hippocampal development: Implications for early memory development and disorder. *Developmental Cognitive Neuroscience, 18*, 57-69.

Greenberg, D. L., Keane, M. M., Ryan, L., & Verfaellie, M. (2009). Impaired category fluency in medial temporal lobe amnesia: The role of episodic memory. *Journal of Neuroscience, 29,* 10900-10908.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13*(6), 933-949.

Johnson, M. K., Hashtroudi, S., & Lindsay, S. D. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28.

Karuza, E. A., & Newport, E. L., Aslin, R. N., Starlings, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language, 127*, 46-54.

Kaufman, A. S., & Kaufman, N. L. (1990). Kaufman Brief Intelligence Test. Circle Pines: MN: American Guidance Service.

Keresztes, A., Bender, A. R., Bodammer, N. C., Lindenberger, U., Shing, Y. L., & Werkle-Bergner, M. (2017). Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *PNAS, 114*(34), 9212-9217.

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 Science, 22*(8), 676-686.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "Enemy of induction"?. *Psychological Science, 19*(6), 585-592.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memory: A model of the hippocampal system. *Psychological Review, 119*(3), 573-616.

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63*(6), 889-901.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Lee, J. K., Fandakova, Y., Johnson, E. G., Cohen, N. J., & Bunge, S. A., & Ghetti, S. (2020). Changes in anterior and posterior hippocampus differentially predict item-space, item-time, and item-item memory improvement. *Developmental Cognitive Neuroscience, 41,* 100741.

Liu, K. Y., Gould, R. L., Coulson, M., Ward, E. V., & Howard, R. J. (2016). Tests of pattern separation and pattern completion in humans: A systematic review. *Hippocampus, 26*(6), 705-717.

Lloyd, M. E., Doydum, A. O., & Newcombe, N. S. (2009). Memory binding in early childhood: Evidence for a retrieval deficit. *Child Development*, *80*(5), 1321–1328. doi.org/10.1111/j.1467-8624.2009.01353.x

Lukowski, A. F., Wiebe, S. A., & Bauer, P. J. (2009). Going beyond the specifics: Generalization of single actions, but not temporal order, at nine months. *Infant Behavior and Development.*

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why are there complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419-457.

Newcombe, N. S., Balcomb, F., Ferrara, K., Hansen, M., & Koski, J. (2014). Two rooms, two representations? Episodic-like memory in toddlers and preschoolers. *Developmental Science*, *17*(5), 743–756.

Newcombe, N. S., Benear, S. L., Ngo, C. T., & Olson, I. R. (in press). Memory in infancy and childhood. In M. Kahana & A. Wagner (Eds.), *Oxford Handbook on Human Memory.*

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.

Ngo, C. T., Lin, Y., Newcombe, N. S., & Olson, I. R. (2019). Building up and wearing down episodic memory: Mnemonic discrimination and relational binding. *Journal of Experimental Psychology: General, 148*(9), 1463-1479.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at time: The hippocampus and concept formation. *Neuroscience Letters, 690*, 31-38.

Montgomery, K. S., Edwards III, G., Levites, Y., Myers, C. E., Gluck, M. A., Setlow, B. & Bizon, J. L. (2016). Deficits in hippocampal-dependent transfer generalization learning accompany synaptic dysfunction in a mouse model of amyloidosis. *Hippocampus, 26*(4), 455-471.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Presentation. *Empirical Methods in Natural Language Processing (EMNLP),* 1532-1543.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C. et al. (2001). Interactive memory systems in the brain. *Nature, 414*, 546-550.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology, 23*(17), 764-773.

Pudhiyidath, A., Roome, H., Coughlin, C., Nguyen, K. V., & Preston, A. R. (2019). Developmental differences in temporal schema acquisition impact reasoning decisions. *Cognitive Neuropsychology, 37*(1-2), 25-35.

Ramsaran, A. I., Schlichting, M. L., & Frankland, P. W. (2019). The ontogeny of memory persistence and specificity. *Developmental Cognitive Neuroscience, 36*: 100591.

Riggins, T. R. (2014). Longitudinal investigation of source memory reveals different developmental trajectories for item memory and binding. *Developmental Psychology*, *50*(2), 449–59.

Riggins, T. R., Canada, K. L., & Botdorf, M. (2020). Empirical evidence supporting neural contributions to episodic memory development in early childhood: Implications for childhood amnesia. *Child Development Perspectives, 14*(1), 41-48.

Riggins, T. R., Geng, F., Botdorf, M., Canada, K., Cox, L., Hancock, G. R. (2018). Protracted hippocampal development is associated with age-related improvements in memory during early childhood. *Neuroimage, 174,* 127-137.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 80*(274), 1926-1928.

Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (p.1-38). The MIT Press.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modeling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B, 372*(1711): 20160049.

Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B., & Preston, A. R. (2017). Hippocampal structure predicts statistical learning and associative inference abilities during development. *Journal of Cognitive Neuroscience, 29*(1), 37-51.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications, 6*: 8151.

Selmeczy, D., Fandakova, Y., Grimm, K., Bunge, S. A., & Ghetti, S. (2019). Longitudinal trajectories of hippocampal and prefrontal contributions to episodic retrieval: Effects of age and puberty. *Developmental Cognitive Neuroscience, 36,* 100599.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron, 60*(2), 378-389.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science, 42*(8), 3100-3115.

Sluzenski, J., Newcombe, N. S., & Kovacs, S. L. (2006). Binding, relational memory, and recall of naturalistic events: a developmental perspective. *Journal of Experimental Psychology: Learn, Memory, & Cognition, 32*(1), 89–100.

Sowell, E. R., Delis, D., Stiles, J. & Jernigan, T. L. (2001) Improved memory functioning and frontal lobe maturation between childhood and adolescence: a structural MRI study. *Journal of International Neuropsychological Society, 7*(3), 312-322.

Tompary, A., Zhou, W., & Davachi, L. (2020). Schematic memories develop quickly, but are not expressed unless necessary. *Scientific Reports, 10,* 16968.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving E. & W. Donaldson, *Organization of memory*. Academic Press.

Tulving, E. (1995). Organization of memory: Quo vadis? In: Gazzaniga MS, ed. The cognitive neurosciences. Cambridge, MA: MIT Press, 839-847.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1-25.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus, 8*(3), 198-204.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 195-202.

Varga, N. L., Stewart, R. A., & Bauer, P. J. (2016). Integrating across episodes: Integrating the long-term accessibility of self-derived knowledge in 4-year-old children. *Journal of Experimental Child Psychology, 145*, 48-63.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Paesschen, W. V., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 80*(277), 376-380.

Winocur, G., Moscovitch, M., Bontempo, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal-neocortical interactions. *Neuropsychologia, 48*(8), 2339-2356.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168-179.

Author Note

We thank Nadhia Engle, Rebecca Adler, Linda Hoffman, and Richard Ho for their assistance with data collection. We are grateful to Dr. James Antony for his advice on the Global Vectors for Word Presentation for the semantic similarity analyses. This work was supported by the National Institutes of Health to C. T. Ngo (F31HD090872), to I. R. Olson (R01 MH091113 and R21 HD098509), and to N. S. Newcombe (R01 HD099165), and by Temple University to N. S. Newcombe (OVPR 161706-24607-02). The content is solely responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contribution

C. T. Ngo and N. S. Newcombe developed the research questions. All authors contributed to the design of the experiment. C. T. Ngo, S. L. Benear, and H. Popal developed the stimuli. H. Popal implemented the task in Python. Data were collected by C. T. Ngo, S. L. Benear and others (see Author Note). Data analyses and result interpretation were led by C. T. Ngo, under the supervision of N. S. Newcombe and I. R. Olson. C. T. Ngo drafted the manuscript, and all authors provided critical revisions. All authors approved the final of the manuscript for submission.

Data Availability

All experimental materials and second-level data have been made publicly available through the Open Science Framework (https://osf.io/gv485/).

SI for Ngo, Benear, Popal, Olson, & Newcombe.

Methods

1.1. Stimulus materials

After initial stimulus selection, we gauged children's familiarity with the categories using a sorting task. Six children (5 4-year-old and 1 5-year-old) who did not participate in the main experiment participated. Line-drawn images of the objects were printed out on A4 papers and cut into small cards. For each block, 90 items were randomly separated into 10 decks of 9 items (1 item from each category per deck). One deck of items was randomly selected to serve as the reference deck and the cards were arranged horizontally on a table. Children were given the other 8 decks one at a time and asked to place each card under an item in the probe set where it best belonged. We repeated the same procedure with the second block. These children performed the sorting task with 100% congruency with our initial assignments.

1.2. Screening Phase

To acquaint children with the task and to ensure that we would only include children who understood the cover story of the "collection game", we administered a short mockup of the experiment with different stimuli from the main experiment. Children were introduced to a character named Gachapin. They were told that "Gachapin was making a collection of different kinds of vegetables and goes to different places to look for vegetables to add to his collection". We then presented 4 encoding trials, each showing Gachapin in a context (e.g., a garden) and paired with a vegetable (e.g., carrot). Different from the encoding phase in the main experiment, we showed 4 encoding trials simultaneously on the same screen. The mockup test phase for Gachapin proceeded in the same manner as the test phase in the main experiment, with the exception that corrective feedback was given for each task. It is important to note that on the generalization test trial of Gachapin, participants were asked to choose one object that Gachapin would add to his collection and were again reminded of the category: “Remember, Gachapin likes vegetables and is collecting different kinds of vegetables”. Subsequently, another encoding-test block proceeded using a different character, category, and set of stimuli. Participants who did not select the target in at least one of the two generalization trials did not proceed to the main experiment (*n*=3: 2 3-year-olds and 1 4-year-old child). The rationale for this exclusion criterion was that failing on a generalization test after explicit instructions about the character’s category and seeing all the encoding trials simultaneously indicated a failure in understanding the task.

Results

2.1. Estimating Semantic Similarity

To approximate the degree of semantic clustering of each category in our whole stimuli set, we calculated two semantic similarity scores from GloVe: (1) *within-category score*: an average pairwise similarity score across 36 pairs for a given category (9 items per category); and (2) *across-categories score*: an average pairwise similarity across 729 pairs for an item from a given category and all items from the other 9 categories learned in the same block (see Figure S1). The overall pattern shows that within-category scores are numerically higher than the across-categories scores for all 20 categories, although to varying degrees for different categories.

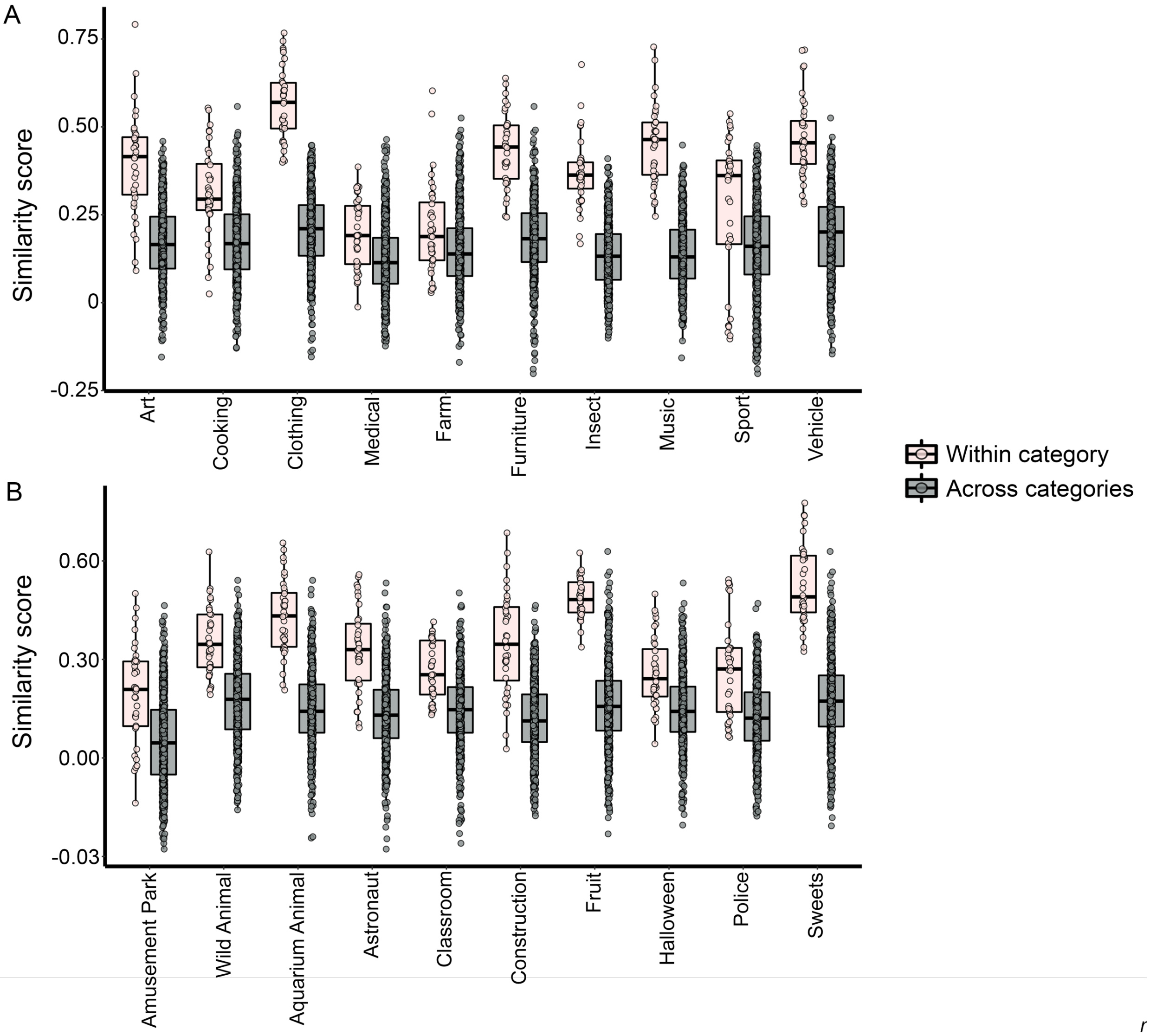

Figure S1. Distributions of within-category and across-categories similarity scores for all pairs of items learned in Block A (A) and Block B (B). Each point represents a pairwise inter-item similarity score.

2.2. Semantic similarity on generalization performance using Latent Semantic Analysis (LSA).

In addition to using GloVe, we evaluated the semantic similarity of words in our stimulus set with Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), which also works by analyzing large text corpora. LSA relies on singular value decomposition, a mathematical matrix decomposition technique similar to factor analysis. Instead of focusing on word-word co-occurrences, LSA evaluates the detailed patterns of word occurrences across many meaning-bearing contexts, such as sentences or paragraphs. Therefore, although using GloVe and LSA achieved the same goal—allowing us to calculate the semantic similarity of any given word pair—the disparate methods by which each technique achieves these ends and the different text corpora yield diverging results in some cases. Similar to our approach using GloVe, we computed a semantic similarity score for every pairwise of items learned in the same block using LSA trained on the TASA general reading first year college corpus. The similarity matrices derived from LSA are shown in Figure S1. This corpus included 37,651 documents, 92,406 terms, and 419 dimensions. The results on generalized linear mixed models predicting generalization success were the consistent with those from GloVe.

For young children, item conceptual specificity, β= 0.43, *SE*= 0.18, *z*= 2.42, *p*= .02, and semantic similarity, β= 1.67, *SE*= 0.74, *z*= 2.26, *p*= .02, were significantly associated with generalization success. Context binding accuracy, β= -0.09, *SE*= 0.17, *z*= -0.50, *p*= .62, was not associated with generalization success, but there was a trend towards significance for item perceptual specificity accuracy, β= -0.30, *SE*= 0.18, *z*= -1.68, *p*= .09 (see Figure 5, top). These results suggest that remembering the specific item identities and the greater degree of semantic relatedness within a category, the more likely younger children succeeded in making generalization judgment. Importantly, a reduced model that included only semantic similarity and item conceptual specificity as fixed effects (AIC= 862.11, BIC= 884.42) did not significantly differ from the full model (AIC= 862.02, BIC= 893.25), $\chi^2(2)$= 4.09, *p*= .13.

For older children, context binding accuracy, *β*= -0.13, *SE*= 0.20, *z*= -0.67, *p*= .50, item conceptual specificity accuracy, *β*= 0.10, *SE*= 0.22, *z*= 0.44, *p*= .66, item perceptual specificity accuracy, *β*= 0.18, *SE*= 0.20, *z*= 0.90, *p*= .37, and semantic similarity, *β*= 0.38, *SE*= 0.86, *z*= 0.44, *p*= .66, were not significantly associated with generalization success in older children.

For young adults, context binding accuracy was significantly associated with generalization success, *β*= .65, *SE*= 0.27, *z*= 2.40, *p*= .02, whereas item conceptual specificity, *β*= .26, *SE*= 0.27, *z*= 0.97, *p*= .33, item perceptual specificity, *β*= 0.02, *SE*= 0.26, *z*= 0.08, *p*= .94, and semantic similarity, *β*= -1.67, *SE*= 1.02, *z*= 1.63, *p*= .10, were not. A reduced model that only included context binding accuracy (AIC= 470.55, BIC= 488.01) did not differ from the full model (AIC= 472.86, BIC= 503.40), χ^2(3)= 3.69, *p*= .30, in predicting generalization success in young adults.

3.6. Task correlations

We conducted bivariate Pearson correlations to test whether performances on each of the four tasks was related to performances on the others, in each age group separately. For younger children, generalization positively correlated with context binding, *r*(30)= .49, *p*= .004, and item conceptual specificity, *r*(30)= .38, *p*= .03, but not with item perceptual specificity, *r*(30)= .09, *p*= .62. Context binding positively correlated with item conceptual specificity, *r*(30)= .57, *p*< .001, and there was a trend for item perceptual specificity, *r*(30)= .32, *p*= .07. Item conceptual and item perceptual specificity positively correlated with each other, *r*(30)= .33, *p*= .07.

For older children, generalization positively correlated with context binding, *r*(36)= .37, *p*= .02, and item conceptual specificity, *r*(36)= .45, *p*= .004, and with item perceptual specificity, *r*(36)= .48, *p*= .002. Context binding positively correlated with item conceptual specificity, *r*(36)= .46, *p*= .003, but not with item perceptual specificity, *r*(36)= .24, *p*= .16. Item conceptual and item perceptual specificity positively correlated with each other, *r*(36)= .35, *p*= .03.

For young adults, generalization positively correlated with context binding, *r*(27)= .55, *p*= .002, with item conceptual specificity, *r*(27)= .60, *p*< .001, and with item perceptual specificity, *r*(27)= .27, *p*= .15. Context binding positively correlated with item conceptual specificity, *r*(27)= .60, *p*< .001, but not with item perceptual specificity, *r*(27)= .27, *p*= .15. Item conceptual specificity and item perceptual specificity positively correlated with each other, *r*(27)= .62, *p*< .001.