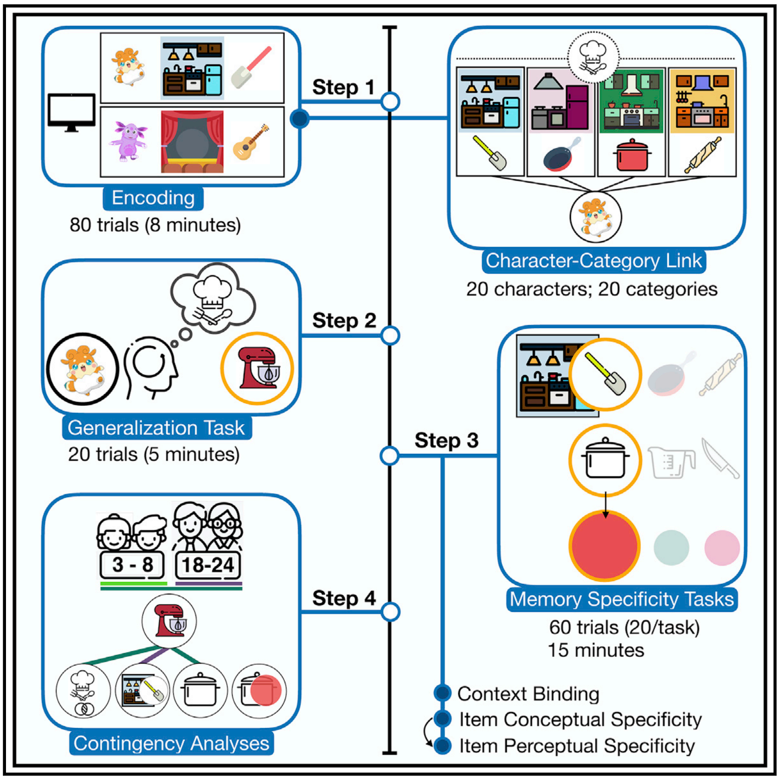
Current Biology

Report

Contingency of semantic generalization on episodic specificity varies across development

Graphical abstract



Authors

Chi T. Ngo, Susan L. Benear, Haroon Popal, Ingrid R. Olson, Nora S. Newcombe

Correspondence

ngo@mpib-berlin.mpg.de

In brief

Do we create general knowledge by drawing on memories of specific past experiences? Children learn categories and gain knowledge at rapid rates, despite fragile episodic memory. Ngo et al. show that, although adults' memories of specific events are related to generalizing, children form new generalizations based on prior semantic knowledge.

Highlights

- The ability to generalize knowledge to new situations improves from age 3 to age 8
- Different aspects of past events promote generalization success across development
- Adults, but not children, rely on context memories of past instances to generalize

Ngo et al., 2021, Current Biology *31*, 2690–2697
June 21, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cub.2021.03.088

CellPress

CellPress

Current Biology

Report

Contingency of semantic generalization on episodic specificity varies across development

Chi T. Ngo,[1,3,4,*] Susan L. Benear,[2,3] Haroon Popal,[2] Ingrid R. Olson,[2] and Nora S. Newcombe[2]
[1]Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
[2]Department of Psychology, Temple University, Philadelphia, PA, USA
[3]These authors contributed equally
[4]Lead contact
*Correspondence: ngo@mpib-berlin.mpg.de
https://doi.org/10.1016/j.cub.2021.03.088

SUMMARY

Semantic memory—general knowledge of ideas and concepts—includes generalization processes that support inference. Episodic memory, on the other hand, preserves the specificity of individual events by binding together unique combinations of elements from an episode and relies on pattern separation to distinguish similar experiences. These two memory systems play complementary roles, supporting different mnemonic goals, but the nature and extent of their interdependence is unclear.[1,2] Some models suggest that new information is encoded initially as hippocampus-dependent episodic memory and then, either through repetition or gist extraction, becomes semantic over time.[3,4] These models also posit a neocortical route to semantic memory acquisition exists that can bypass the hippocampus.[3] Both proposed routes are slow learning mechanisms, yet generalization can occur rapidly. Recent models suggest that fast generalization relies, in part, on the retrieval of individual but related episodes.[5,6] Such episodic memory gating mechanisms render fast generalization contingent on the memory specificity of instances, a pattern that has been observed in adults.[7,8] None of these models take into account the observation that generalization and episodic specificity have asynchronous developmental profiles, with generalization emerging years before episodic memory.[9,10] We ask two questions about generalized and specific memory during early childhood: first, is rapid generalization contingent on remembering specific past memories? And second, does the strength or nature of this contingency differ across development? We found that the interdependence of generalization and episodic memory varies across development: generalization success in adults, but not in children, was contingent on context binding.

RESULTS AND DISCUSSION

To chart the development of generalization and episodic memory specificity before and after age 6, when episodic memory appears robust,[11,12] we devised a memory task for children ages 3 to 8 years (n = 70) and also administered it to young adults (N = 29; STAR Methods). In this paradigm, we assessed generalization and memory specificity processes from the same set of experiences. Participants viewed 20 cartoon characters making collections of their favorite objects. Characters were embedded in 80 events, each showing a character in a context and with an object. Each character was seen with four objects from the same category (e.g., musical instruments) in four semantically congruent contexts (e.g., performance stages). We operationalized generalization as the ability to make novel inferences about characters based on commonalities among the places and objects seen with them (i.e., the category). We treated episodic specificity as multifaceted: 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) in three-alternative forced-choice tasks (Figure 1).

Data from five participants (two 4-year-olds, two 5-year-olds, and one 7-year-old) whose average accuracy across all four memory tasks did not exceed chance level (33%) were excluded from analyses. Overall chance-level performance likely reflects poor general understanding of the procedure or inattentiveness during encoding. Performance did not differ by sex (all p values > 0.12) or relate to verbal skills in any age groups (all p values > 0.22). Correlations between tasks are illustrated in Figure S1.

Age-related differences

We tested whether performance on each task increases as a function of age across 3–8 years. A linear regression of age (in months) on accuracy showed that, with increasing age, children were better at generalization ($R^2(63) = 0.20$; $p = 2.19 \times 10^{-4}$), item conceptual specificity ($R^2(63) = 0.22$; $p = 8.70 \times 10^{-5}$), and item perceptual specificity ($R^2(63) = 0.12$; $p = 0.005$). Interestingly, age was not associated with context binding accuracy ($R^2(63) = 0.03$; $p = 0.20$; Figure 2A). Among young adults, age was not related to memory performance (all p values > 0.46). Second, we compared children's memory performance to that of young adults. We used three categories: younger children (aged 3–5); older children (aged 6–8); and adults. A mixed 3

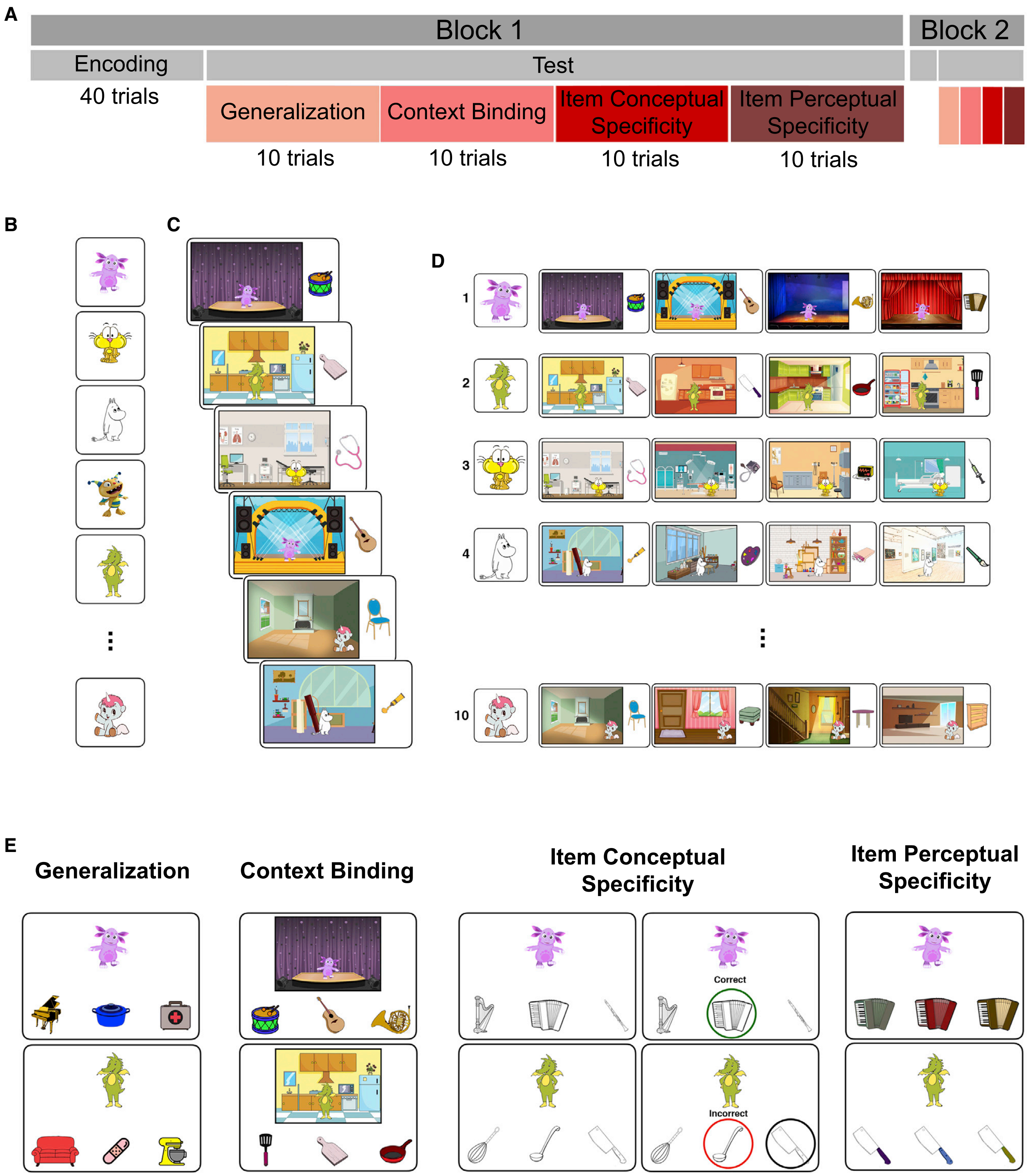


Figure 1. A schematic depiction of the main memory task procedure

The overall procedure included two blocks (A); each block included a character introduction phase (B), the encoding phase (C) with a pre-determined character-category assignment (D), and the test phase (E).

(age groups) × 4 (tasks) ANOVA yielded a main effect of age (F(2, 91) = 17.79; p < 0.001; η^2 = 0.13), a main effect of task (F(3, 272) = 54.71; p < 0.001; η^2 = 0.18), and a significant age by task interaction (F(6, 273) = 6.25; p < 0.001; η^2 = 0.04). Post hoc tests showed that younger children were less likely to make accurate generalizations than older children (t = −5.09; p_{holm} < 0.001) and adults (t = −7.28; p_{holm} < 0.001); older children and adults did not differ on generalization (t = −2.64; p_{holm} =

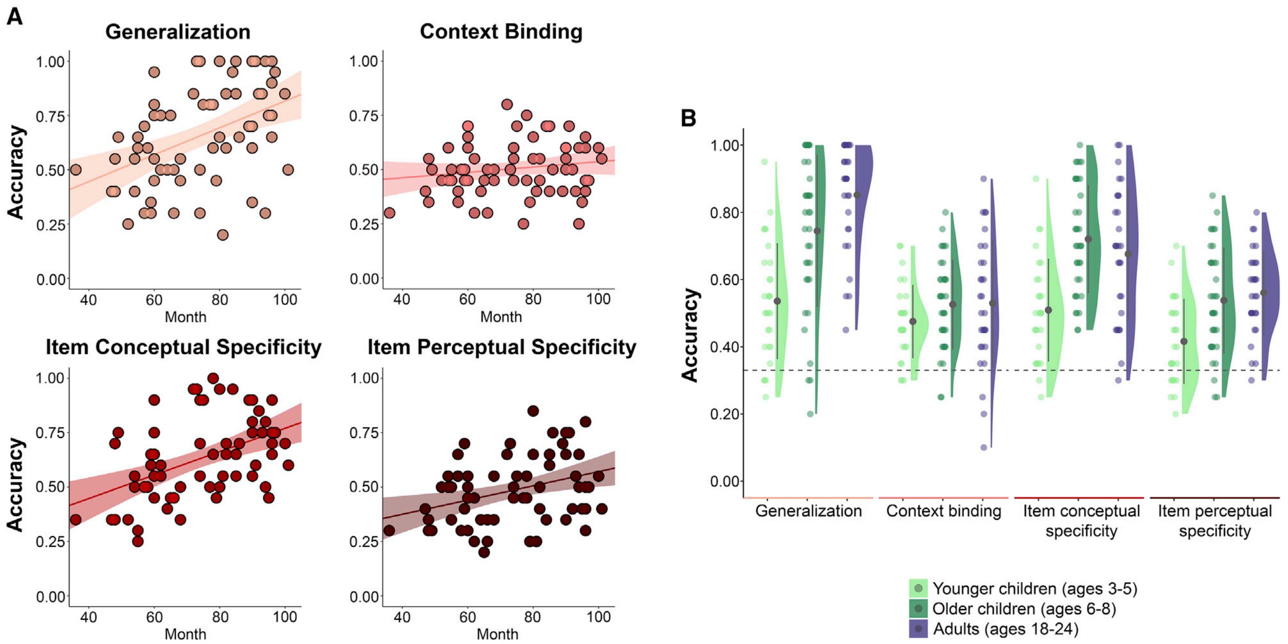


Figure 2. Age patterns in memory performances separated by tasks

(A) Scatterplots of memory accuracy on the generalization, context binding, item conceptual specificity, and item perceptual specificity tasks shown on the y axes and age (measured in months) shown on the x axes.

(B) Distributions of participants' accuracy on each task separated by age group, including younger children (ages 3–5), older children (ages 6–8), and young adults (ages 18–24). Black dots indicate the means, black lines indicate the error bars, and colored dots represent individual participants.

See also Figure S1.

0.24). Similarly, for item conceptual specificity, both older children (t = −5.15; $p_{holm} < 0.001$) and adults (t = −3.85; p_{holm} = 0.006) outperformed younger children, but their performances did not differ (t = 1.09; p_{holm} = 1.00). For item perceptual specificity, again older children and young adults did not differ (p_{holm} = 1.00), and adults outperformed younger children (t = −3.33; p = 0.03), but the difference between older and younger children was nonsignificant (p = 0.06). Surprisingly, there were no age differences in context binding among the three age groups (p > 0.86). Despite performing worse than their older counterparts in three out of four tasks, the younger children's accuracy exceeded chance level (0.33) on all four tasks (all p values < 0.002; Figure 2B).

Age-related improvements in generalization from early to middle childhood corroborate previous findings using different paradigms, including associative inference[13–15] and temporal regularity.[16] Although older children and adults did not differ in our study, previous studies reported an increase in associative inference or detection of recurring patterns from age 6 to 14.[15] Generalization may continue to change during middle childhood. We also found that younger children were less able to remember objects' identities and the perceptual details than older children and adults. The literature has shown few age-related differences in item conceptual memory,[10,17,18] but perhaps this age pattern only applies when conceptual interference among items is low.

Age changes in item perceptual specificity are consistent with previous work.[19,20] Past research on pattern separation has primarily used variations in single-object exemplars as stimuli, such as different rubber ducks,[21] creating interference along both conceptual and perceptual dimensions. We found that accuracy for objects' perceptual details was significantly associated with conceptual specificity (Figure S4A). Nonetheless, the age patterns in perceptual memory specificity among children persisted when we only examined trials in which the object identities were remembered (Figure S4B). These findings suggest that parsing the sources of item-level memory specificity is important for charting memory development.

Associations between age and context binding were not detected. Previous studies have consistently reported age-related improvements in context binding—or relational binding in general—throughout early and middle childhood.[12,17,22,23] We speculate that the nonsignificant age effects in our paradigm could be due to the unusually high number of item-context associations, together with strong semantic congruency for the item-context pairs associated with each character. In addition, participants learned about characters in an interleaved fashion, which is thought to be beneficial for generalization by increasing between-category discriminability.[24,25] Perhaps the interleaving dampened adults' memories for specific item-context pairs as a cost of promoting generalization. Further, children were exposed to all four types of questions prior to encoding in the "screening" procedure, whereas adults were not, which may have dampened age-related differences in context binding.

What determines generalization?

Our primary question was whether generalization depends on episodic specificity and whether this relationship depends on age. We asked (1) what aspects of episodic memory specificity would predict generalization success, (2) whether within-category semantic similarity would promote generalization, and (3) whether these factors interacted with age. We began by defining within-category semantic similarity. Semantic relatedness varied across different characters (Figure 3B). We quantified this variation using global vectors for word representation (GloVe)[26] to estimate the semantic similarity among the items in our stimulus set (Figure 3A; STAR Methods: Quantification and statistical analysis). GloVe is a vector space model that can be "trained" on a corpus of words by building a co-occurrence matrix and predicting the total co-occurrences between a target word and a context word. The co-occurrence of two words from large bodies of text is an index of their semantic relatedness.

We then conducted a generalized linear mixed effects model with age (as a continuous variable), context binding, item conceptual specificity, item perceptual specificity, semantic similarity, age*context binding, age*item conceptual specificity, age*item perceptual specificity, and age*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 response (successful or unsuccessful) conditional on the attributes of each participant and each category by adding them to the models as random effects. To ensure that ceiling effects did not impact the contingency patterns, we excluded participants whose generalization performance or item conceptual specificity performance reached ceiling level (100%; 9 adults and 8 older children).

Generalization success was significantly related to age (β = 0.33; z = 2.15; p = 0.032), semantic similarity (β = 0.16; z = 2.00; p = 0.046), item conceptual specificity (β = 0.27; z = 2.14; p = 0.032), and an age*context binding interaction (β = 0.39; z = 2.94; p = 0.003). No other predictors or interactions reached significance (all p values > 0.35; Figure 4). Increasing age and higher within-category semantic similarity were associated with higher probability of generalization. Intact memory specificity for the object identities seen with a character increased the probability of generalization for that character. Importantly, in adults, remembering the specific contexts of the objects was associated with generalization for a given character. In children, this link was not observed: generalization success was untethered to their context memory.

Our design targeted different aspects of episodic memory. Across all ages, greater degree of within-category semantic relatedness increased the probability of generalization success. The conceptual common ground that links together related episodes is important for generalization, suggesting a role for pre-existing semantic knowledge in promoting generalization. These findings align with the idea that category coherence is important for category learning and generalization.[27] Further, accurate item conceptual specificity was associated with greater generalization. Generalization benefitted from accurate memory specificity for the object identities from the individual episodes.

Crucially, contingency between episodic memory specificity and generalization varies across development. Adults' generalization relied on remembering specific item-context representations, consistent with the notion that rapid generalization relies on retrieving specific instances[7,8] and that there was no generalization-specificity trade-off.[28,29] These results suggest that adults' ability to rapidly generalize do not rely on abstraction—a process by which memories for the specific instances are lost, but the emergent average representation supports generalization across episodes.[30] Instead, it relies on their ability to remember specifics and then integrate overlapping elements across episodes.

However, contingency between generalization and context binding was not observed in either early or middle childhood. Memory for specific what-where relational structures is one key signature of episodic memory capacity.[31,32] Based on these 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 are available at different points in development. In a mature system, fast generalization can occur on the fly based on a chain-like retrieval process of the individual but related episodes. Preservation of rich contextual memories played a role in generalization for young adults, consistent with previous studies.[6,7] However, this link was not observed in children. Without a full constellation of robust episodic memory capacities, younger children may rely on the aspects of a specific episode that they do encode and retain, along with the support of overall semantic structures that tie together related episodes. These patterns align with previous findings that memory for individual items develops much earlier than item-context or item-item relational memory.[12,18] Even in our youngest children, memory specificity at the item level was tied to generalization success, namely conceptual specificity, suggesting a certain kind of contingency between semantic memory acquisition and episodic memory.

The development of generalization and episodic memory processes is linked to the maturational courses of several brain regions and networks. Several models have posited that the dentate gyrus and CA3 subfields of the hippocampus are involved in individuated memories.[33,34] Aligned with these ideas, development of late-maturing subfields, including the dentate gyrus and CA3, is associated with binding ability[35] and pattern separation processes.[19,36] Age differences in the structure[37] and functional recruitment[38] of the prefrontal cortex have also been linked to episodic memory improvements in late childhood and adolescence. In adults, recent research has shown that the hippocampus also contributes to the integration of related events to form new generalizable memories using a variety of paradigms.[27] The hippocampus may interact with the medial prefrontal cortex (mPFC) to integrate episodes with shared elements.[39] Indeed, associative inference in children is associated with gray matter volume in hippocampal head and mPFC.[40] 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 that occur in infancy and childhood.

An important question for the future is whether generalization relies on different neural substrates at different stages of neural development. It has been suggested that infants rely on the early-developing monosynaptic pathway linking the entorhinal cortex to CA1 to perform fast generalization.[34,41] Uneven maturational rates for intrahippocampal pathways and mPFC suggest that generalization might rely on different mechanisms in infancy, in childhood, and in adulthood. Specifically, some forms of rapid statistical learning and generalization could rely on the monosynaptic pathway early on in life,[34] whereas inference-based

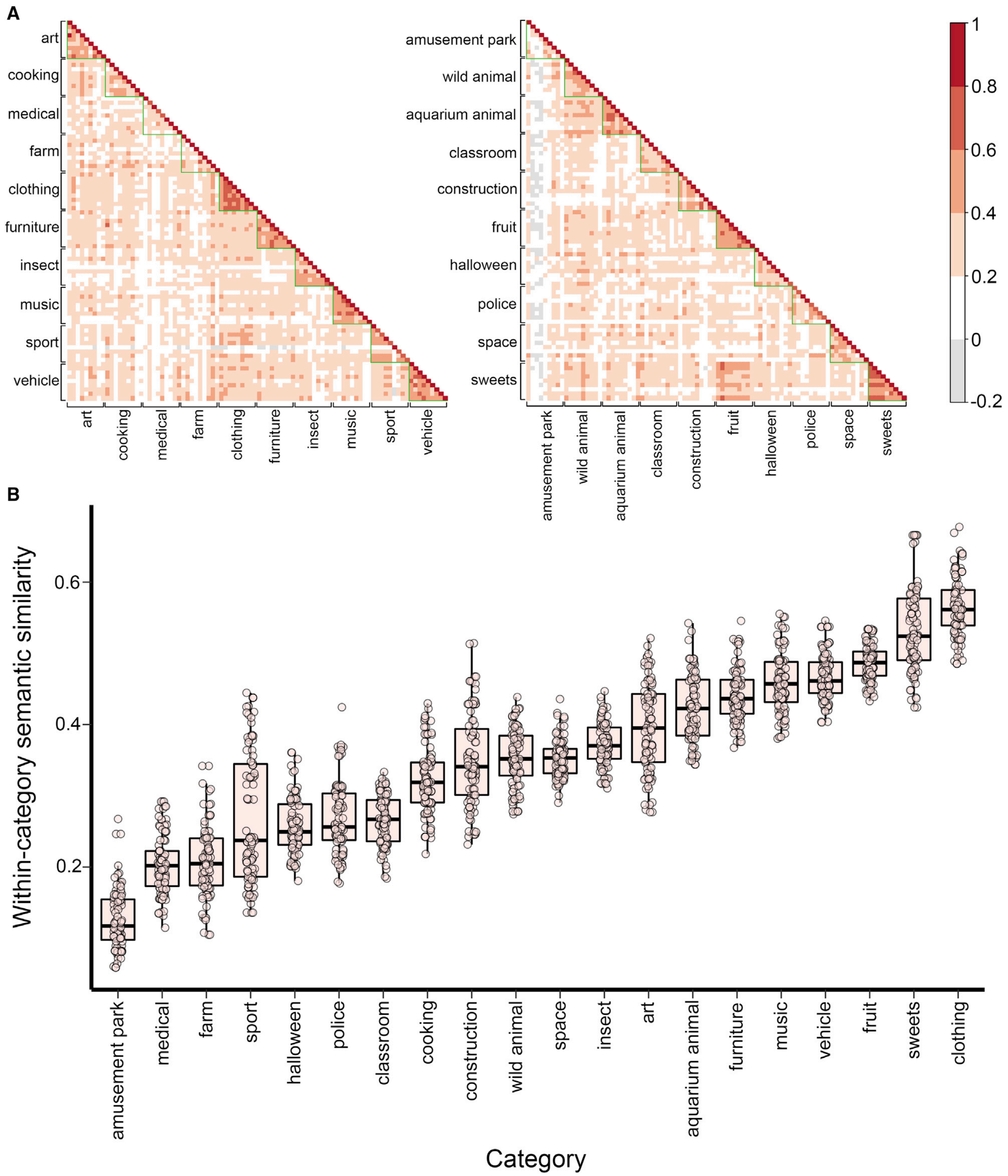


Figure 3. Semantic similarity between every pair of items within the stimulus set

(A) A matrix of all pairwise similarity scores for the 90-item stimulus set in blocks A (left) and B (right) calculated from global vectors for word representation (GloVe). Every cell represents an inter-item semantic similarity score, with darker colors representing higher scores.

(B) A distribution of participant-specific semantic similarity scores, defined as the mean of the four encoding items and the generalization target for each category. Higher similarity scores indicate that the items within a given category were closer to one another in semantic space (more similar). Each black horizontal bar indicates the group median, the whiskers extend to 1.5 times the interquarter range, and every dot represents a participant.

See also Figure S3.

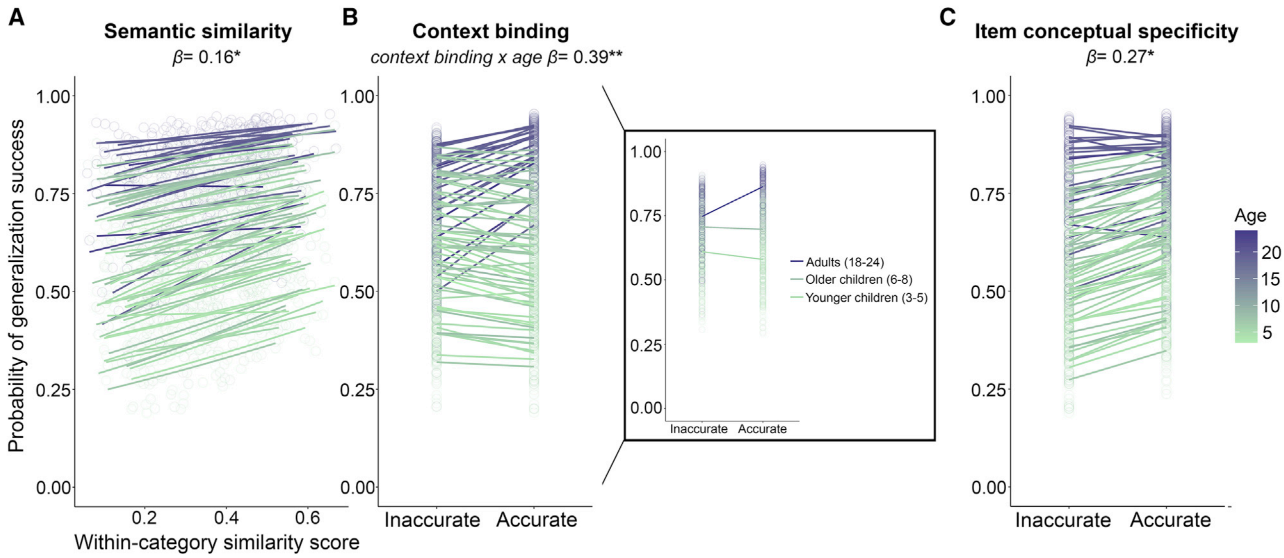


Figure 4. Age patterns in the contingency of generalization on memory specificity and semantic similarity

Distributions of the estimated probability of generalization success (y axes) by each significant fixed effect and interaction with age from the generalized linear mixed effects model. For semantic similarity (A), within-category semantic similarity scores are plotted on the x axis. Greater within-category similarity score is associated with higher probability of generalization success for the respective character. For context binding (B) and item conceptual specificity (C), accurate and inaccurate trials are plotted on the x axes. The association between context binding success and generalization success depends on age, such that accurate memory for context binding was associated with greater probability of generalization success in adulthood, but this pattern was not observed in childhood. A callout box depicts the same data, except age was grouped for visualization purposes. For item conceptual specificity, accuracy was associated with greater probability of generalization success for the corresponding character. Each line represents an individual participant; each dot denotes an individual trial. Color intensity represents age. Significance notation: *p < 0.05; **p < 0.01. See also Figure S4.

generalization later in life may recruit wider hippocampal circuitry in coordination with the prefrontal cortex.[42]

Based on this research, we suggest two next steps. First, a cross-sectional design limits us from understanding developmental change in generalization and specificity. Charting the potential lead-lag between the two memory functions would elucidate the dependency between semantic memory acquisition and episodic memory specificity. Perhaps behavioral gains in context binding allow one to rely on the rich contextual details of individual episodes for rapid generalization. Second, age-related differences in categorical knowledge need to be separately assessed and related to memory. In addition, two limitations of the current work should be corrected in future work. The design can be improved by randomizing the order of the context binding and the item conceptual and perceptual specificity tasks to circumvent unintended order effects in performance. Another limitation in the current work is that the children were exposed to examples of all four memory tasks in the screening procedure, whereas the adults were not administered the screening procedure. This procedural difference may have introduced uneven strategy deployment to the encoding phase between the age groups.

In conclusion, our study suggests that there may be multiple routes to inference-based generalization. Specifically, memory specificity for the rich context in an idiosyncratic episode may support fast generalization in adults, but not in children. The developmental asynchrony of semantic and episodic memory challenges the notion that semantic knowledge acquisition and rapid generalization are necessarily gated by episodic memory. This developmental phenomenon has important implications for contemporary models of memory that characterize the ontology of processes supporting complementary memory functions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subjects
  - Materials and Equipment
- METHOD DETAILS
  - Sorting Task
  - Materials
  - Procedure
  - Screening Procedure
  - Materials
  - Procedure
  - Main Experiment
  - Verbal Intelligence Assessment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Age-related Differences in Memory Performance
  - Semantic Similarity Analyses
  - Contingency Analyses on Memory Performances

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.03.088.

ACKNOWLEDGMENTS

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.N. (F31HD090872), I.R.O. (R01 MH091113, R56 MH091113, and R21 HD098509), and N.S.N. and I.R.O. (R01 HD099165) and by Temple University to N.S.N. (OVPR 161706-24607-02). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 10, 2020
Revised: February 10, 2021
Accepted: March 26, 2021
Published: April 21, 2021

REFERENCES

1. Tulving, E. (1972). Episodic and semantic memory. In Organization of Memory, E. Tulving, and W. Donaldson, eds. (Academic).
2. Schacter, D.L., and Tulving, E. (1994). What are the memory systems of 1994? In Memory Systems, D.L. Schacter, and E. Tulving, eds. (The MIT Press), pp. 1–38.
3. McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419–457.
4. Winocur, G., Moscovitch, M., and Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal-neocortical interactions. Neuropsychologia *48*, 2339–2356.
5. Kumaran, D., and McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. *119*, 573–616.
6. Koster, R., Chadwick, M.J., Chen, Y., Berron, D., Banino, A., Düzel, E., Hassabis, D., and Kumaran, D. (2018). Big-loop recurrence within the hippocampal system supports integration of information across episodes. Neuron *99*, 1342–1354.e6.
7. Banino, A., Koster, R., Hassabis, D., and Kumaran, D. (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. Sci. Rep. *6*, 31330.
8. Tompary, A., Zhou, W., and Davachi, L. (2020). Schematic memories develop quickly, but are not expressed unless necessary. Sci. Rep. *10*, 16968.
9. Keresztes, A., Ngo, C.T., Lindenberger, U., Werkle-Bergner, M., and Newcombe, N.S. (2018). Hippocampal maturation drives memory from generalization to specificity. Trends Cogn. Sci. *22*, 676–686.
10. Newcombe, N.S., Benear, S., Ngo, C.T., and Olson, I.R. (2021). Memory in infancy and childhood. In Oxford Handbook on Human Memory (Oxford University Press), M. Kahana and A. Wagner, Eds).
11. Ngo, C.T., Newcombe, N.S., and Olson, I.R. (2018). The ontogeny of relational memory and pattern separation. Dev. Sci. *21*, https://doi.org/10.1111/desc.12556.
12. Riggins, T. (2014). Longitudinal investigation of source memory reveals different developmental trajectories for item memory and binding. Dev. Psychol. *50*, 449–459.
13. Varga, N.L., Stewart, R.A., and Bauer, P.J. (2016). Integrating across episodes: investigating the long-term accessibility of self-derived knowledge in 4-year-old children. J. Exp. Child Psychol. *145*, 48–63.
14. Bauer, P.J., and San Souci, P. (2010). Going beyond the facts: young children extend knowledge by integrating episodes. J. Exp. Child Psychol. *107*, 452–465.
15. Schlichting, M.L., Guarino, K.F., Schapiro, A.C., Turk-Browne, N.B., and Preston, A.R. (2017). Hippocampal structure predicts statistical learning and associative inference abilities during development. J. Cogn. Neurosci. *29*, 37–51.
16. Pudhiyidath, A., Roome, H.E., Coughlin, C., Nguyen, K.V., and Preston, A.R. (2020). Developmental differences in temporal schema acquisition impact reasoning decisions. Cogn. Neuropsychol. *37*, 25–45.
17. Lloyd, M.E., Doydum, A.O., and Newcombe, N.S. (2009). Memory binding in early childhood: evidence for a retrieval deficit. Child Dev. *80*, 1321–1328.
18. Sluzenski, J., Newcombe, N.S., and Kovacs, S.L. (2006). Binding, relational memory, and recall of naturalistic events: a developmental perspective. J. Exp. Psychol. Learn. Mem. Cogn. *32*, 89–100.
19. Canada, K.L., Ngo, C.T., Newcombe, N.S., Geng, F., and Riggins, T. (2019). It's all in the details: relations between young children's developing pattern separation abilities and hippocampal subfield volumes. Cereb. Cortex *29*, 3427–3433.
20. Ngo, C.T., Newcombe, N.S., and Olson, I.R. (2019). Gain-loss framing enhances mnemonic discrimination in preschoolers. Child Dev. *90*, 1569–1578.
21. Liu, K.Y., Gould, R.L., Coulson, M.C., Ward, E.V., and Howard, R.J. (2016). Tests of pattern separation and pattern completion in humans-a systematic review. Hippocampus *26*, 705–717.
22. Lee, J.K., Fandakova, Y., Johnson, E.G., Cohen, N.J., Bunge, S.A., and Ghetti, S. (2020). Changes in anterior and posterior hippocampus differentially predict item-space, item-time, and item-item memory improvement. Dev. Cogn. Neurosci. *41*, 100741.
23. Ngo, C.T., Lin, Y., Newcombe, N.S., and Olson, I.R. (2019). Building up and wearing down episodic memory: mnemonic discrimination and relational binding. J. Exp. Psychol. Gen. *148*, 1463–1479.
24. Birnbaum, M.S., Kornell, N., Bjork, E.L., and Bjork, R.A. (2013). Why interleaving enhances inductive learning: the roles of discrimination and retrieval. Mem. Cognit. *41*, 392–402.
25. Kornell, N., and Bjork, R.A. (2008). Learning concepts and categories: is spacing the "enemy of induction"? Psychol. Sci. *19*, 585–592.
26. Pennington, J., Socher, R., and Manning, C.D. (2014). GloVe: Global Vectors for Word Presentation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 1532–1543.
27. Zeithamova, D., and Bowman, C.R. (2020). Generalization and the hippocampus: more than one story? Neurobiol. Learn. Mem. *175*, 107317.
28. Gershman, S.J., Schapiro, A.C., Hupbach, A., and Norman, K.A. (2013). Neural context reinstatement predicts memory misattribution. J. Neurosci. *33*, 8590–8595.
29. L Varga, N., Gaugler, T., and Talarico, J. (2019). Are mnemonic failures and benefits two sides of the same coin?: Investigating the real-world

consequences of individual differences in memory integration. Mem. Cognit. *47*, 496–510.

30. Altmann, G.T.M. (2017). Abstraction and generalization in statistical learning: implications for the relationship between semantic types and episodic tokens. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160060.
31. Tulving, E. (2002). Episodic memory: from mind to brain. Annu. Rev. Psychol. *53*, 1–25.
32. Johnson, M.K., Hashtroudi, S., and Lindsay, D.S. (1993). Source monitoring. Psychol. Bull. *114*, 3–28.
33. Norman, K.A., and O'Reilly, R.C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol. Rev. *110*, 611–646.
34. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049.
35. Riggins, T., Geng, F., Botdorf, M., Canada, K., Cox, L., and Hancock, G.R. (2018). Protracted hippocampal development is associated with age-related improvements in memory during early childhood. Neuroimage *174*, 127–137.
36. Keresztes, A., Bender, A.R., Bodammer, N.C., Lindenberger, U., Shing, Y.L., and Werkle-Bergner, M. (2017). Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. Proc. Natl. Acad. Sci. USA *114*, 9212–9217.
37. Sowell, E.R., Peterson, B.S., Thompson, P.M., Welcome, S.E., Henkenius, A.L., and Toga, A.W. (2003). Mapping cortical change across the human life span. Nat. Neurosci. *6*, 309–315.
38. Selmeczy, D., Fandakova, Y., Grimm, K.J., Bunge, S.A., and Ghetti, S. (2019). Longitudinal trajectories of hippocampal and prefrontal contributions to episodic retrieval: effects of age and puberty. Dev. Cogn. Neurosci. *36*, 100599.
39. Schlichting, M.L., and Preston, A.R. (2015). Memory integration: neural mechanisms and implications for behavior. Curr. Opin. Behav. Sci. *1*, 1–8.
40. Bauer, P.J., Dugan, J.A., Varga, N.L., and Riggins, T. (2019). Relations between neural structures and children's self-derivation of new knowledge through memory integration. Dev. Cogn. Neurosci. *36*, 100611.
41. Gómez, R.L., and Edgin, J.O. (2016). The extended trajectory of hippocampal development: implications for early memory development and disorder. Dev. Cogn. Neurosci. *18*, 57–69.
42. Zeithamova, D., and Preston, A.R. (2010). Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. J. Neurosci. *30*, 14676–14684.
43. Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J.K. (2019). PsychoPy2: experiments in behavior made easy. Behav. Res. Methods *51*, 195–203.
44. JASP Team (2020). JASP, version 0.14.1.
45. RStudio Team (2020). RStudio: integrated development for R (RStudio).
46. Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R.H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., et al. (2020). "Package 'lme4'," CRAN (R Foundation for Statistical Computing).
47. MathWorks (2010). MATLAB, version 7.10.0 (R2010a) (MathWorks).
48. Kaufman, A.S., and Kaufman, N.L. (1990). Kaufman Brief Intelligence Test (American Guidance Service).
49. Grober, E., and Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. J. Clin. Exp. Neuropsychol. *13*, 933–949.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Memory Task	Custom program	https://github.com/hspopal/generalization
PsychoPy, v3.0.0	Peirce et al.[43]	https://psychopy.org; RRID: SCR_006571
JASP, v0. 14.1	JASP Team[44]	https://jasp-stats.org/; RRID: SCR_015823
RStudio, v1.4.1103	RStudio Team[45]	https://www.rstudio.com; RRID: SCR_000432
lme4 package (Rstudio)	Bates et al.[46]	https://github.com/lme4/lme4/; RRID: SCR_015654
MATLAB, v9.8	MathWorks[47]	https://www.mathworks.com/products/matlab.html; RRID: SCR_001622
Global Vectors for Word Representation (GloVe)	Pennington et al.[26]	https://nlp.stanford.edu/projects/glove/
Other		
Kaufman Brief Intelligence Test, Second Edition (KBIT-2)	Kaufman and Kaufman[48]	N/A
American National Adult Reading Test (AMNART)	Grober and Sliwinski[49]	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Chi (Zoe) Ngo (ngo@mpib-berlin.mpg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All experimental materials, second-level data, and scripts for the analyses have been made publicly available through the Open Science Framework (https://osf.io/gv485/). The code for the memory experiment is available via Github (https://github.com/hspopal/generalization).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects

Child Sample in the Sorting Task

Six children (five 4-year-olds and one 5-year-old) who did not participate in the main experiment participated in our sorting task (see *Step-by-Step* Method details).

Child Sample in the Main Experiment

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; M_{month} = 86.24 ± 8.46, range = 72-101) recruited from 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 Screening 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 memory performance separated by testing format, see Figure S2).

Young Adult Sample in the Main Experiment

The young adult sample consisted of 29 undergraduate students (18 females; 11 males; $M_{age(years)}$ = 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. All young adults were given partial course credit. This experiment was approved by the Temple University Institutional Review Board committee.

Materials and Equipment

All participants were administered the memory task from MacBook laptops running macOS v10.11 or later. The PsychoPy memory task was installed on these laptops in advance and displayed to children either in-person or via screen-sharing over Zoom. To conduct the experiment, the PsychoPy program, stimuli files, and the task scripts were locally downloaded onto the laptops in advance.

METHOD DETAILS

Sorting Task

After our initial stimulus selection, we gauged children's familiarity with the categories using a sorting task. Six children (5 4-year-olds and 1 5-year-olds) who did not participate in the main experiment participated in this task.

Materials

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).

Procedure

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.

Screening Procedure

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.

Materials

Stimuli were selected in the same manner as those used in the Memory Task (see Main Experiment, Materials), except that the screening procedure was created and administered using PowerPoint v16.16.19.

Procedure

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: two 3-year-olds and one 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 comprehending the task.

Main Experiment

Materials

PsychoPy version 3.0.0 was used to program and present the Memory Task in full screen on laptop computers. Randomization between subjects was done in regards to the order of trials and the selection of objects (without replacement) for each of the task sections. 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; 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. These additional stimuli were semantically unrelated to those used for the study and test phases of the experiment.

In-person 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 (Figure 1A). The test phase consisted of four memory tasks described below.

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 (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 (5 s, 0.5 s 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 (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: (a) generalization; (b) context binding; (c) item conceptual specificity; and (d) item perceptual specificity. These were administered in a fixed order across participants (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 2 s 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.

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.

Verbal Intelligence Assessment

Materials

Children were administered the verbal portion of the Kaufman Brief Intelligence Test, second edition (KBIT-243), whereas young adults were given the American National Adult Reading Test (AMNART[49]), as measures of general verbal skills. One child was not administered the KBIT due to fatigue.

Procedure

KBIT-2. Fifty-eight children were tested in person and twelve children were virtually tested.

In-person testing format

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.

Virtual testing format

The experimenter shared their own screen on Zoom with the participant, on which was displayed a scanned-in version of the KBIT-2, displayed on screen rather than on paper as it would be in person. Otherwise, the experimenter administered the test verbally in the same manner as when in person. When participants responded to the questions by pointing to one of the options on the screen, participants' parents were instructed to respond with the letter corresponding to the image the participant pointed to (A-F) to indicate to the experimenter which option the participant had selected. Parents were specifically instructed to not name objects that appeared on screen, and to refer only to their corresponding letter label.

AMNART

All young adults were tested in person. The 45-item AMNART is an American version of the National Adult Reading Test. This test measures the ability to read aloud irregular English words. Pronunciation errors were tallied by the experimenter and an AMNART-estimated IQ score was calculated using Grober and Sliwinski's formula, which accounts for years of education.[49]

QUANTIFICATION AND STATISTICAL ANALYSIS

We used JASP v0.14.1 and RStudio v1.4.1103 to conduct all of our statistical analyses. The significance level of 0.05 was applied to all analyses, with the exception of Posthoc tests which employed a Holm correction to the significance level. We quantified accuracy as the proportion of trials in which the targets were selected in our three-alternative forced choice (AFC) test. First, we identified participants whose average accuracy across the four memory tasks did not exceed chance level (33%). These participants were not included in the subsequent analyses.

To test whether memory performance differed by sex, we conducted an independent-samples t tests to compare accuracy on each task between male and female participants. Further, we tested whether verbal skills were related to memory by conducting bivariate Pearson correlations between verbal score (KBIT-2 and AMNART for children and adults, respectively) and accuracy on each task accuracy. A separate correlation was conducted for each of the three age groups: younger children (aged 3-5), older children (aged 6-8) and adults (aged 18-24).

Age-related Differences in Memory Performance

To test for age-related differences in memory performances, we tested whether performance on each of the memory tasks (generalization, context binding, item conceptual specificity, and item perceptual specificity) was related with age in children by conducting Pearson correlations between age (measured in months) and accuracy. The same analysis was performed in adults, except that age was measured in years. Further, to test whether memory performance differed by age and whether the age effects interacted with memory tasks from early childhood into young adulthood, we conducted a mixed 3 (age groups: younger children, older children, adults) x 4 (tasks) ANOVA. Children aged 3-5 were categorized as younger children. Children aged 6-8 were categorized as older children. The adult group consisted of participants ages 18-24. All four tasks were included: generalization, context binding, item conceptual specificity, and item perceptual specificity.

Finally, 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 (younger children, older children, adults) separately.

Semantic Similarity Analyses

Global Vectors for Word Representation (GloVe)[26] was employed using MATLAB v9.8 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 (Figure 3A). To approximate the degree of semantic clustering of each category in our whole stimuli set, we calculated two semantic similarity scores from GloVe: (i) within-category score: an average pairwise similarity score across 36 pairs for a given category (9 items per category); and (ii) 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 (Figure S3). The overall pattern shows that within-category scores are higher than across-category scores for all 20 categories, although to varying degrees for different categories.

Given that for each category, we randomly assigned four items that appeared at encoding, and one generalization target at test, we computed a participant-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 (Figure 3B).

Contingency Analyses on Memory Performances

To test what aspects of episodic memory specificity would predict generalization success and whether within-category semantic similarity would promote generalization, for our three age groups separately (younger children, older children, and young adults). We used the *lmer* package in RStudio v1.4.1103 to conduct a generalized linear mixed effects model with age, semantic similarity, context binding accuracy, item conceptual specificity accuracy, and item perceptual specificity accuracy, age*semantic similarity, age*context binding accuracy, age*item conceptual specificity accuracy, and age*item perceptual specificity accuracy as fixed effects, with participant and category were included as random effects, to predict generalization success on a trial-by-trial basis. Age and semantic similarity were continuous variables. Context binding, item conceptual specificity, and item perceptual specificity accuracy were binary variables (0 for incorrect and 1 for correct responses). 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.

To test whether remembering an object's perceptual details was strongly tied to participants' conceptual memory for that object, we conducted a generalized linear mixed effects model predicting item perceptual specificity with item conceptual specificity, age, and item conceptual specificity*age as fixed effects and with participants and categories as random effects (Figure S4A).