Generalization from early to middle childhood

Tydings M. McClary[1+], Simon Ciranka[2,3], Elisa S. Buchberger[1], Bernhard Spitzer[2,4], Ulman Lindenberger[1,5], Chi T. Ngo[1*], & Markus Werkle-Bergner[1+*]

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

[2] *Research Group Adaptive Memory and Decision Making, Max Planck Institute for Human Development, Berlin, Germany*

[3] *Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany*

[4] *Biopsychology, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany*

[5] *Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany, and London, UK*

[+]Corresponding authors: Tydings M. McClary (mcclary@mpib-berlin.mpg.de) and Markus Werkle-Bergner (werkle@mpib-berlin.mpg.de)

[*]Co-senior authorship

ORCID IDs:
T.M.M.: 0009-0009-9960-8211; S.C.: 0000-0002-2067-9781; E.S.B.: 0000-0001-5377-9722; B.S.: 0000-0001-9752-932X; U.L.: 0000-0001-8428-6453; C.T.N.: 0009-0005-7229-8883; M.W.-B.: 0000-0002-6399-9996

Acknowledgements

The research was conducted within the project Lifespan Rhythms of Memory and Cognition (RHYME; PI: M.W.-B.) at the Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin. T.M.M. is a fellow of the International Max Planck Research School on Learning, Institutions, and Future Evolution (IMPRS LIFE). M.W.-B. received financial support from the Jacobs Foundation through an Early Career Research Fellowship. C. T. N. received financial support from the German Research Foundation (Project #NG 191/2-1) and the Jacobs Foundation (2021-1417-99). We further acknowledge support by the Max Planck Dahlem Campus of Cognition (MPDCC). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. All necessary support, including funding, facilities, and IRB approval are in place for the proposed research. The study was approved by the Ethics Committee of the Max Planck Institute for Human Development. All authors declare no conflicts of interest.

Data availability

Experimental stimuli, de-identified raw data, second-level data, and analysis scripts will be made publicly available through Open Science Framework (https://osf.io/89gwp/) upon acceptance of our manuscript.

1. Abstract

Generalization enables individuals to apply knowledge derived from past experiences in novel situations, yet it is unclear whether common experimental paradigms probe the same underlying ability in early to middle childhood. We assessed 84 children aged 4, 6, and 8 years on four paradigms, namely statistical learning, associative inference, transitive inference, and categorization, and examined age effects and cross-task concordance. Accuracy increased with age in every task, but inter-task correlations were weak, suggesting that each paradigm might tap a partly distinct facet of generalization. Focused analyses of the associative and transitive inference task showed that older children's success increasingly relied on memory for directly learned pairs. In younger children, the weak memory–generalization link diverged by task: 4- and 6-year-olds often solved transitive inferences without accurate pair memory, whereas such increased memory-independent generalization was not evident in associative inference. Computational modeling supported this pattern, indicating that younger children used value-based reinforcement learning to solve transitive problems, whereas 8-year-olds employed a pair-memory strategy. Taken together, our findings show that generalization develops through multiple pathways, underscoring the need for multi-indicator and model-based approaches to capture its multiple facets.

2. Introduction

Imagine watching a tennis match in which Novak Djokovic defeats Rafael Nadal. You might infer that Djokovic would likely also beat other players whom Nadal has previously defeated, such as Roger Federer. In another scenario, imagine moving to a new city and meeting your neighbor walking a Labrador retriever. A few days later, you see a different person walking the same dog, prompting the inference that this person is most likely acquainted with your neighbor and, therefore, associated with them. Such inferences exemplify generalization — a fundamental cognitive capacity that enables individuals to extrapolate from prior knowledge in service of behaving adaptively in new situations.

Generalization is thought to rely on cognitive and neural mechanisms that abstract commonalities across individual episodes and apply this knowledge to novel stimuli or situations. However, the nature of these mechanisms is actively debated (for a review, see Taylor et al., 2021), in part because generalization encompasses multiple learning processes (Zeithamova & Bowman, 2020). Some suggested mechanisms exploit statistical regularities among experiences, such as co-occurrences, and are assumed to take place automatically and without the need for feedback (Forest et al., 2023). Others depend on the retrieval of individual, but overlapping specific instances from the past; some reflect categorization, while others rely on hierarchical and ordinal structures in the world (Zeithamova et al., 2012). Investigating the development of generalization provides a valuable lens through which to disentangle the cognitive and neural mechanisms that underlie its different forms.

Research on early development of generalization uses a wide range of behavioral paradigms, including statistical learning, categorization, and episodic inference, such as associative and transitive inference. The capacity to detect statistical regularities is evident very early in life. For instance, infants and toddlers are capable of building generalizable semantic knowledge (discussed in Keresztes et al., 2018; Ramsaran et al., 2019). They can extract and retain spoken words over longer periods of multiple weeks by the age of 8 months (Jusczyk & Hohne, 1997) and can readily infer category memberships (Sloutsky, 2003). A landmark study by Saffran et al. (1996) found that 8-month-old infants could quickly detect statistical regularities in a continuous auditory stream of syllables with uneven transitional probabilities. These findings

have since been replicated in the visual domain (Kirkham et al., 2002) and even in newborns (Bulf et al., 2011), suggesting an early-developing capacity for statistical regularity detection.

Nevertheless, both statistical learning and other forms of generalization continue to develop throughout childhood (Ngo et al., 2021) and adolescence (Schlichting et al., 2017). For example, *associative* inference, or the ability to infer a novel relationship (e.g., sheep - dog) from previously learned associations (e.g., sheep - park, dog - park), has been shown to increase between ages 3 to 5, and to be positively related to future thinking via flexibly combining past events (Richmond & Pan, 2013). Associative inference and statistical learning are positively correlated after controlling for age, both of which continue to improve from age 6 to adulthood (Schlichting et al., 2017).

Another form of generalization, *transitive* inference, involves inferring relationships between items that have not been experienced together (transitive pair, e.g. A<D) based on learned relationships from overlapping pairings (premise pairs, e.g. A<B, B<C, and C<D), with first investigations of this phenomenon in children dating back to Piaget (1924/1928). Age-related patterns are mixed and depend on the paradigm and criteria used (Chapman & Lindenberger, 1988, 1992b), as shown by findings where successful transitive inference can appear as early as around 16 months or as late as the age of 10. A potential reason for this diversity is that different transitive inference experiments are solvable via different strategies that vary in complexity, where some are available to younger children and others not. In one study, infants saw an agent choosing colored balls in a transitive order (e.g., red over yellow and a yellow over green) and subsequently showed preferential looking for violations of expected selections – selecting green over red – compared to the expected ones – selecting red over green (Mou et al., 2014). 4-year-old children succeed at transitive inference tasks with extensive corrective feedback and when the order of premise relations is spatially correlated (Bryant & Trabasso, 1971), an experimental feature that children can exploit to succeed in the task. Yet another study using spatially uncorrelated premise relations during learning found that accuracy on transitive inference exceeds chance performance only by the age of 10 (Townsend et al., 2010). Nevertheless, when investigating different transitive inference task procedures, a clear positive age trend emerges: preschoolers perform above chance, but subpar to fourth graders, who perform worse than adults (Kallio, 1982).

Category learning is another facet of generalization that unfolds across infancy and childhood, moving from basic perceptual groupings to increasingly complex and abstract forms. By three to four months, infants already form categories based on perceptual similarity (Quinn et al., 1993), and recent evidence shows that visual object categorization is a robust ability during infancy (Spriet et al., 2022). As development progresses, children learn more prototype-based, with category size, structure, and complexity influencing acquisition (Minda & Smith, 2001). By middle childhood, categorization reflects the integration of perceptual, attentional, and conceptual processes (Sloutsky & Fisher, 2011), with children being able to use multiple features, including linguistic labels, to differentiate between categories (Deng & Sloutsky, 2013).

Charting the developmental profile of generalization requires understanding its multiple forms and their interdependence. A key question is whether generalization depends on retaining overlapping memories or whether alternative routes exist. Evidence from neuroimaging studies provides a starting point, implicating the hippocampus, a central region for memory formation and retrieval, in several forms of generalization, including statistical learning (Schapiro et al., 2012), transitive inference (Heckers et al., 2004), associative inference (Preston et al., 2004), and categorization (Bowman & Zeithamova, 2018) Neurocomputational models seek to explain *how* the hippocampus supports rapid generalization beyond its traditional role in episodic memory (Kumaran & McClelland, 2012; Schapiro et al., 2017; Sučević & Schapiro, 2023). For instance, the *Recurrency and Episodic Memory Results in Generalization* model (REMERGE; Kumaran & McClelland, 2012) proposes that recurrent connections in hippocampal subfield CA3 support "on-the-fly" generalization by reactivating overlapping episodic memories. The *Complementary Hippocampal Operations for Representing Statistics and Episodes* model (C-HORSE*;* Schapiro et al., 2017) distinguishes between the trisynaptic pathway that supports specific memories, and the monosynaptic pathway that extracts regularities and fosters generalization (Sučević & Schapiro, 2023). Notably, the monosynaptic pathway develops earlier (Gómez & Edgin, 2016; Lavenex & Banta Lavenex, 2013), inviting the hypothesis that generalization can occur in the face of relatively weak memory specificity early in life (Keresztes et al., 2018; Ellis et al., 2021; Newcombe et al., 2024).

Episodic inference tasks, such as associative and transitive inference, provide an opportunity to examine the link between memory and generalization. If such dependency exists in the case of associative inference, inferring the relationship between items (AC) based on an overlapping pair mate (B) depends on the memories of the individual pairs (AB and BC). Thus far, not much is known about such memory-inference relationship. Richmond and Pan (2013), found no difference in inference performance between preschoolers with perfect versus imperfect recognition on direct pairs, suggesting that inference does not necessarily rely on remembering directly learned pairs. Schlichting and colleagues (2017) only analyzed the inference trials where participants had accurate memory of both direct pairs. Given the differences in data-analytic approaches between the two studies, the available evidence does not allow adjudicating the extent to which remembering previously encoded relations benefits inference, and whether this benefit differs by age.

In the case of transitive inference, transitivity should depend on premise pair memory if generalization relies on the retrieval of overlapping instances. Previous work suggests that children can provide answers indicative of transitive inference in the absence of memory for the premise pairs (Brainerd & Reyna, 1992). This stands in opposition to operational reasoning in transitive inference, which relies on explicitly remembering premise pairs (Piaget et al., 1977; Chapman & Lindenberger, 1988, 1992a, 1992b). Following up on these controversies, it has been suggested that assigning implicit values to items during feedback-based learning is an alternative mechanism for transitivity, enabling inference through value comparison (Jensen et al., 2015; see also "functional reasoning" in Piaget et al., 1968/1977). Here, individual item values are updated via feedback on premise pairs and are subsequently used to guide decisions on transitive pairs. Recent modeling work indicates that learning individual item values provides a better account of transitive inference in adults than remembering pairwise associations between items (Ciranka et al., 2022). Thus, dependent on task parameters, different processing pathways can support successful inference. Based on this evidence, one might also expect that generalization does not develop unitarily across tasks and paradigms.

The current study pursued three goals. First, we examined age-related differences in generalization across four tasks, namely, associative inference, transitive inference, statistical learning, and categorization, in children aged 4, 6, and 8 years. Second, we tested whether

individual differences in task performance are correlated, which would suggest shared underlying processes. If the different forms of generalization reflect the same set of learning mechanisms, performances on these tasks should strongly correlate within and across age groups. Correlations among tasks within age groups turned out to be relatively low, prompting a third, exploratory goal: to investigate differential generalization-memory contingencies between the two episodic inference tasks: associative and transitive inference.

To preview, we observed significant age-related improvements across all generalization tasks, most consistently between ages 4 and 6. In both associative and transitive inference tasks, intact pair memories supported generalization, but their benefits were greater for older than younger children. Computational modeling of transitive inference further revealed age-related shifts in strategy use: younger children more frequently learned through simple reinforcement learning, whereas older children increasingly drew on explicit retrieval of pairwise associative memories.

3. Methods

3.1. Participants

Participants aged 4, 6, and 8 were recruited from the Berlin metropolitan area for this study via the in-house participant database CASTELLUM of the Max Planck Institute for Human Development (Bengfort et al., 2022). A total of 87 children participated in the current study, which took place on the premises of the Max Planck Institute for Human Development in Berlin, Germany. Out of 33 4-year-olds, 3 were excluded entirely due to lack of understanding or non-compliance with task instructions on two or more tasks, resulting in a total of 84 participants: 30 4-year-olds (M = 54.9 months, SD = 2.9 months, 14 females), 27 6-year-olds (M = 79.2 months, SD = 3.4 months, 13 females), and 27 8-year-olds (M = 102.2 months, SD = 2.9 months, 14 females). Prior to the lab visit, participants were screened according to a set of inclusion criteria, which required that they be native German speaker, free of chronic illnesses, sleep disorders, learning disorder, or diagnosed psychiatric or neurological conditions, not taking medication that affects the central nervous system, free of red-green deficiency or color blindness, and free of any other illness at the timepoint of each study session. Legal guardians gave informed consent for their children, and children verbally agreed to participate when asked during the first session. Families received monetary compensation of 10€/hour as well as a 5€ completion bonus for taking part in both sessions. In addition, children received a child-friendly research certificate and a small gift for their participation. The study was approved by the Ethics Committee of the Max Planck Institute for Human Development.

3.2. Study procedure

The task battery consisted of four tasks and was divided into two separate sessions that took place within 42 days. Children performed the Associative Inference (AI) and Transitive Inference (TI) task in session 1 and the Temporal Regularity (TR) and Categorization (CAT) task in session 2. Note that data from the TR and CAT tasks are partially used (for task procedure of the TR and CAT task see Supplementary Methods). The task order within each session and the session order were fixed across children. All tasks were written and presented in MATLAB (Version 9.7 [R2019b], https://de.mathworks.com/products/matlab) on a 1920x1080 resolution Monitor using the Psychophysics Toolbox (http://psychtoolbox.org/). Data collection took place between January and August 2022.

3.2.1. Associative Inference Task

The AI task was adapted from previous studies that assessed associative inference in children (Richmond & Pan, 2013; Schlichting et al., 2017). Stimuli consisted of 30 animals and 15 scenes from flaticon (https://www.flaticon.com/) that were divided into 15 ABC triads made up of two animals (A and C) and one scene (B). From every triad, there were two separate animal-scene pairings that were fixed across participants. The task consisted of three repetitions of encoding and direct memory test cycles, followed by an associative inference test (Figure 1). Children were told that they would meet a group of befriended animals who planned to go on a trip. They were instructed to watch closely where each animal wanted to go. Each encoding phase consisted of 30 animal-scene pairs. On each trial, children saw an animal on the left side and a scene on the right side of the screen for 3.5 seconds, followed by a 0.5 second cross-haired inter-trial interval (Figure 1, top). To help maintain children's attention, cheerful and child-friendly instrumental music was played in the background during each encoding phase. The order of AB and BC pairs in the encoding phase was pseudorandomized such that all AB pairs were shown first followed by all BC pairs in the same order based on their triad membership. Each of the encoding phases had a different trial order. However, the order of all encoding trials was fixed across participants. Children were not informed that there was a relation between the AB and BC pairs within the same triad, nor that they would later be asked to make any inference judgments.

After each encoding phase, a direct associative memory test followed which consisted of a self-paced 3 alternative-forced-choice (AFC) test with all 30 animal-scene pairs from encoding (Figure 1, bottom left). On each trial, an animal appeared at the top of the screen, with three scenes shown below. Children were asked to point to the scene the animal had "wanted to visit". The choices included one target scene and two lure scenes that were previously paired with different animals. The experimenter recorded the child's response on the keyboard with no corrective feedback provided. The trial order for the direct associative memory test differed between blocks but were fixed across all participants.

Immediately after the three repetitions of encoding and direct memory test, children performed an associative inference test that measured their ability to identify indirectly related animal pairs via a shared scene (which animal [A] goes to the same scene [B] as the probe animal [C], see Figure 1). Each indirect pair of a triad was tested, resulting in 15

associative inference trials. Children were told to choose which of the three animals at the bottom of the screen they wanted to take a trip to the same place as the probe animal presented at the top of the screen. The cue animal on the top was always the C item of a triad (i.e. shown second in every encoding block). The three options included a target (the animal from the same triad as the cue) and two lures (animals from different triads). No feedback was given during the associative inference test. The trial order was fixed across all participants. A schematic depiction of the task procedure is shown in Figure 1.

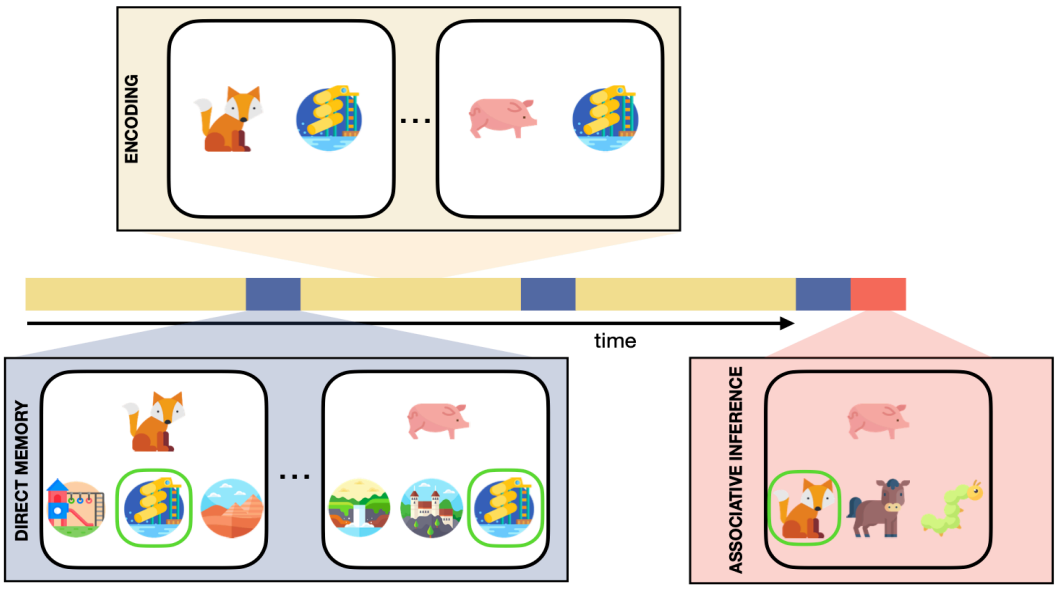


Figure 1. A schematic depiction of the Associative Inference Task. During encoding (top), children learned 30 animal-scene pairs from 15 triads consisting of two animals and one scene. In the direct memory test (blue), children were instructed to select the correct scene for each animal from three options. After three cycles of encoding and direct memory test, children completed the associative inference test (bottom right), in which they chose the animal that wanted to go to the same place as the animal shown at the top of the screen, despite never having seen the two animals together. Correct options indicated here with green rectangles are for visualization purposes only.

3.2.2. Transitive Inference Task

Two additional 4-year-olds participated but were not included in the analyses for this task due to failure to complete the testing phase (n=1) and failure to comply with the experimenter's instructions (n=1), resulting in n=28 4-year-olds.

The structure for this task was adapted from previous studies that examined similar age ranges (Bryant & Trabasso, 1971; Townsend et al., 2010). Stimuli consisted of five cartoon human characters in different-colored clothing, created with CrazyTalkAnimator 3.0 (https://crazytalk.reallusion.com/animator.html). Each training trial consisted of a dynamic and engaging video featuring two characters jumping.

The task consisted of one practice round, two separate training phases, and a test phase (Figure 2). Children were first told that they would meet five friends who were participating in a jumping tournament, which entailed a series of one-on-one competitions. Prior to each one-on-one match, children were asked to point to the friend whom they thought would be a better (i.e., higher) jumper. After children responded, the experimenter recorded the response and showed an animation of the friends jumping at differential heights, which provided children with feedback on their pre-match judgment. They heard a cheering sound only on correct trials. In addition, a winner's badge was displayed on top of the winner. This feedback interval lasted 1.5 seconds and was followed by a .5 second ITI with a fixation-cross (see Figure 2A top for an example trial). One example trial, using cartoon dogs instead of human characters, preceded the main task in order to acquaint the children with the task procedure.

The main experiment consisted of two training phases and a test phase. In the first training phase, children were first shown each premise pair individually (AB, BC, CD, DE) in ascending order, with each pair repeating three times. In a second cycle, the first training phase continued until children reached 80% accuracy after a minimum of another 3 trials per pair. Note that this learning criterion excludes the first trial of the first cycle given that children's responses would be based on guessing. Between each premise pair, participants were given the opportunity to take self-paced breaks.

The second training phase immediately followed the first training phase and used the same procedure, except that the four premise pairs were presented in a random order. Training continued until children either (a) selected the correct winning jumper on six successive trials for each pair, or (b) achieved at least 80% accuracy for each pair across a minimum of 24 trials. If neither criterion was met after 50 trials, the second training phase was terminated.[1] In both training phases, the position of the winner was counterbalanced across trials. The order of training trials was held constant across participants for the minimum of trials needed to advance.

The test phase immediately followed the second training phase and consisted of a 2AFC task and an order reconstruction task (Figure 2B). The 2AFC task consisted of 40 trials: four premise pairs and six untrained transitive pairs (AC, AD, AE, BD, BE, CE), with each pair tested 4 times with varying positions across trials. On each trial, children were shown two characters and were asked to choose the better jumper based on what they had learned without any corrective feedback. One random order list of test trials was created and administered to all children.

[1] When data collection began, we observed that most 4-year-olds did not meet the criterion for the second training phase. At that time, the first training phase consisted of only one learning cycle until 80% accuracy was reached. To strengthen learning before the second training phase, we adapted the procedure to include two cycles. However, the poprotion of children meeting the second-phase criterion did not differ between the one-cycle and the two-cycle procedures. Therefore, the samples were pooled (see Supplementary Table 1 and Supplementary Figure S1).

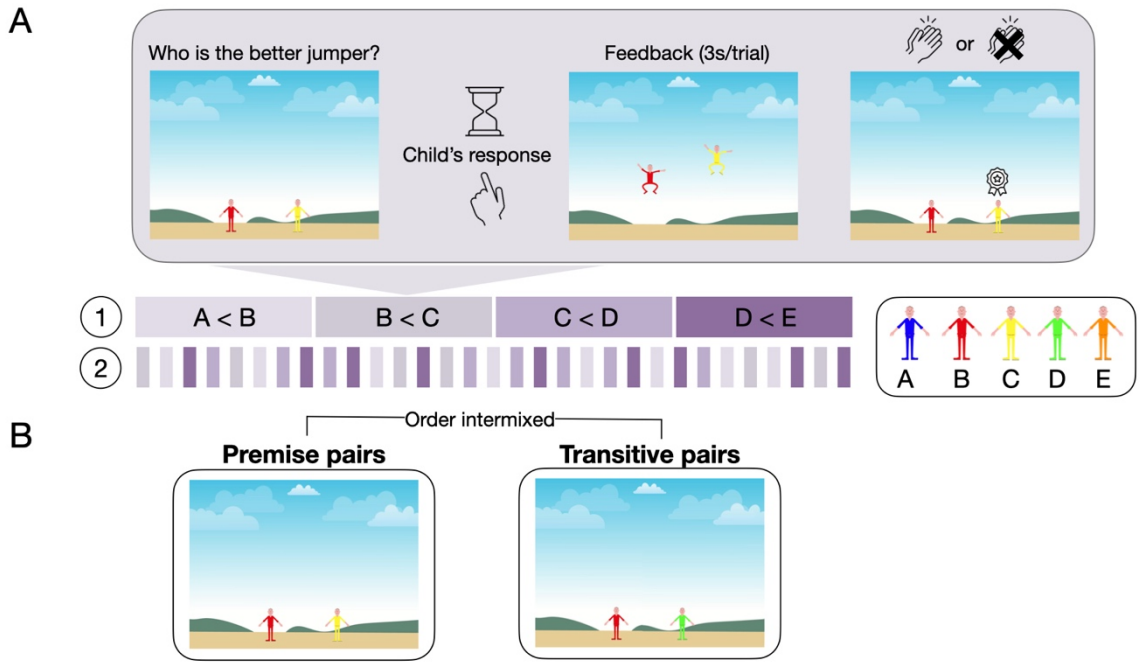


Figure 2. A schematic depiction of the Transitive Inference Task. ***(A)*** Training Phase: Children learned the relationship between five jumpers across individual premise (A<B, B<C, C<D, D<E). The first training phase continued until children reached 80% accuracy on each pair. In the second training phase, premise pairs were presented in an intermixed order, and training continued until children either selected the correct jumper on six consecutive trials per pair or achieved 80% accuracy per pair. Different training phases are denoted by the circled numbers on the left. ***(B)*** Test Phase: In the 2AFC task, children inferred the relationships non-adjacent, transitive pairs (e.g., BD) without corrective feedback.

3.3. Statistical analysis

All statistical analyses were conducted in R (Version 4.2.1; R Core Team, 2020) using RStudio (Version 1.3.1073; RStudio Team, 2020). A significance threshold of $\alpha = .05$ was applied to all analyses. Where appropriate, p-values were corrected for multiple comparisons using the Bonferroni-Holm method (Holm, 1979), and corrected values were compared against $\alpha = .05$. Accuracy was defined as the proportion of trials in which children responded correctly, with the exception of the AI Task where accuracy on direct memory was quantified as the proportion of triads in which children responded correctly either 2 out of 3 blocks or in the

last block. This additional metric was chosen to account for children who may have become fatigued in the last cycle, and those who required multiple repetitions to learn the associations. To assess group-level performance relative to chance, one-sided one-sample *t*-tests were conducted against chance levels of .5 for 2AFC and .33 for 3AFC tasks.

3.3.1. Generalization performance

To examine age differences, we fit general linear models (GLMs) for each of the four tasks with age group as a discrete predictor. Follow-up contrasts between parameter estimates for individual age groups from these linear models were conducted using the emmeans package (Lenth, 2022). To account for heteroscedasticity between age groups, robust standard errors were applied when contrasting parameters with the vcovHC function from the sandwich package (Zeileis, 2004; Zeileis et al., 2020) using the HC3 estimator (Long & Ervin, 2000). Further, to assess cross-task correlations, Pearson correlations between each pair of tasks and corresponding Bayes factors were calculated (Wagenmakers et al., 2016). To rule out that weak correlations were due to low reliability, split-half reliabilities and disattenuated correlations were calculated for each task (see Supplementary Methods).

3.3.2. Comparison of direct memory and inference performance

To test whether children differed between direct memory and inference performance in the AI and TI task, we fit linear mixed models (LMMs) using the lmer function from the lme4 package using restricted maximum likelihood estimation and the Nelder-Mead optimizer (Bates et al., 2015). We included participants as random intercept term to account for the dependency within participants and used age and pair type (direct vs. inference) as discrete predictors. Degrees of freedom were approximated using the Satterthwaite method. Discrete predictors were effect coded using the stats::contr.sum function to ensure main effect interpretation with interactions included. Significant effects were determined by investigating the respective type III sum-of-squares (SS) ANOVA-table using the stats::anova function. Follow-up contrasts between parameter estimates were conducted using the emmeans package (Lenth, 2022). To account for heteroscedasticity between age groups, robust standard errors were used when contrasting parameters using the vcovCR function from the clubSandwich package (Pustejovsky, 2022) using the CR2 adjustment type (Huang et al., 2022).

3.3.3. Dependence of inference on direct memory

Since some previous work suggests that generalization depends on direct memory of exemplars, while other work suggests alternative routes to generalization are possible, we tested whether inference relies on direct pair memory in both the AI and TI task. This measures the extent to which inference draws on specific pair memories. To evaluate this, we fit separate generalized linear mixed models (GLMMs) of the binomial family to the binary choice data in the inference test. Models were estimated with the glmer function from the lme4 package using a logit link function and the bobyqa optimizer (Bates et al., 2015). Participants were included as a random intercept to account for within-subject dependence. Age and direct pair memory (correct vs. incorrect) served as discrete predictors.

Direct pair memory in the AI task was defined analogously to direct memory accuracy: correct responses on both direct pairs of a triad in at least 2 out 3 blocks or in the final block. For the TI task, direct pair memory was coded as correct when all direct pairs required for a given inference pair was answered correctly on at least 3 of 4 trials in the test phase (e.g., for transitive pair AC, both AB and BC were answered correctly in ≥3 of 4 trials). Confidence intervals for odds estimates in GLMMs were based on asymptotic normality and denoted by CI_{asymp}. Categorical predictors were effect-coded using the stats::contr.sum function to allow interpretation of main effects in the presence of interactions. Significant effects were determined via type III Wald chi-square tests implemented in rstatix::Anova, and follow-up contrasts between age groups were conducted with the emmeans package (Lenth, 2022).

To explore age-related differences in reliance on direct memory for inference, we constructed 2x2 contingency tables reflecting of the marginal proportions of inference trials in which children were either correct (D+I+) or incorrect (D-I-) on both inference and the corresponding direct memory trials. Alternatively, participants could be only correct on one but not the other trial type, like directly remembering items, but fail at generalization (D+I-), or failing to directly remember, but succeed in generalization (D-I+; see Figure 5A). These tables allowed us to assess the degree to which memory and inference co-occurred. Direct memory accuracy for each inference pair was defined using the same metric as in the GLMM analysis. Participant-level contingency tables were first computed and then aggregated within age groups to form group-level tables.

We tested for effects using permutation tests on the proportion data. For each test, 10,000 permutations were generated, and the observed difference in proportions was compared against the null distribution of differences. Effects were considered statistically significant if the observed difference fell within the upper or lower 2.5% of the null distribution, corresponding to $p < .05$.

First, we examined whether age effects were present across all cells for each task separately, that is, whether children of different age groups differed in their overall response pattern within each task. Next, we focused on the off-diagonal cells, as these provided insights into the relationship between memory and inference. Specifically, the D+I- represents trials where children remembered the direct pairs but failed to use this knowledge for inference – a phenomenon we refer to as the knowledge-behavior gap (see also Blake et al., 2014, for a similar phenomenon in social psychology). Conversely, the D-I+ represents trials where children succeeded in inference despite lacking direct memory of the underlying pairs – a pattern we refer to as an alternative route to inference.

To characterize response patterns in these off-diagonal cells, we first asked: (1) whether the proportions of D+I- and D-I+ differed between tasks, and (2) whether these differences varied by age. We then computed a difference score between D+I- and D-I+, such that positive values indicated a pattern consistent with the knowledge-behavior gap, whereas negative values indicated a pattern consistent with the alternative route to inference. This difference score was used to test (a) whether one response pattern predominated within each task, (b) how the difference score varied between tasks, and (c) whether such variation differed by age.

3.3.4. Models of Transitive Inference

To test whether children relied on memory for direct pairs or instead generalized via value representations acquired through trial-and-error reinforcement learning, we formulated models based on each mechanism. These models examined how children generalized from experiences such as remembering that jumper A outperformed B and B outperformed C even though A and C never competed directly.

The reinforcement learning model (RL) assumes that participants update a value representation for each jumper based on whether their prediction of the higher jumper was correct. In contrast, pair-level learning models assume that participants directly store pairwise

relations and “stitch” these memories together to generalize from past to novel comparisons (Ciranka et al., 2022).

In RL models, value estimates (*Q*) are updated using a delta rule that adjusts the implicit value representation of the winning jumper *Q(i)* and the losing jumper Q(j) on trial *t* based on their relative difference:

$$d_t(i,j) = \eta[Q_t(i) - Q_t(j)], \qquad (1)$$

where η is a free parameter that determines how strongly the value difference influences learning. By introducing separate learning rates for the winning (α^+) and losing (α^-) jumper, the reinforcement learning model captures asymmetries in updating, which enables transitive inference through implicit value representations rather than direct memory for pairwise relations:

$$Q_{t+1}(i) = Q_t(i) + \alpha^+[1 - d_t(i,j) - Q_t(i)], \qquad (2)$$

$$Q_{t+1}(j) = Q_t(j) + \alpha^-[-1 + d_t(i,j) - Q_t(j)]. \qquad (3)$$

Finally, the RL model predicts a probability of choosing jumper i over jumper j using a softmax function of the value difference. The temperature parameter τ_{RL} governs the degree of stochasticity in choice behavior, with lower values leading to more deterministic choices and higher values producing greater randomness:

$$CP_{RL,t} = \frac{1}{1 + \exp(-(Q(i)_t - Q(j)_t)/\tau_{RL})}. \qquad (4)$$

The pair memory model (Pair) explicitly memorizes outcomes of jumper pairs. Unlike the reinforcement learning model, it does not rely on prediction errors to update implicit value estimates of each jumper. Instead, it tracks the outcome of observed pairs and stitches these memories together, based on their relative rank, to perform transitive inference.

For each neighboring pair n, the model records which jumper won and tracks a winning count in the variables U_n and L_n. When jumper U_n won against jumper L_n on a given trial, U_n is incremented by γ, a free parameter that captures memory fidelity with $\gamma < 1$ allowing for imperfect memory). If L_n wins, L_n is updated in the same fashion:

$$U_{n,t+1} = U_{n,t} + \gamma. \qquad (5)$$

The preference p, for U_n over L_n at time t ($p_{U>L,t}$) is then given by the proportion of remembered wins for each jumper:

$$p_{U_n>L_n,t} = \frac{U_{n,t}}{U_{n,t} + L_{n,t}}. \qquad (6)$$

This mechanism, however, only generates judgments for jumpers who directly competed against one another. Transitive inference becomes possible when the model incorporates memory recall process that stitches pair preferences, linking non-adjacent items. In the model this is implemented by summing the inferred preferences between jumpers i and j across all intermediate pairs; that is, by summing over preferences $p_{n,t}$ of the pairs connecting i and j:

$$p_{i>j,t} = \frac{\sum_{p_{n,t\in M}} (p_{n,t} - .5)}{|i-j|^{\lambda+1}} + .5, \qquad (7)$$

Here, $|i-j|$ denotes the true ranked distance between the jumpers. For example, when asked whether A is better than D, the model retrieves from memory that A was better than B, then that B was better than C, and finally that C was better than D. The parameter λ serves as a forgetting parameter, modeling failures to retrieve the linking pair preferences. We defined two variants of the pair memory model: (a) Pair_0 in which λ is fixed at a high value, allowing learning of neighboring pairs only and thus preventing transitive inference, and (b) Pair_+ in which λ is a free parameter. Note that when $|i-j| = 1$, equation (7) reduces to equation (6). To convert preferences into choice probabilities, we applied a logistic choice rule with additional pair decision noise τ_{pair}:

$$CP_{pair,t} = \frac{1}{1 + \exp(-p_{i>j,t}/\tau_{pair})} \qquad (8)$$

The models were fitted by minimizing the negative log-likelihood of the model given each participant's single-trial responses. To assess whether RL or pair memory better captured participants' choices, we compared models using the lowest Bayesian Information Criterion (BIC) to identify the model with the lowest BIC for every participant. Based on these BIC values, we then computed protected exceedance probability (pxp) within each age group to determine which of the models provide a better description of the majority of participants.

4. Results

4.1. Age differences across generalization tasks

We first examined age differences in generalization capacities from early to middle childhood. Mean accuracy on generalization trials by age group for all tasks is shown in Figure 3A. We note that because the CAT task was very extensive, we obtained insufficient data from 4-year-olds, as the majority were unable to complete it. Across tasks, all age groups performed above chance except for 4-year-olds on the TR Task (see Supplementary Table 3). On all tasks with sufficient 4-year-old data, 4-year-old were less accurate than both 6-year-olds and 8-year-olds. Further, on all tasks except the CAT and TR task, 6-year-olds performed significantly below the level of 8-year-olds (all $ps < .03$, see Supplementary Table 4). Together, these findings suggest consistent age-related differences in generalization, particularly between 4- and 6-year-olds, across multiple behavioral paradigms.

4.2. Low inter-task correlations question the notion of a common generalization factor

Another aim of this study was to investigate whether the tasks tapped into a common generalization factor. If such a factor existed, inter-individual differences in performance should be highly correlated across tasks. To examine this, we computed Pearson correlations of generalization performance across all tasks (Figure 3B, lower triangle). When collapsing across age, performance on AI, TI, and TR were significantly correlated. However, when controlling for age by calculating the correlations within each age group separately, the only remaining significant correlation was between the AI and TR task in 6-year-olds. This correlation did not survive correction for multiple comparisons. All correlations and Bayes factors are provided in Supplementary Table 5.

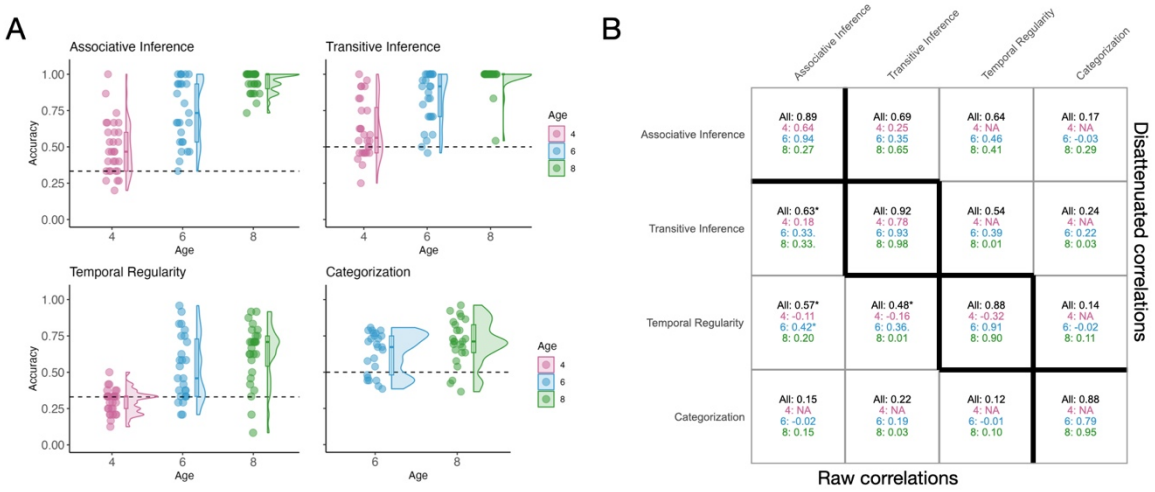


Figure 3. ***(A)*** Accuracy on individual generalization tasks by age group. Horizontal dashed lines indicate chance level for each task (0.33 for AI and TR, 0.5 for TI and CAT). Note that no data are available for 4-year-olds on the CAT task (see Supplementary Methods). ***(B)*** Correlation Matrix: Raw inter-task correlations across all four tasks are shown on the lower triangle, task reliabilities are shown on the diagonal, and disattenuated correlations are shown on the upper triangle. All coefficients are shown both collapsed across age (top, in black) and within age groups (bottom, color-coded). Note that 4-year-olds were excluded from correlations with the CAT task and from disattenuated correlations with the TR task due to negative reliability (see also Supplementary Table 10). Significance levels (only for raw correlations): $'.': p < .1, '*': p < .05$.

The low coherence between tasks may reflect the limited sensitivity of the measures to capture sufficient variance. Floor performance of 4-year-olds in the TR task and their inability to complete the CAT task, combined with the near-ceiling performance of 8-year-olds in the TR and AI tasks, suggest that these measures may not be well-suited for assessing generalization across this age range. Further, the weak correlations may reflect poor reliability. Therefore, we computed split-half reliabilities and disattenuated correlations (see Supplementary Methods). We found that reliabilities for all tasks were moderate to high (except for the TR task in 4-year-olds and AI task in 8-year-olds) and that reliability disattenuation did not change the correlations to a great extent (Figure 3B, diagonal and upper triangle), such that the observed correlations were not a result of low reliability.

These findings suggest that the tasks may not measure the same generalization ability. To explore other potential reasons for the lack of strong correlations, we focused on the AI and TI task. These tasks were selected because (1) they are often grouped under inferential reasoning skills (Zeithamova et al., 2012), and (2) they provide a direct mean of assessing the role of specific memories in supporting generalization.

4.3. Age- and task-dependent differences in direct memory and generalization performance across inference tasks

For both the AI and TI tasks, we compared mean accuracy across age groups for direct memory and inference trial types to test whether the two tasks followed similar age patterns (see Figure 4A, top row). Pair type was being either direct or inference, depending on the task. In AI, direct pairs referred to accuracy on direct memory pairs, whereas inference pairs referred to associative inference trials. In TI, direct pairs referred to premise pairs, whereas inference pairs referred to transitive test trials. Accuracy on direct pairs in AI was calculated by including only trials where children responded correctly on at least two out of three direct test blocks for a given animal, or on the last test block for that animal (see Methods for details).

We first compared performance between direct and inference pairs within each task using separate LMMs on the aggregated accuracy data. In the AI task, we observed a significant Age X Pair type interaction $\left(F(2,\ 81) = 5.75,\ p = .005,\ \eta_p^2 = .12\right)$. Follow-up comparisons showed that for both pair types, 4-year-olds were less accurate than to their older counterparts, and 6-year-olds were less accurate than 8-year-olds (see Supplementary Table 6). To identify the source of the interaction, we tested whether the effect of pair type differed between age groups. Children in all age groups performed significantly better on direct pairs than on inference pairs. However, this difference was less pronounced in 8-year-olds than in younger children ($\$ 4: \beta = 0.16, 95\% CI\ [0.1, 0.22], t(81) = 4.92, p_{Holm} < .001$; $6: \beta = 0.14, 95\% CI\ [0.07, 0.2], t(81) = 4.21, p_{Holm} < .001$; $8: \beta = 0.03, 95\% CI\ [0.01, 0.06], t(81) = 2.79, p_{Holm} = .013$). Note that 8-year-olds' accuracy was near ceiling, reducing sensitivity to detect differences between pair types. Nonetheless, the results indicate that (i) accuracy on both direct and inference pairs improves across early to middle childhood, and (ii) the performance gap between remembering specific pairs and inferring associations narrows with age (Figure 4B).

In the TI task, we found no Age X Pair Type interaction ($F(2,79) = 1.10,\ p = 0.337,\ \eta_p^2 = .03$). However, there were significant main effects of age ($F(2,79) = 33.05,\ p < .001,\ \eta_p^2 = .46$), and pair type ($F(1,79) = 4.00,\ p = .0490, \eta_p^2 = .05$). Overall, 8-year-olds outperformed both younger groups, and 6-year-old children outperformed 4-year-old children (all $p_{Holm}s < .001$, see Supplementary Table 7 for detailed results). Across age groups, accuracy on inference pairs was slightly higher than on direct ($\beta = 0.03, 95\% CI\ [0.00, 0.06], t(79) = 2.02, p_{Holm} = .047$; also see Figure 4A, right panel). These results suggest that in TI, children across ages tend to perform better on inference then direct pair memories, with overall performance increasing with age.

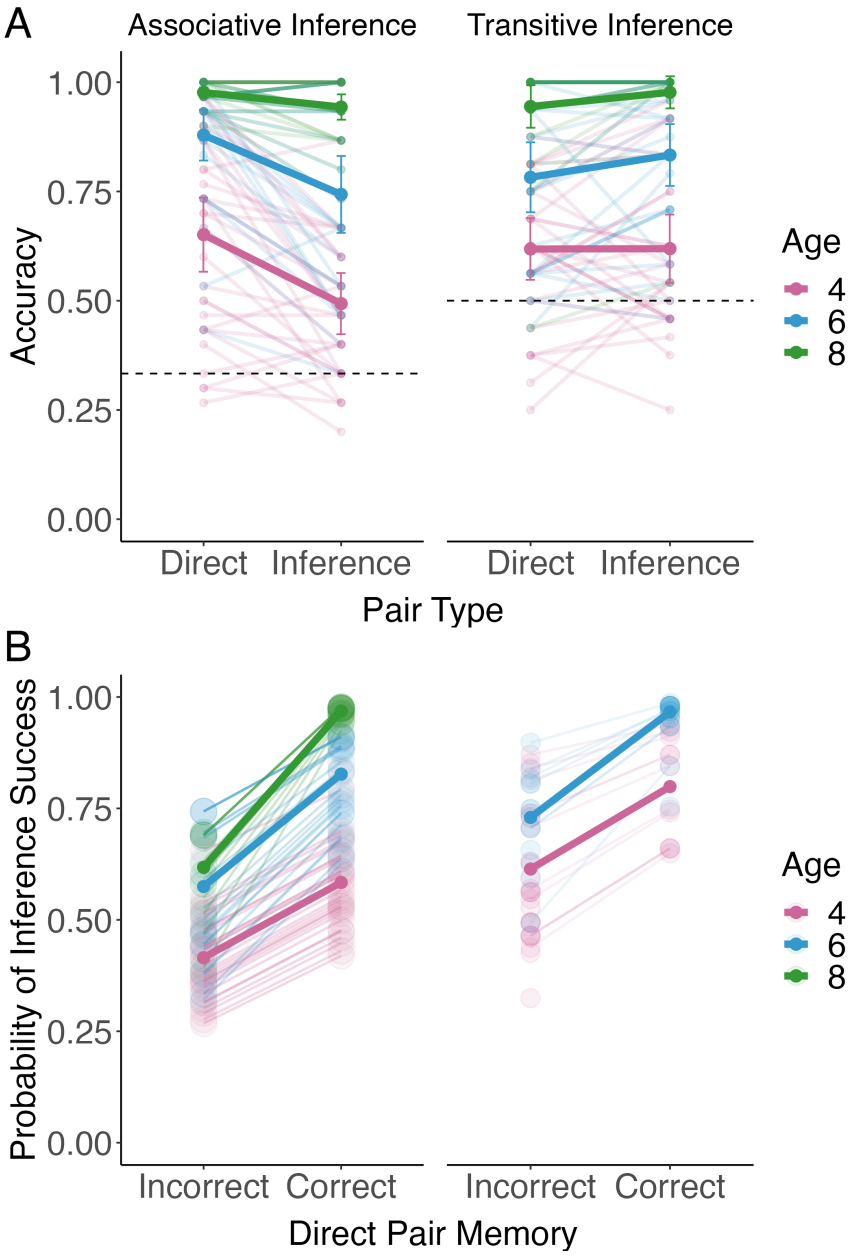


Figure 4. ***(A)*** Accuracy on direct and inference pair in the AI and TI tasks. In the AI task, accuracy was lower for inference than direct pairs, whereas in the TI task, there was a tendency for the reverse. Dotted lines denote chance level for each task. Error bars show 95% CIs of the standard error of the mean on the group level. ***(B)*** Predicted probability of inference success from GLMMs, conditional on whether the underlying direct pair memory was correct. In both tasks, remembering the required direct pairs increased the probability of successful inference. Note that 8-year-olds were excluded from TI analysis due to ceiling performance. Solid dots and lines represent age group means, faded dots and lines represent individual participants.

Interestingly, although age-related differences were similar across tasks, the performance patterns between tasks differed. In the AI task, performance was overall better on direct than inference pairs, while the opposite was true for the TI task. This suggests that remembering direct pairs is easier in the AI task, while inferring pair relationships is easier in the TI task.

4.4. The dependence of inference on directly learned pairs varies by age

Subsequently, we examined whether children's success in making inferences was contingent upon their recollection of direct experiences and whether this relationship differed by age. Furthermore, if children exhibited differential reliance on direct memory for inferences in one task but not in the other, this would help explain the weaker-than-expected correlation between the tasks. To this end, we implemented separate GLMMs for each task, incorporating direct pair memory (an index of memory for the relevant direct pairs, see Methods for a detailed description) and age as discrete predictors for investigating inference success on a trial-by-trial basis.

Predicted probabilities from the GLMMs are shown in Figure 4B (bottom row). For the AI task, we found a significant interaction between age and direct pair memory ($\chi^2(2) = 13.8,\ p = .001$). To examine this interaction, we performed two sets of follow-up comparisons using the estimated marginal means. First, we tested whether the effect of age differed depending on whether the underlying direct pairs were remembered. We observed that the odds of inference success did not differ between age groups when the pairs were not remembered (all $p_{Holm} > .2$, also see Supplementary Table 8). However, when the underlying pairs were remembered, older children consistently showed higher odds of success compared to younger counterparts ($8-6: odds = 6.43,\ 95\%CI_{asym}[3.12,13.28],\ z = 5.04,\ p_{Holm} < .001$; $8-4: odds = 21.9,\ 95\%CI_{asym}[10.489,45.74],\ z = 8.22,\ p_{Holm} < .001$; $6-4: odds = 3.4,\ 95\%CI_{asym}[1.91,6.06],\ z = 4.16,\ p_{Holm} < .001$; see Supplementary Table 8 for all follow-up comparisons). Second, we investigated whether the effect of direct pair memory itself differed by age group. Across all age groups, remembering both direct associations in a triad increased the odds of inference success (Figure 4B left plot; see Supplementary Table 9 for detailed results). This effect was significantly stronger for 8-year-olds than for their younger counterparts ($8-6: odds = 5.38,\ 95\%CI_{asym}[1.49,19.46],\ z = 2.57,\ p_{Holm} = .021$; $8-4: odds = 9.61, 95\%CI_{asym}[2.83,32.65],\ z = 3.63,\ p_{Holm} =$

$.001$) . In contrast, 6- and 4-year-olds did not significantly differ ($odds = 1.79$, $95\%CI_{asym}$ $[0.86,3.72]$, $z = 1.55$, $p_{Holm} = .121$). Taken together, these results indicate that when children failed to remember the underlying pairs, their performance was similar across all ages. However, 8-year-olds benefited more from having intact memories of the direct pairs, suggesting that they relied more heavily on directly learned information than 6- and 4-year-olds when performing associative inference.

For the TI Task, we added an additional independent variable, distance, to index how far apart two items of a transitive pair are in the hierarchy (e.g., two items separating A and D), as we were interested in how the distance might affect inference success. Thus, we included direct pair memory, distance, age, and their pairwise interactions as fixed effects in a GLMM, using subject as a random intercept. Because nearly all 8-year-olds scored perfectly on transitive pairs (see Figure 4A, right plot), we did not include them in this analysis. We found no significant interaction between distance and direct pair memory (distance x direct pair memory: $\chi^2(1) = 0.00$, $p = .947$), however age interacted with both other predictors (age x direct pair memory: $\chi^2(1) = 7.71$, $p = .006$; age x distance: $\chi^2(1) = 5.89$, $p = .015$). To explore these interactions, we performed follow-up comparisons on the estimated marginal means. First, we examined the interaction between direct pair memory and age. Both age groups showed a greater probability of success on transitive inference if they remembered the required direct pairs ($4: odds = 2.48, 95\%\ CI_{asymp}$ $[, 5.07], z = 2.49$, $p_{Holm} = .013$; $6: odds = 10.47, 95\% CI_{asymp}$ $[4.53,24.21], z = 5.49, p_{Holm} < .001$). However, this effect was stronger in 6-year-olds than in 4-year-olds ($odds = 4.22$, $95\%CI_{asymp}$ $[1.53,11.67]$, $z = 2.78$, $p_{Holm} = .011$; also see Figure 4B right plot). Put differently, when children did not remember the required direct pairs, 4- and 6-year-olds did not significantly differ in their transitive inference success ($odds = 1.73$, $95\%CI_{asymp}$ $[0.89,\ 3.37]$, $z = 1.63, p_{Holm} = .10$). However, when these pairs were remembered, 6-year-olds outperformed 4-year-olds ($odds = 7.32$, $95\%CI_{asymp}$ $[2.79,\ 19.22]$, $z = 4.04, p_{Holm} < .001$). Thus, while direct pair memory benefits transitive inference in both groups, its influence appears stronger in 6-year-olds.

Next, we investigated the interaction between distance and age. Distance was positively associated with transitive inference performance in 6-year-olds ($odds = 2.10$,

$95\%CI_{asymp}$ [1.19,3.70], $z = 3.14$, $p_{Holm} = .005$), but not in 4-year-olds ($odds = 1.16$, $95\%CI_{asymp}$ [0.68,1.99], $z = 0.68$, $p_{Holm} = .50$). Moreover, this difference between age groups was significant ($odds = 1.80$, $95\%CI_{asymp}$ [1.01,3.23], $z = 2.43$, $p_{Holm} = .031$). These findings indicate that only 6-year-olds benefit from a greater distance between jumpers when making transitive inferences, regardless of whether they remember the required direct pairs.

Notably, across both tasks, remembering the underlying information benefited all age groups, although the benefit was smaller in younger children. This pattern suggests that although direct pair memory improves inference success in general, their link is weaker in younger children. Consequently, younger children derive less advantage from direct memory when making inferences.

4.5. Age-related differences in the link between direct memory and inference suggest distinct underlying mechanisms

We next aimed to understand the link between direct pair memory and inference by examining contingency tables that show the proportion of inference trials for which (a) both direct memory and inference were correct (D+I+), (b) both were incorrect (D–I–), (c) only direct memory was correct (D+I–), or (d) only inference was correct (D–I+; see Figure 5A). By comparing these outcomes, we sought to determine whether the weaker link between direct memory and inference stems from (1) having direct memory but failing to make the corresponding inference (“knowledge–behavior gap”; D+I–), or (2) successfully making inferences without intact direct memory (“alternative route to inference”; D-I+). Such patterns would suggest that different mechanisms govern performance on these tasks.

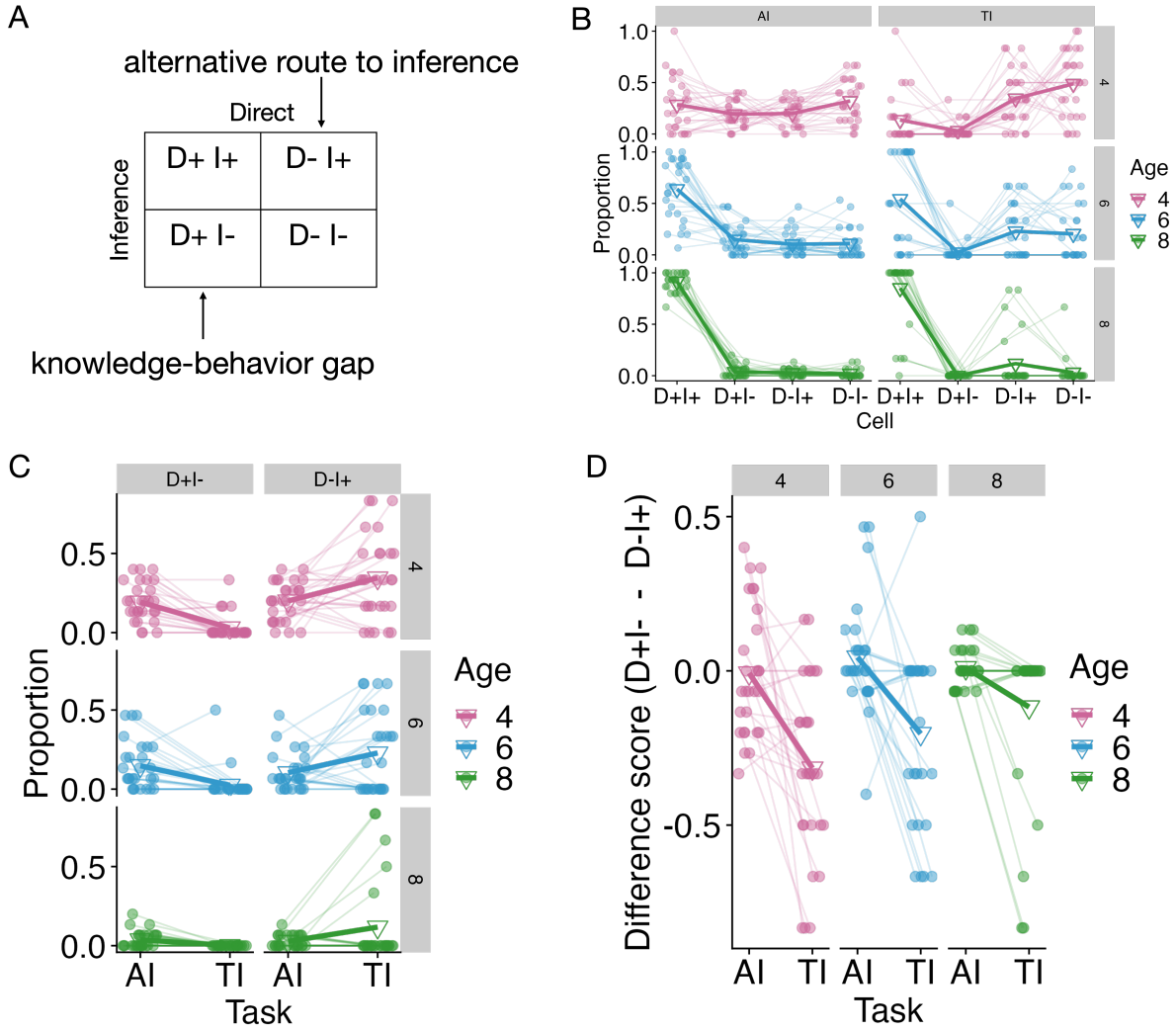


Figure 5. ***(A)*** Contingency tables for both the AI and TI task and for each age group individually. The values within these cells reflect the proportion of inference trials in which direct pair memory for the corresponding inference trial was correct or not (horizontal) and inference was correct or not (vertical). Note the off-diagonal cells that inform about the possible reasons for a weaker link between direct pair memory and inference, namely via knowledge-behavior gap or alternative route to inference. ***(B)*** Proportions in the individual cells divided by age and task. ***(C)*** Off-diagonal cells and their difference in proportions between tasks. ***(D)*** Difference score from subtracting D-I+ from D+I- in the individual tasks to determine the driving factor for the weaker link in younger children. Positive values denote higher proportions in the D+I- while negative values higher proportions in the D-I+ cell. The negative value in younger children in the TI task indicates that the weaker link stems from an alternative route to inference. ***(B)-(D)*** Solid triangles and lines denote age group averages, faded dots and lines denote individual participants.

The proportions in each cell, broken down by age and task, are shown in Figure 5B. For a detailed description of age-related differences in these cells, please refer to the Supplementary Results. To explore how the off-diagonal cells (D+I–, D–I+) vary by task and age, we first compared the proportions between tasks.

We found that, in the D+I– cell, proportions were higher in the AI task compared to the TI task, whereas the reverse was true in the D–I+ cell (both $p < .01$). Next, we examined whether these differences depended on age. Across all age groups, D–I+ proportions were similarly higher in the TI task relative to the AI task. However, for the D+I– cell, the difference between the associative and transitive tasks was larger for 4- and 6-year-olds than for 8-year-olds (both $p < .012$; see Figure 5C).

Overall, these results suggest that the alternative route to inference is more pronounced in the TI task across all ages, while the knowledge–behavior gap is more pronounced in the AI task only among 4- and 6-year-olds compared to 8-year-olds. This distinction highlights that younger children's weaker link between memory and inference may arise from different mechanisms in the two tasks.

In a final step, we investigated the driving factor behind the weaker link between direct pair memory and inference in younger children by comparing the proportions of the off-diagonal cells within each task and examining whether these differences vary between tasks and across ages. Specifically, we created a difference score by subtracting the proportion in the D–I+ cell from the D+I– cell (Figure 5D).

First, examining the difference score across both tasks, irrespective of age, we found that this score was negative in the TI task, indicating the proportions in the D-I+ cell were consistently higher than in the D+I- cell ($p < .001$). In contrast, in the AI task this score was not significantly different from zero ($p = .211$). Comparing the two tasks showed that the difference score was significantly more negative in the TI than in the AI task ($p < .001$).

Next, we analyzed how the difference score varied with age. In the TI task, 4-year-olds showed more negative scores than 8-year-olds ($p = .005$) and a trending difference compared to 6-year-olds ($p = .079$; all other $p > .13$). Finally, testing whether the difference score varied differently between the tasks for different age groups revealed that 4-year-olds' difference in scores between tasks was significantly greater than 8-year-olds' ($p = .020$), with a non-

significant trend between 6- and 8-year-olds ($p = .082$). In other words, 4-year-olds displayed a greater disparity between the TI and AI tasks than 8-year-olds, while 6-year-olds showed a similar trend that did not reach significance.

From these findings, we concluded that in the AI task, the lack of a difference in proportions makes it difficult to pinpoint whether the knowledge–behavior gap or alternative route explains the weaker link between direct memory and inference. However, in the TI task—particularly for younger children—this weaker link appears to stem from an alternative route to inference (i.e., correct inferences in the absence of direct pair memory). Thus, these results suggest that younger children may rely on an alternate mechanism to perform transitive inferences without having to rely directly on memory for individual pairs, a possibility we further explore in the following section.

4.6. Alternative route to transitive inference by value assignment prominent in younger children

To better understand the alternative route to inference in the TI task, we explored which mechanisms might explain how children solve these problems without relying solely on pair memory. One such mechanism rests on "functional reasoning" (Piaget et al., 1968/1977; Chapman & Lindenberger, 1988, 1992b) and can be modelled by assigning a value to individual items. This can be contrasted with strategies that instead rely on remembering each pair of items and their neighbors in the transitive structure. To this end, we used three of the models described in Ciranka and colleagues (2022) to investigate whether children utilize item values via reinforcement learning, rather than relying on pair memory.

These three models include one reinforcement learning model (RL), which applies distinct learning rates for winning and losing items and uses relative value differences to regulate value updates (Ciranka et al., 2022), and two pair memory models that differ in how they use pair-level information. In the first pair memory model (Pair_0), the system can only learn about direct pairs and cannot support inferences, whereas the other pair memory model (Pair_+) links direct pair memories, facilitating transitive inference (see Methods).

The modeling results indicate that children varied in their reliance on pair memory and reinforcement learning across age groups (Figure 6C). For 8-year-olds, model comparison using BIC and pxp values revealed that the pair memory model Pair_+ provided the best fit

(51.12 ± 4.20; 0.66). For both 6- and 4-year-olds, however, model RL fit best according to BIC ($6: 101.47 \pm 4.18$; $4: 139.56 \pm 2.93$) and pxp ($6: > .99$; $4: 0.75$). We also note that pxp values for 8- and 4-year-olds suggest possible heterogeneity within these groups (see Figure 6C), indicating the best-fitting model may not fully capture every child's individual behavior. Nonetheless, simulated responses generated by the best-fitting models closely matched the observed mean response patterns (Figure 6A, B).

These findings shed light on the alternative route to inference in younger children. Among 8-year-olds, the best-fitting model ($Pair_+$) uses pairwise item associations to infer transitive relations—a logical benchmark for children who are near ceiling performance. Younger children, however, appear to rely more on reinforcement learning of item values instead of pair memory when they cannot otherwise infer transitive relationships (but see 4-year-olds being partly explained by model $Pair_0$ that can only learn direct pair comparisons and is the most basic model). Thus, a higher proportion of correct inferences without correct direct-pair memory in younger children appears to stem from this alternative value-based mechanism, allowing them to perform transitive inferences without explicitly recalling underlying item pairs.

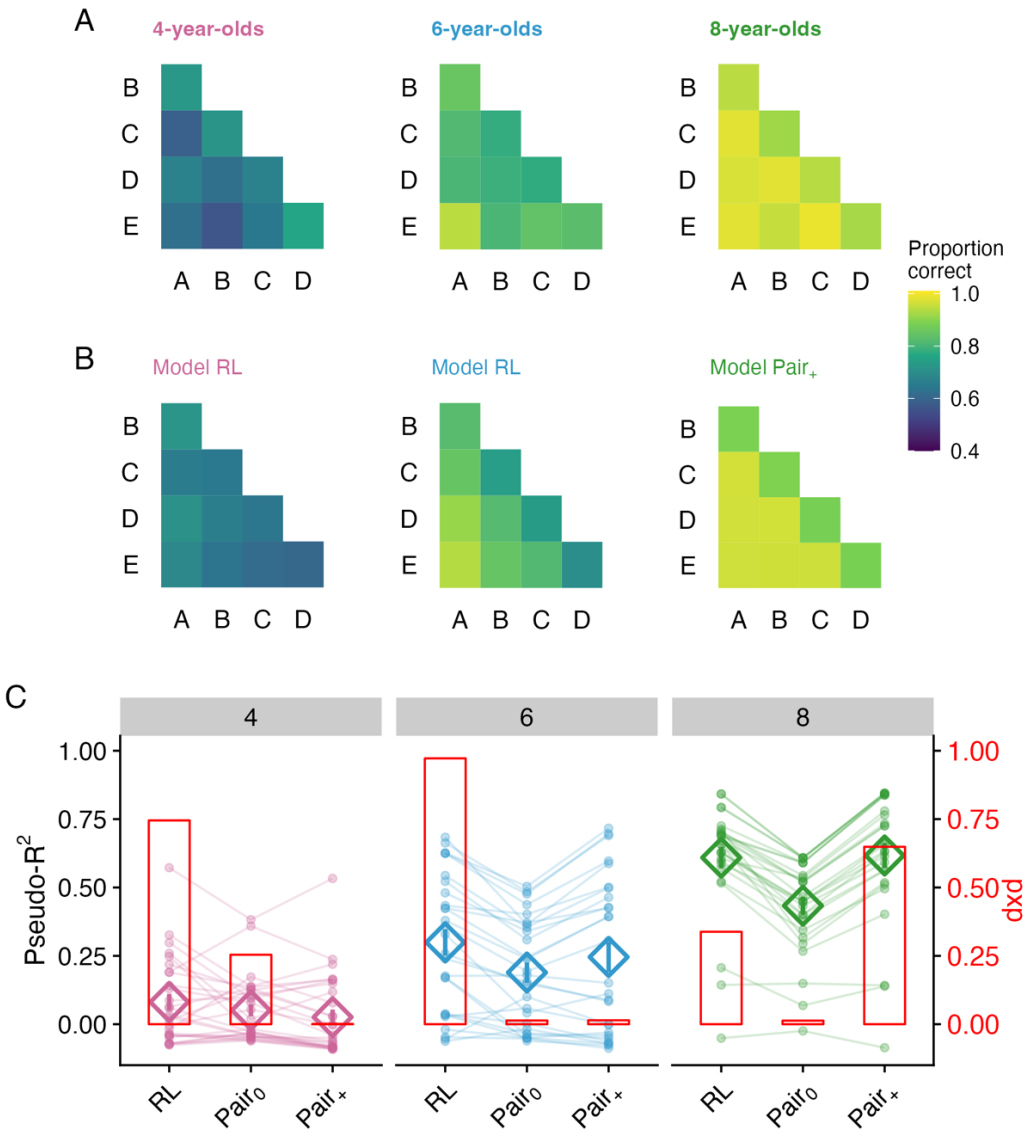


Figure 6. ***(A)*** Mean proportion correct across all trials in the experiment divided up by pair and age group. ***(B)*** Mean proportion correct simulated for each age group separately using the best-fitting model and the corresponding parameter estimates. ***(C)*** Model comparison for each age group individually. Pseudo-R^2 is inversely related to BIC, where larger values indicate a better fit (left y-axis: diamonds show mean value across children in a given age group for the respective model, error bars show mean ± sem, faded dots show individual participants, and faded lines connect a single participants' Pseudo-R^2 for one model to that of the next). The red bar graphs that are superimposed denote the protected exceedance probability (pxp) of each model, meaning the probability that a model describes the majority of participants best (right y-axis).

5. Discussion

There has been a long-standing interest in understanding generalization in both animals and humans, with recent efforts seeking to integrate diverse conceptualizations of this capacity (Ghirlanda & Enquist, 2003; Taylor et al., 2021; Wu et al., 2024). In this study, we examined the development of generalization by (1) assessing age-related differences across several tasks, (2) testing whether inter-individual differences in performance were correlated across tasks in early and middle childhood, and (3) probing the link between memory and generalization in two inference tasks. Four key findings emerged.

First, remembering directly learned associations benefitted both inference tasks, but performance was not correlated between them, and response patterns differed for direct versus inference pairs. Although both tasks are often assumed to rely on a similar mechanism (Zeithamova et al., 2012), our data revealed only a weak association, even after reliability disattenuation, potentially reflecting distinct underlying processes. Notably, accuracy was higher for direct pairs in the TI task, while the reverse was true for the AI task, further underscoring their potential mechanistic differences. From a neurocomputational perspective, both tasks can be solved by the mechanism proposed in REMERGE (Kumaran & McClelland, 2012). However, to our knowledge, no prior study has tested whether behavioral performance on both tasks is related. A recent meta-analysis reported an overlapping activation cluster in the left hippocampus for both associative and transitive inference (Zhang et al., 2022), suggesting the recruitment of similar neural regions for both tasks. Reverse inference from neural correlates alone should always be taken with caution though (Poldrack, 2006), highlighting the need for converging behavioral evidence.

Second, memories of specific pairs benefited both associative and transitive inference accuracy. The current results do not align with Richmond and Pan's (2013) findings, which showed no between-person association between direct memory and associative inference in 3-5-year-olds. Here, we found that within-person, trial-level analyses revealed a stronger link. Methodological differences likely explain the discrepancy: Richmond and Pan (2013) examined between-person differences, while we analyzed within-person variations on a trial-by-trial level. Other studies looking at trial-by-trial variations have also shown that generalization is tied to the memory for specific instances in children (Buchberger et al., 2024;

Karjack et al., 2025; Ngo et al., 2021). For example, memory for character-object pairings increased the likelihood of generalizing which new item a character would select (Buchberger et al., 2024; Ngo et al., 2021). In another study that used an associative inference paradigm to investigate knowledge derivation based on overlapping facts, Bauer and San Souci (2010) showed that children integrated knowledge more readily when related facts are remembered correctly. Together, these findings demonstrate that direct memory supports generalization across multiple inference tasks.

Third, the degree to which direct memory benefits generalization in inference tasks increased with age. Older children drew on prior experiences more efficiently than young children. Two observations support this claim. First, among trials for which directly learned pairs were remembered, younger children were still less likely to infer correctly than their older counterparts. And second, the increase in probability of inference success between incorrect and correct pair memory was stronger in older compared to younger children. This suggests that direct memory and inference is more tightly linked among older children, potentially reflecting differential developments of the neural underpinnings of memory and generalization throughout childhood. Neurocomputational models have proposed that the hippocampal subfields CA3 and dentate gyrus are involved in memory for specific relational details (Norman & O'Reilly, 2003; Schapiro et al., 2017), whereas CA1 is involved in generalization and inference (Schapiro et al., 2017; Sučević & Schapiro, 2023).These subfields mature at different rates (Keresztes et al., 2017; Lavenex & Banta Lavenex, 2013; Riggins et al., 2018), with CA1 developing earlier than CA3 and DG. This asynchronous maturation may limit the extent to which these subregions cooperate in early development. In adults, there is neural evidence of both integrated representations in CA1 and differentiated representations in CA3 and dentate gyrus (Molitor et al., 2021; Schlichting et al., 2015). Recent reinforcement learning studies suggest that from late childhood to adulthood, both general and specific representations guide decisions (Nussenbaum & Hartley, 2025). Thus, it is likely that over development, children become better at forming differentiated representations which are in turn also utilizable for episodic inferences.

Fourth, the weak association between memory and inference in younger children during transitive inference stems from accurate inference despite poor memory for the constituent pairs. A weak coupling between generalization and memory specificity can result from two

types of mismatches: (1) successful generalization despite inaccurate memory specificity (alternative route), or (2) unsuccessful generalization despite accurate memory specificity (knowledge-behavior gap). In the AI task, these patterns were relatively balanced. However, in the TI task, successful generalization in the absence of accurate memory was prevalent in younger children, suggesting that they may have relied on a different strategy to solve the task (cf. Chapman & Lindenberger, 1988). Employing a computational model revealed that the alternative route involved using individual item values, rather than relational pair memory, to guide choices.

Computational modelling has the potential to provide insights that are hard to derive from behavioral data alone (Schlesinger & McMurray, 2012). In the present case, comparing a reinforcement learning model to a memory model that can in principle both perform transitive inference across age groups revealed important differences. 8-year-old childrens' behaviour was best explained by a model that relies on memory for individual pairs and linking them for inference ($Pair_+$). 4- and 6-year-olds' performance were best explained by a reinforcement learning model akin to "functional reasoning" (Chapman & Lindenberger, 1988; Piaget et al., 1968/1977) that performs transitive inference without requiring that children explicitly learn the transitive structure. This suggests that different age groups may use strategies with different complexity when performing the same transitive inference task. However, because 8-year-olds' performance was near ceiling, the findings in this age group should be interpreted cautiously: the $Pair_+$ model appears to account for near-perfect performance, but limited variance makes it restricts potential conclusions. Model comparison also suggests that part of the 8-year-old age group might be explained by the RL model and, therefore, hence we cannot conclusively disentangle both models. Similarly, 4-year-olds may in part be explained by $Pair_0$, the model that only learns item pairings and cannot perform inference. Therefore, it seems that strategies for solving the transitive inference task are more heterogeneous in younger children, in line with similar recent findings in younger participants (Giron et al., 2023; Schuck et al., 2022).

It is still actively debated whether encoding- or retrieval-based mechanisms predominantly support inference (for an extensive review, see Zeithamova et al., 2012). One prominent theory, integrative encoding, suggests that new information is encoded together with existing memories given overlapping elements (Shohamy & Wagner, 2008; Zeithamova & Preston,

2010). In contrast, retrieval-based accounts propose that memory representations are coded individually in an episodic-like way, including their relationship to one another (O'Reilly & Rudy, 2001). These are then retrieved together from partial input by means of pattern completion for generalization (McClelland et al., 1995; Norman & O'Reilly, 2003). Assigning values to stimuli and using them to guide future choices can be seen as a type of encoding mechanism, while linking memories of pair relationships at test as a retrieval mechanism. A recent study that examined whether integrated representations of overlapping events are used to perform associative inference found no evidence for the encoding of such integrated representations in children (Schlichting et al., 2022). Similarly, another study investigating associative inference using eye tracking to probe whether memory integration took place, found that only adults showed signs of integration, while children did not (Bauer et al., 2021). Relating this to the current study, it is likely that integrated memories were not formed, but that children used separate memories at retrieval in the AI task. We speculate that this underscores one divergence between the two tasks used here: successful associative inference relies predominantly on retrieval mechanisms, whereas successful transitive inference, at least in 4- and 6-year-olds, relies predominantly on an encoding mechanism of associating items with values, bypassing the use of individual memories (also see Berens & Bird, 2022, for an encoding-based account of transitive inference in adults).

Finally, despite similar age-related trends across tasks, the results do not support generalization being a unitary construct. Age-related increases in performance persisted on nearly all tasks and most consistently from ages 4 to 6. This task-invariant age pattern corroborate previous findings that used single-task designs in developmental samples: 4-year-olds performed better than chance on both the transitive (Bryant & Trabasso, 1971; Kallio, 1982) and associative inference (Richmond & Pan, 2013) task, while temporal regularity performance did not exceed chance until the age of 6 (Raviv & Arnon, 2018). One possibility for the lack of inter-task associations within individual age groups could have been limited task reliability. However, we found that reliabilities were high across the entire sample and only tended to be lower in 4-year-olds, and even there the influence on the disattenuated correlation was minimal (excluding the negative reliability in the TR task due to chance performance). Therefore, we ruled out low reliability as reason for not finding strong inter-task correlations. A different possibility is that our sample size was too small. However, if the

tasks measure a common mechanism, we would have expected to detect a strong effect even with our limited sample. Lastly, there indeed might not be a common factor of generalization as expected. This mirrors recent findings in exploration and risk-taking, where indicators previously thought to capture these constructs do not cohere and rather measure differing aspects of these constructs (Anvari et al., 2024; Pedroni et al., 2017). In the current case, the basis for using these tasks mainly comes from a variety of neuroimaging studies in adults (reviewed in Zeithamova & Bowman, 2020) and accompanying neurocomputational models (Kumaran & McClelland, 2012; Schapiro et al., 2017; Sučević & Schapiro, 2023) suggesting involvement of the hippocampus in rapid generalization. Direct behavioral evidence supporting a common underlying construct is scarce, though, in both adults and children. Thus, important differences between tasks indicate that generalization may be more multifaceted than previously assumed. However, further studies are needed to corroborate these claims.

Several limitations should be noted. First, task procedures were modified mid-study for the TR and TI tasks. Although no performance differences were found between procedural versions, small sample sizes may impair confidence in this conclusion. Second, the study was not designed to support joint computational modelling of transitive and associative inference or any other tasks and, thus, we could not implement parallel models across tasks. Future work should thus design comparable generalization tasks to allow direct model-based comparisons that is better suited to inform about different mechanisms of generalization. Third, in addition to the above-mentioned possibilities for the null result, the combination of measures might not have been sensitive enough to capture sufficient variability across tasks. Having encountered both floor and ceiling effects and substantial drop out from the youngest age group in the CAT task tied with previously reported low reliability in the TR task (Arnon, 2020) and in the current 4-year-old sample indicate measurement problems. Task difficulty and age-appropriateness may also have varied, obscuring potential inter-task associations.

In sum, the present results highlight early to middle childhood as an important period for the development of generalization capacities. However, generalization might not be a unitary process: different tasks appear to rely on distinct cognitive and, potentially, neural mechanisms. A closer examination of task designs and the application of multi-indicator approaches will be crucial for understanding the multifaceted nature of generalization across

development. Further, we find consistency in that children's inference is tied to memories for the underlying information and that this association becomes stronger with age. This suggests a maturational process of brain regions responsible for these processes working more closely together later in development. Moreover, the modeling results suggest that younger children's success in transitive inference may rather arise from item-value-based rather than pair-memory-based strategies. Applying computational models across multiple generalization tasks within individuals will be fruitful to determine if generalization is task-specific across ontogeny and to which degree it still reflects a unified process. Together, the present findings suggest that the capacity for generalization develops along multiple routes in early and middle childhood, based on cognitive prerequisites that evolve during ontogeny.

6. References

Anvari, F., Billinger, S., Analytis, P. P., Franco, V. R., & Marchiori, D. (2024). Testing the convergent validity, domain generality, and temporal stability of selected measures of people's tendency to explore. *Nature Communications*, *15*(1), 7721. https://doi.org/10.1038/s41467-024-51685-z

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68–81. https://doi.org/10.3758/s13428-019-01205-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bauer, P. J., Cronin-Golomb, L. M., Porter, B. M., Jaganjac, A., & Miller, H. E. (2021). Integration of memory content in adults and children: Developmental differences in task conditions and functional consequences. *Journal of Experimental Psychology: General*, *150*(7), 1259–1278. https://doi.org/10.1037/xge0000996

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), Article 4. https://doi.org/10.1016/j.jecp.2010.05.012

Bengfort, T., Hayat, T., & Göttel, T. (2022). Castellum: A participant management tool for scientific studies. *Journal of Open Source Software*, *7*(79), 4600. https://doi.org/10.21105/joss.04600

Berens, S. C., & Bird, C. M. (2022). Hippocampal and medial prefrontal cortices encode structural task representations following progressive and interleaved training

schedules. *PLOS Computational Biology*, *18*(10), e1010566. https://doi.org/10.1371/journal.pcbi.1010566

Blake, P. R., McAuliffe, K., & Warneken, F. (2014). The developmental origins of fairness: The knowledge–behavior gap. *Trends in Cognitive Sciences*, *18*(11), 559–561. https://doi.org/10.1016/j.tics.2014.08.003

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*, e59360. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization. *The Journal of Neuroscience*, *38*(10), Article 10. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Brainerd, C., & Reyna, V. F. (1992). The memory independence effect: What do the data show? What do the theories claim? *Developmental Review*, *12*(2), 164–186. https://doi.org/10.1016/0273-2297(92)90007-O

Bryant, P. E., & Trabasso, T. (1971). Transitive Inferences and Memory in Young Children. *Nature*, *232*(5311), Article 5311. https://doi.org/10.1038/232456a0

Buchberger, E. S., Joechner, A., Ngo, C. T., Lindenberger, U., & Werkle-Bergner, M. (2024). Age differences in generalization, memory specificity, and their overnight fate in childhood. *Child Development*, *95*(4). https://doi.org/10.1111/cdev.14089

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage in the development of transitivity. *Developmental Psychology*, *24*(4), 542–551. https://doi.org/10.1037/0012-1649.24.4.542

Chapman, M., & Lindenberger, U. (1992a). How to detect reasoning-remembering dependence (and how not to). *Developmental Review*, *12*(2), 187–198. https://doi.org/10.1016/0273-2297(92)90008-P

Chapman, M., & Lindenberger, U. (1992b). Transitivity judgments, memory for premises, and models of children's reasoning. *Developmental Review*, *12*(2), 124–163. https://doi.org/10.1016/0273-2297(92)90006-N

Ciranka, S., Linde-Domingo, J., Padezhki, I., Wicharz, C., Wu, C. M., & Spitzer, B. (2022). Asymmetric reinforcement learning facilitates human inference of transitive relations. *Nature Human Behaviour*, *6*(4), 555–564. https://doi.org/10.1038/s41562-021-01263-w

Deng, W., & Sloutsky, V. M. (2013). The role of linguistic labels in inductive generalization. *Journal of Experimental Child Psychology*, *114*(3), 432–455. https://doi.org/10.1016/j.jecp.2012.10.011

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2021). Evidence of hippocampal learning in human infants. *Current Biology*, *31*(15), 3358-3364.e4. https://doi.org/10.1016/j.cub.2021.04.072

Forest, T. A., Schlichting, M. L., Duncan, K. D., & Finn, A. S. (2023). Changes in statistical learning across development. *Nature Reviews Psychology*, *2*(4), 205–219. https://doi.org/10.1038/s44159-023-00157-0

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. https://doi.org/10.1006/anbe.2003.2174

Giron, A. P., Ciranka, S., Schulz, E., Van Den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, *7*(11), 1955–1967. https://doi.org/10.1038/s41562-023-01662-1

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. https://doi.org/10.1016/j.dcn.2015.08.009

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), Article 2. https://doi.org/10.1002/hipo.10189

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. JSTOR.

Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for Heteroskedasticity Resulting from Between-Group Differences in Multilevel Models. *Multivariate Behavioral Research*, 1–21. https://doi.org/10.1080/00273171.2022.2077290

Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P., & Terrace, H. S. (2015). Implicit Value Updating Explains Transitive Inference Performance: The Betasort Model. *PLOS Computational Biology*, *11*(9), e1004523. https://doi.org/10.1371/journal.pcbi.1004523

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' Memory for Spoken Words. *Science*, *277*(5334), 1984–1986. https://doi.org/10.1126/science.277.5334.1984

Kallio, K. D. (1982). Developmental change on a five-term transitive inference. *Journal of Experimental Child Psychology*, *33*(1), 142–164. https://doi.org/10.1016/0022-0965(82)90011-X

Karjack, S., Newcombe, N. S., & Ngo, C. T. (2025). The dependence of children's generalization on episodic memory varies with age and level of abstraction. *Nature Communications*, *16*(1), 8894. https://doi.org/10.1038/s41467-025-63934-w

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. *Proceedings of the National Academy of Sciences*, *114*(34), Article 34. https://doi.org/10.1073/pnas.1710654114

Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2018). Hippocampal Maturation Drives Memory from Generalization to Specificity. *Trends in Cognitive Sciences*, *22*(8), 676–686. https://doi.org/10.1016/j.tics.2018.05.004

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Lavenex, P., & Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioural Brain Research*, *254*, 8–21. https://doi.org/10.1016/j.bbr.2013.02.007

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.7.5) [R package]. https://CRAN.R-project.org/package=emmeans

Long, J. S., & Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, *54*(3), 217. https://doi.org/10.2307/2685594

McClelland, J. L., McNaughton, B. L., & 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. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775

Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A., & Preston, A. R. (2021). Memory Reactivation during Learning Simultaneously Promotes Dentate Gyrus/$CA_{2,3}$ Pattern Differentiation and CA_1 Memory Integration. *The Journal of Neuroscience*, *41*(4), 726–738. https://doi.org/10.1523/JNEUROSCI.0394-20.2020

Mou, Y., Province, J. M., & Luo, Y. (2014). Can infants make transitive inferences? *Cognitive Psychology*, *68*, 98–112. https://doi.org/10.1016/j.cogpsych.2013.11.003

Newcombe, N. S., Benear, S. L., Ngo, C. T., & Olson, I. R. (2024). Memory in Infancy and Childhood. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory, Two Volume Pack* (1st ed., pp. 1547–1575). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190917982.013.53

Ngo, C. T., Benear, S. L., Popal, H., Olson, I. R., & Newcombe, N. S. (2021). Contingency of semantic generalization on episodic specificity varies across development. *Current Biology*, *31*(12), Article 12. https://doi.org/10.1016/j.cub.2021.03.088

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. https://doi.org/10.1037/0033-295X.110.4.611

Nussenbaum, K., & Hartley, C. A. (2025). Reinforcement learning increasingly relates to memory specificity from childhood to adulthood. *Nature Communications*, *16*(1), 4074. https://doi.org/10.1038/s41467-025-59379-w

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345. https://doi.org/10.1037/0033-295X.108.2.311

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*(11), 803–809. https://doi.org/10.1038/s41562-017-0219-x

Piaget, J. (1928). *Judgement and Reasoning in the Child* (M. Gabain, Trans.; 1st ed.). Kegan Paul, Trench, Trübner & Co. Ltd. (Original work published 1924)

Piaget, J., Grize, J.-B., Szeminska, A., & Bang, V. (1977). *Epistemology and Psychology of Functions*. Springer Netherlands. https://doi.org/10.1007/978-94-010-9321-7 (Original work published 1968)

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. E. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus*, *14*(2), Article 2. https://doi.org/10.1002/hipo.20009

Pustejovsky, J. E. (2022). *_clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections_* (Version R package version 0.5.7) [R]. https://CRAN.R-project.org/package=clubSandwich

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for Representations of Perceptually Similar Natural Categories by 3-Month-Old and 4-Month-Old Infants. *Perception*, *22*(4), 463–475. https://doi.org/10.1068/p220463

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsaran, A. I., Schlichting, M. L., & Frankland, P. W. (2019). The ontogeny of memory persistence and specificity. *Developmental Cognitive Neuroscience*, *36*, 100591. https://doi.org/10.1016/j.dcn.2018.09.002

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*(4), e12593. https://doi.org/10.1111/desc.12593

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version R package version 2.2.5) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Richmond, J. L., & Pan, R. (2013). Thinking about the future early in life: The role of relational memory. *Journal of Experimental Child Psychology*, *114*(4), 510–521. https://doi.org/10.1016/j.jecp.2012.11.002

Riggins, T., 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. https://doi.org/10.1016/j.neuroimage.2018.03.009

RStudio Team. (2020). *RStudio: Integrated Development Environment for R. RStudio* (Version 1.3.1073) [Computer software]. RStudio, PBC. http://www.rstudio.com/

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlesinger, M., & McMurray, B. (2012). The past, present, and future of computational models of cognitive development. *Cognitive Development*, *27*(4), 326–348. https://doi.org/10.1016/j.cogdev.2012.07.002

Schlichting, M. L., Guarino, K. F., Roome, H. E., & Preston, A. R. (2022). Developmental differences in memory reactivation relate to encoding and inference in the human brain. *Nature Human Behaviour*, *6*(3), 415–428. https://doi.org/10.1038/s41562-021-01206-5

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), Article 1. https://doi.org/10.1162/jocn_a_01028

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*(1), 8151. https://doi.org/10.1038/ncomms9151

Schuck, N. W., Li, A. X., Wenke, D., Ay-Bryson, D. S., Loewe, A. T., Gaschler, R., & Shing, Y. L. (2022). Spontaneous discovery of novel task solutions in children. *PLOS ONE*, *17*(5), e0266253. https://doi.org/10.1371/journal.pone.0266253

Shohamy, D., & Wagner, A. D. (2008). Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron*, *60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*(6), 246–251. https://doi.org/10.1016/S1364-6613(03)00109-8

Sloutsky, V. M., & Fisher, A. V. (2011). The Development of Categorization. In *Psychology of Learning and Motivation* (Vol. 54, pp. 141–166). Elsevier. https://doi.org/10.1016/B978-0-12-385527-5.00005-X

Spriet, C., Abassi, E., Hochmann, J.-R., & Papeo, L. (2022). Visual object categorization in infancy. *Proceedings of the National Academy of Sciences*, *119*(8), e2105866119. https://doi.org/10.1073/pnas.2105866119

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *eLife*, *12*, e77185. https://doi.org/10.7554/eLife.77185

Taylor, J. E., Cortese, A., Barron, H. C., Pan, X., Sakagami, M., & Zeithamova, D. (2021). How do we generalize? *Neurons, Behavior, Data Analysis, and Theory*, *1*. https://doi.org/10.51628/001c.27687

Townsend, E. L., Richmond, J. L., Vogel-Farley, V. K., & Thomas, K. (2010). Medial temporal lobe memory in childhood: Developmental transitions. *Developmental Science*, *13*(5), 738–751. https://doi.org/10.1111/j.1467-7687.2009.00935.x

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*(2), 413–426. https://doi.org/10.3758/s13428-015-0593-0

Wu, C. M., Meder, B., & Schulz, E. (2024). Unifying Principles of Generalization: Past, Present, and Future. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-021524-110810

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software*, *11*(10). https://doi.org/10.18637/jss.v011.i10

Zeileis, A., Köll, S., & Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software*, *95*(1), 1–36. https://doi.org/10.18637/jss.v095.i01

Zeithamova, D., & Bowman, C. R. (2020). Generalization and the hippocampus: More than one story? *Neurobiology of Learning and Memory*, *175*, 107317. https://doi.org/10.1016/j.nlm.2020.107317

Zeithamova, D., & Preston, A. R. (2010). Flexible Memories: Differential Roles for Medial Temporal Lobe and Prefrontal Cortex in Cross-Episode Binding. *Journal of Neuroscience*, *30*(44), 14676–14684. https://doi.org/10.1523/JNEUROSCI.3250-10.2010

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience*, *6*, 70. https://doi.org/10.3389/fnhum.2012.00070

Zhang, X., Qiu, Y., Li, J., Jia, C., Liao, J., Chen, K., Qiu, L., Yuan, Z., & Huang, R. (2022). Neural correlates of transitive inference: An SDM meta-analysis on 32 fMRI studies. *NeuroImage*, *258*, 119354. https://doi.org/10.1016/j.neuroimage.2022.119354

7. Supplementary Materials

7.1. Supplementary Methods

7.1.1. Temporal Regularity Task

One additional 4-year-old did not attend the second session, resulting in 29 4-year-olds included in the analysis for this task. The design of this task was adapted from a previous study that investigated statistical learning in children (Schlichting et al., 2017). Stimuli consisted of 12 cartoon aliens grouped into four unique triplets that were held constant across participants. Children were told that they would see aliens boarding a spaceship one at a time to return to their home planet and that they should pay close attention to all the aliens boarding the ship. The encoding phase began with an empty spaceship interior shown for 5 seconds followed by each triplet being shown 24 times, resulting in 288 encoding trials. On each trial, an alien appeared in the middle of the screen on top of the spaceship background and remained there for 1 second before moving to the left side of the screen and disappearing (0.5 seconds). After the alien disappeared, the next trial immediately followed, such that the entrance and exit of the aliens appeared continuous. During the entire encoding phase, children heard ambient space sounds in the background, intended to keep children engaged during encoding. Unbeknownst to the participants, we manipulated the transitional probability between two successive aliens, such that the transitional probabilities were 1.0 and 0.33 for within- and across-triplets aliens, respectively (see Supplementary Figure 2A). No instructions about the triplet structure were given prior to encoding to ensure that the transitional probabilities are the only way of demarcating triplets. The order of the triplets was pseudorandomized such that neither the same triplet (ABC) nor the same pair of triplets (ABCDEF) would appear in immediate succession. The order of the encoding trials was held constant across participants.

Immediately after the encoding phase, children performed a 3AFC triplet completion task (see Supplementary Figure 2B). They were first told that some alien species are befriended and that these like to sleep in the same room on the spaceship. Importantly, they were then told that befriended alien species always boarded the spaceship right after one another. Participants' task was then to help sort the aliens into their correct rooms by selecting the alien that always appeared after the first two shown on the top of the screen. On each test trial, children saw two aliens from the same triplet appear after each other on the top of the

screen with a 1.5 second inter-stimulus interval followed by all three options appearing simultaneously at the bottom of the screen.[2] Children were asked to select the correct alien who belongs with the two probe aliens. The three options included the target item and two lures. The two lures were aliens from two other triplets who appeared in the second and third position of their respective triplets. Each triplet was tested six times with a different combination of lures, resulting in a total of 24 test trials. One randomized order of the test trials was created, with the constraint that the same triplet is never tested in succession and that the position of targets is counterbalanced. All participants received the same test trial order.

7.1.2. Categorization Task

All 4-year-olds were excluded for this task because only very few were able to complete the task in its entirety ($n = 5$). Further, two 6- and 8-year-olds were excluded due to not finishing the task, resulting in 25 children in each age group included in the analysis. The design for the CAT Task was adapted from previous studies in young adults (Bowman et al., 2020; Bowman & Zeithamova, 2018). In this task, children first learned to categorize fish exemplars into two species of fish – salt vs. freshwater fish – based on their features, with corrective feedback. In the test phase, they were then asked to categorize new exemplars of each species. Stimuli were cartoon fish images, each having 8 features: mouth, eyes, gills, back fin, side fin, under fin, scales, and tail fin (see Supplementary Figure S3A for an example). One prototype of each species was predetermined, such that the two prototypes differed on each of the 8 features. Based on these prototypes, all fish exemplars within each species were created such that the exemplars share all but 1 (1-off), 2 (2-off), or 3 (3-off) from their respective prototype (Supplementary Figure S3A). The task procedure consisted of a 5-block training phase and a test phase.

[2] An earlier version of the task included a static presentation during the test phase, meaning that children saw both cues and all three choice options simultaneously. We noticed that 4-year-olds were at chance, so we adapted the test phase to mimic the encoding phase more closely by showing the cue aliens sequentially. However, this did not influence accuracy and 4-year-olds still performed at chance level so that we pooled the samples together (see Supplementary Table 2).

During training, participants repeatedly categorized four different exemplars of each species, all of which were 2-off exemplars. Note that the trained exemplars differed on four features from the other trained exemplars of their species (for a schematic overview of the feature assignment for training, see Supplementary Figure S3B). Importantly, the prototypes did not appear during the training phase. On each trial, children were shown one fish exemplar on a computer screen and asked whether the fish on the screen lives in fresh or salt water. Following the response, the experimenter recorded the child's response and provided corrective feedback in an interval of 1.5 seconds. Only the correct responses were also accompanied by a cheering sound. The next trial proceeded after a 0.5 second fixation-cross ITI (Supplementary Figure S3C).

Each of the five training blocks consisted of 8 training fish, each presented six times, resulting in 48 trials per block and 240 training trials in total. Participants were given self-paced breaks between two blocks. The set of training exemplars and the trial order was held constant across participants. The order of presentation in each block was pseudo-randomized with the constraint that the same exemplar does not appear in succession, and that no more than three members of the same species would be shown in succession. The same peudo-random presentation order across all five training blocks was held constant for each participant. The training phase lasted approximately 45 minutes.

The test phase immediately followed the training phase and consisted of 68 trials (Supplementary Figure S3B). Children were told that they would continue to categorize members of the two species but without feedback. Important for measuring generalization was children's categorization accuracy on the untrained items, including both prototypes (each tested twice) and untrained exemplars of each species. Among the untrained exemplars, there were eight items for every distance (1-off, 2-off, and 3-off) from their respective prototype, each tested once. All 1-off items were the same across participants as there are only eight possible exemplars per species. However, 2- and 3-off exemplars were selected randomly for each participant. The 2-off items were selected such that four exemplars would always differ in four dimensions from one another (same structure as the training items, see Supplementary Figure S3B). The 3-off items were selected such that two exemplars would always differ in exactly 6 dimensions from each other. Additionally, all of the trained exemplars were tested twice. The order of test trials was randomized such that,

in addition to the same restrictions for the training phase, no more than three items of the same type (e.g. 3-off item) would appear in succession. Lastly, the experimenter asked the participant to verbally report on their general categorization strategy. Children were subsequently asked to assign each feature to the species that they thought the given feature was most indicative. A schematic depiction of the task is shown in Supplementary Figure S3C. To familiarize children with the stimuli, the experimenter showed participants one example of a fish image, completely distinct from the experimental materials, prior to the training phase and successively directed their attention to each of the 8 features. To increase children's engagement throughout the task, they were told that the more fish they categorize, the fuller their own digital aquarium becomes.

7.1.3. Task reliability and disattenuation of correlation coefficients

To check for reliability in our experimental tasks, we computed split-half reliabilities individually for every task, first for the entire sample and then for individual age groups. This was done differently for each task. In the AI task, inference trials were split into even and odd trials. In the TI task, inference trials were split such that each half contained the same amount of trials for each transitive pair, since each pair was tested on 4 different trials. One half contained the first and third trial, the other half the second and fourth trial. In the TR task, similar to the TI task, test trials were split such that each half contained the same amount of trials for each triplet, since each triplet was tested on 6 different trials. One half contained the first, third and fifth trial, the other half the second, fourth and sixth trial. In the CAT task, trials were split such that each half contained an even amount of prototype, training, 1-off, 2-off, and 3-off items per category. Prototype items were repeated twice, resulting in the first prototype trial being in the first half and the second trial being in the second half. All other items were distinct and tested only once, but could share their item type (for example training item or 2-off item) which were tested on 8 trials. Therefore, the first half contained the first, third, fifth, and seventh trial of a given item type, the second half the second, fourth, sixth, and eighth trial. We then took the mean accuracy for each participant for both halves and correlated these, once for the entire sample and then for each individual age group. In a last step, these correlations were adjusted using the Spearman-Brown formula for split-half reliabilities.

To calculate disattenuated correlations, the individual correlations were divided by the square root of the product of reliabilities ($R_{AB} = \frac{r_{AB}}{\sqrt{r_{AA} * r_{BB}}}$) using the psych::correct.cor function (Revelle, 2022). This disattenuated coefficient was calculated for the entire sample and for each individual age group separately, always taking the respective raw inter-task correlation.

7.2 Supplementary Results

To assess age-related differences in response patterns for each task, we compared the proportions of trials in each cell (D+I+, D+I–, D–I+, D–I–) individually between age groups.

In the AI task, comparisons revealed that older children consistently showed higher proportions in the D+I+ cell than younger children (all $p < .01$), whereas the opposite was observed in all other cells (except for a non-significant difference between 4- and 6-year-olds in the D+I– cell, $p = .133$; all other $p < .05$; see Figure 5B).

In the TI task, older children's proportions were again higher in the D+I+ cell (all $p < .001$) compared to younger children, whereas the D–I– cell showed the reverse pattern (all $p < .01$). For the off-diagonal cells, age differences were mostly non-significant, except that 4-year-olds had higher proportions than 8-year-olds in the D-I+ cell (D–I+; $p < .001$). The differences between 4- and 6-year-olds, as well as between 6- and 8-year-olds, approached significance in that same cell ($p = .053$ and $p = .065$, respectively), and 4-year-olds showed higher proportions than 8-year-olds in the D+I- cell (p = .017). No other differences in proportions reached significance (both p > .12).

7.3 Supplementary Tables

Supplementary Table 1

Results from Fisher's exact test on passing the second learning phase criterion based on the administered version of the first learning phase in the Transitive Inference Task.

Age (Years)	Odds Ratio	*p*	95% CI
4	1.16	1.00	[0.17, 9.49]
6	0.50	1.00	[0.04, 4.05]
8	0.00	1.00	[0.00, 77.91]

Note. Odds Ratio refers to the success on passing the second learning phase criterion. P-values adjusted for 3 tests according to Bonferroni-Holm method.

Supplementary Table 2

Results from independent t-test comparing accuracy between static and sequential test versions in the Temporal Regularity Task.

Age (years)	Static			Sequential			*t*	*df*	*p*	*d*
	n1	*M1*	*SD1*	*n2*	*M2*	*SD2*				
4.00	8	0.26	0.05	21	0.31	0.09	-1.57	27	.384	-0.70
6.00	11	0.55	0.23	16	0.51	0.23	0.41	25	1.00	0.15
8.00	9	0.62	0.24	18	0.63	0.20	-0.16	25	1.00	-0.06

Note. P-values adjusted for 3 tests according to Bonferroni-Holm method.
M = mean, n = sample size, SD = standard deviation, df = degrees of freedom.

Supplementary Table 3

Results from one-sided one-sample t-tests against chance for all tasks (.33 for temporal regularity and associative inference; .5 for transitive inference and categorization).

Task	Age (years)	*n*	*M*	*SD*	*t*	*df*	*p*	*d*	95% CI
Temporal Regularity	4	29	0.30	0.08	-2.23	28	.983	-0.41	[0.27, Inf]
Temporal Regularity	6	27	0.53	0.22	4.51	26	< .001***	0.87	[0.45, Inf]
Temporal Regularity	8	27	0.63	0.21	7.46	26	< .001***	1.43	[0.56, Inf]
Associative Inference	4	30	0.49	0.19	4.68	29	< .001***	0.85	[0.44, Inf]
Associative Inference	6	27	0.74	0.22	9.58	26	< .001***	1.84	[0.67, Inf]
Associative Inference	8	27	0.94	0.07	43.25	26	< .001***	8.32	[0.92, Inf]
Transitive Inference	4	28	0.62	0.20	3.13	27	.002**	0.59	[0.55, Inf]
Transitive Inference	6	27	0.83	0.18	9.69	26	< .001***	1.87	[0.77, Inf]
Transitive Inference	8	27	0.98	0.09	26.73	26	< .001***	5.14	[0.95, Inf]
Categorization	6	25	0.63	0.14	4.69	24	< .001***	0.94	[0.58, Inf]
Categorization	8	25	0.71	0.16	6.69	24	< .001***	1.34	[0.66, Inf]

Note. P-values adjusted for the number of tests per task according to Bonferroni-Holm method.
M = mean, n = sample size, SD = standard deviation, df = degrees of freedom, 95% CI entails upper and lower limit .

Supplementary Table 4

Results from follow-up contrasts between age groups from the linear models fit to the respective task.

Task	contrast	beta	*SE*	*t*	*df*	*p*	95% CI
Temporal Regularity	age8 - age6	0.10	0.06	1.71	80	.092	[-0.02, 0.22]
Temporal Regularity	age8 - age4	0.33	0.04	7.61	80	< .001***	[0.24, 0.42]
Temporal Regularity	age6 - age4	0.23	0.05	4.91	80	< .001***	[0.14, 0.32]
Associative Inference	age8 - age6	0.20	0.05	4.44	116	< .001***	[0.11, 0.29]
Associative Inference	age8 - age4	0.45	0.04	12.16	116	< .001***	[0.38, 0.52]
Associative Inference	age6 - age4	0.25	0.05	4.56	116	< .001***	[0.14, 0.36]
Transitive Inference	age8 - age6	0.14	0.04	3.71	104	< .001***	[0.07, 0.22]
Transitive Inference	age8 - age4	0.36	0.04	8.52	104	< .001***	[0.27, 0.44]
Transitive Inference	age6 - age4	0.21	0.05	4.18	104	< .001***	[0.11, 0.32]
Categorization	age8 - age6	0.07	0.04	1.74	85	.086	[-0.01, 0.16]

Note.
P-values adjusted for the number of tests per task according to Bonferroni-Holm method
SE = standard error, df = degrees of freedom, 95% CI entails upper and lower limit.

Supplementary Table 5

Inter-task correlations for each age group seperately.

Age (years)	Task 1	Task 2	*r*	*t*	*p*	Method	p.adj	BF	95% CI
4	Transitive Inference	Associative Inference	.18	0.93	.360	Pearson	1.00	0.57	[-0.21, 0.52]
6	Transitive Inference	Associative Inference	.33	1.74	.095	Pearson	1.00	1.70	[-0.06, 0.63]
8	Transitive Inference	Associative Inference	.33	1.77	.089	Pearson	1.00	1.80	[-0.05, 0.63]
4	Transitive Inference	Temporal Regularity	-.16	-0.80	.432	Pearson	1.00	0.14	[-0.51, 0.24]
6	Transitive Inference	Temporal Regularity	.36	1.94	.063	Pearson	0.88	2.37	[-0.02, 0.65]
8	Transitive Inference	Temporal Regularity	.01	0.03	.977	Pearson	1.00	0.24	[-0.38, 0.38]
6	Transitive Inference	Categorization	.18	0.94	.358	Pearson	1.00	0.58	[-0.21, 0.53]
8	Transitive Inference	Categorization	.03	0.16	.878	Pearson	1.00	0.28	[-0.36, 0.41]
4	Associative Inference	Temporal Regularity	-.11	-0.60	.555	Pearson	1.00	0.15	[-0.46, 0.26]
6	Associative Inference	Temporal Regularity	.42	2.32	.029*	Pearson	0.43	4.57	[0.05, 0.69]
8	Associative Inference	Temporal Regularity	.20	1.03	.313	Pearson	1.00	0.65	[-0.19, 0.54]
6	Associative Inference	Categorization	-.00	-0.02	.985	Pearson	1.00	0.24	[-0.38, 0.38]
8	Associative Inference	Categorization	.14	0.67	.508	Pearson	1.00	0.44	[-0.27, 0.50]
6	Temporal Regularity	Categorization	.00	0.00	.997	Pearson	1.00	0.24	[-0.38, 0.38]
8	Temporal Regularity	Categorization	.07	0.32	.753	Pearson	1.00	0.32	[-0.33, 0.44]

Note. p.adj reflects adjusted p-values according to Bonferroni-Holm Method. No correlations shown between the categorization and the other tasks for 4-year-olds due to unsufficient amount of data in the categorization task. BF reflects the Bayes Factor in favor of the alternative hypothesis (correlation is greater than 0).

Supplementary Table 6

Results from follow-up contrasts between age groups for the respective test types in the Associative Inference Task.

contrast	pair type	beta	*SE*	*t*	*df*	*p*	95% CI
age8 - age6	direct	0.10	0.03	3.32	116	.001**	[0.04, 0.17]
age8 - age4	direct	0.36	0.04	8.16	116	< .001***	[0.27, 0.45]
age6 - age4	direct	0.26	0.05	4.82	116	< .001***	[0.15, 0.37]
age8 - age6	inference	0.20	0.05	4.44	116	< .001***	[0.11, 0.29]
age8 - age4	inference	0.45	0.04	12.16	116	< .001***	[0.38, 0.52]
age6 - age4	inference	0.25	0.05	4.56	116	< .001***	[0.14, 0.36]

Note. Results are from the average across learning blocks. P-values are adjusted for 6 tests according to Bonferroni-Holm method.
DM = direct memory (last block), AI = associative inference, SE = standard error, df = degrees of freedom, 95% CI entails upper and lower limit.

Supplementary Table 7

Results from follow-up contrasts between age groups at test in the Transitive Inference Task.

contrast	beta	*SE*	*t*	*df*	*p*	95% CI
age8 - age6	0.15	0.04	3.84	79	< .001***	[0.07, 0.23]
age8 - age4	0.34	0.04	8.88	79	< .001***	[0.27, 0.42]
age6 - age4	0.19	0.05	3.97	79	< .001***	[0.09, 0.28]

Note. Results are averaged across pair types. P-values are adjusted for 3 tests according to Bonferroni-Holm method.
SE = standard error, df = degrees of freedom, 95% CI entails upper and lower limit.

Supplementary Table 8

Results from follow-up contrasts of odds of inference success between age groups when pair memory was correct or incorrect in the Associative Inference Task.

contrast	pair memory	odds ratio	*SE*	*z*	*p*	95% CI
age8 / age6	correct	6.43	2.38	5.4	< .001***	[3.12, 13.28]
age8 / age4	correct	21.90	8.23	8.22	< .001***	[10.49, 45.74]
age6 / age4	correct	3.40	1.00	4.16	< .001***	[1.91, 6.06]
age8 / age6	incorrect	1.20	0.72	.30	.767	[0.37, 3.91]
age8 / age4	incorrect	2.28	1.29	1.45	.293	[0.75, 6.93]
age6 / age4	incorrect	1.90	0.67	1.83	.202	[0.96, 3.80]

Note. Contrasts on the odds scale are given by dividing the respective odds, however tests are performed on the log odds ratio scale. P-values are adjusted for 6 tests according to Bonferroni-Holm method.
SE = standard error, 95% CI entails upper and lower limit (both given on the odds scale).

Supplementary Table 9

Results from follow-up contrasts on increase of odds of inference success when pair memory is correct vs. incorrect in the Associative Inference Task.

Age (years)	contrast	odds ratio	*SE*	*z*	*p*	95% CI
8	pair memory: correct / incorrect	19.00	11.02	5.8	< .001***	[6.10, 59.23]
6	pair memory: correct / incorrect	3.53	1.07	4.17	< .001***	[1.95, 6.39]
4	pair memory: correct / incorrect	1.98	0.45	3.2	.008**	[1.27, 3.08]

Note. Contrasts on the odds scale are given by dividing the respective odds, however tests are performed on the log odds ratio scale. P-values are adjusted for 6 tests according to Bonferroni-Holm method (3 tests not shown, see main text for results).
SE = standard error, 95% CI entails upper and lower limit (both given on the odds scale).

Supplementary Table 10

Split-half reliability for each task for all age groups combined as well as individually by age.

Task	all	4-yo	6-yo	8-yo
Temporal Regularity	0.88	-0.32	0.91	0.90
Associative Inference	0.89	0.64	0.94	0.27
Transitive Inference	0.92	0.78	0.93	0.98
Categorization	0.88		0.79	0.95

Note. No reliability for the Categorization task for 4-year-olds due to unsufficient amount of data. Negative reliability for 4-year-olds in the Temporal Regularity Task is non-interpretable given their performance was at chance level.

7.4 Supplementary Figures

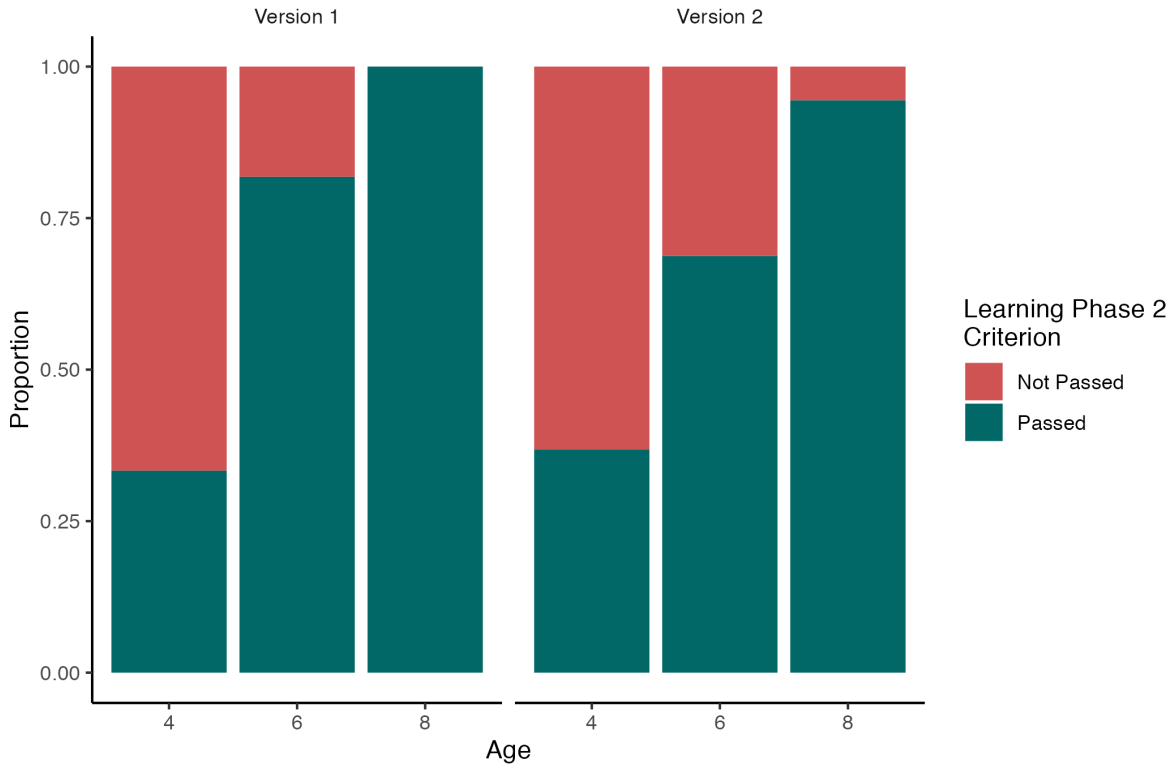


Figure S1. Proportion of children passing the criterion of the second learning phase of the Transitive Inference Task, separated by the version of the first learning phase. Version 1: $n_4 = 9, n_6 = 11, n_8 = 9$; Version2: $n_4 = 19, n_6 = 16, n_8 = 18$.

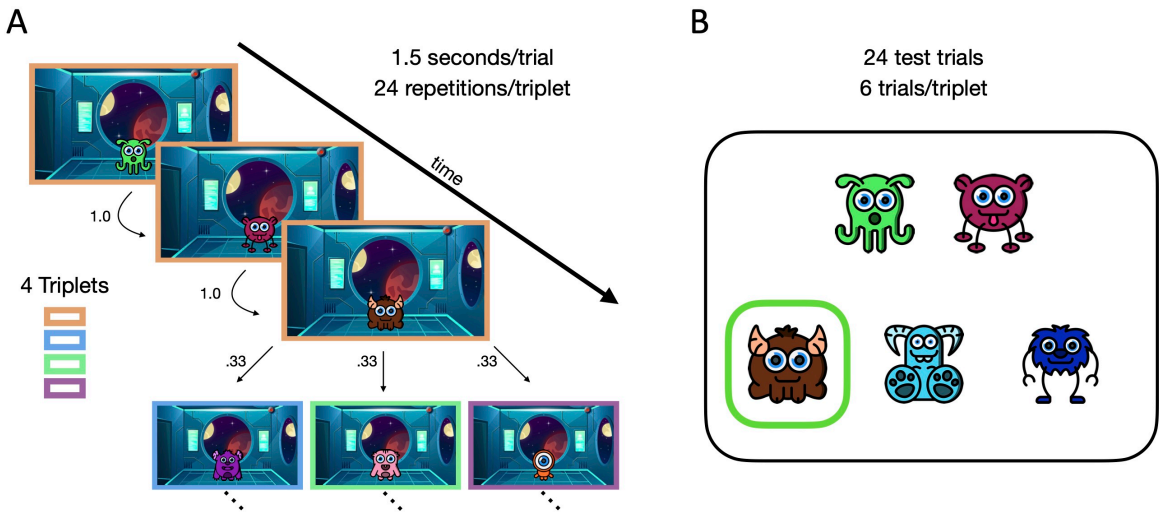
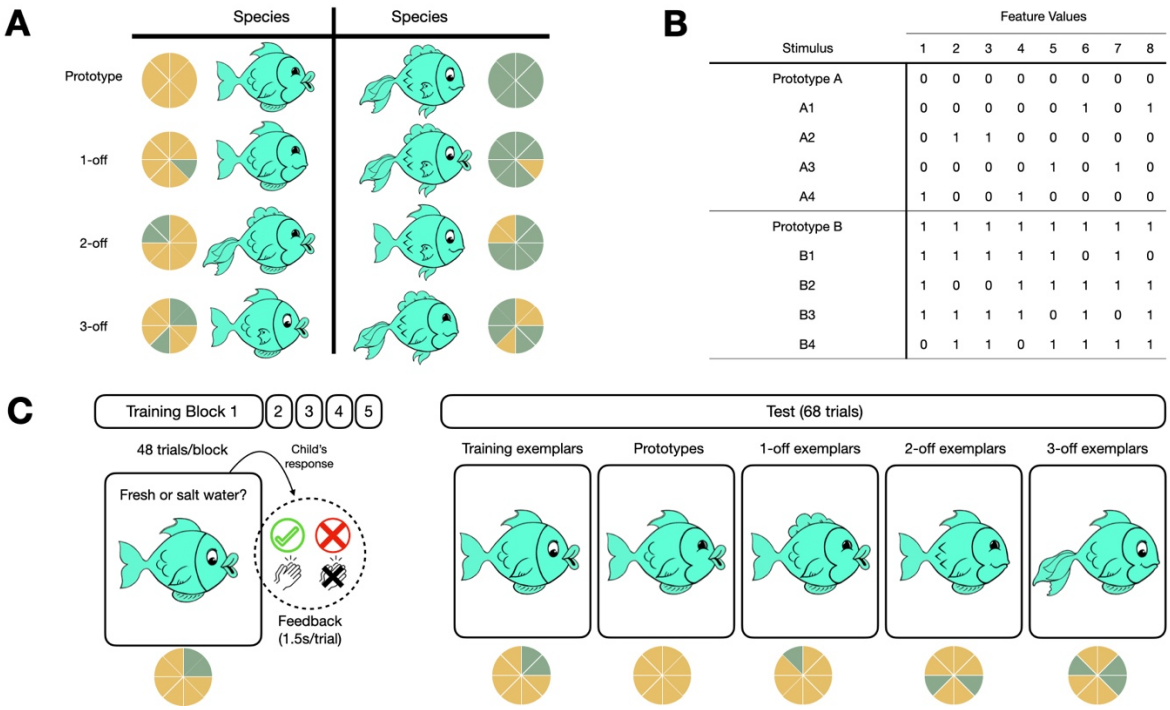


Figure S2. A schematic depiction of the Temporal Regularity Task. ***(A)*** During encoding, children viewed a continuous stream of aliens entering a spaceship one at a time. Unbeknownst to them, the sequence contained four triplets (indicated with different border colors), each presented 24 times. Within each triplet, the transitional probability was 1.0, while between triplets, it was .33. ***(B)*** At test, children first saw two aliens at the top of the screen appear sequentially, followed by the three aliens at the bottom. They were instructed to select the alien that completed the triplet (green rectangle denotes correct option, not shown to participants).



Stimulus	Feature Values							
	1	2	3	4	5	6	7	8
Prototype A	0	0	0	0	0	0	0	0
A1	0	0	0	0	0	1	0	1
A2	0	1	1	0	0	0	0	0
A3	0	0	0	0	1	0	1	0
A4	1	0	0	1	0	0	0	0
Prototype B	1	1	1	1	1	1	1	1
B1	1	1	1	1	1	0	1	0
B2	1	0	0	1	1	1	1	1
B3	1	1	1	1	0	1	0	1
B4	0	1	1	0	1	1	1	1

Figure S3. Categorization Task stimulus design. ***(A)*** Each species prototype shared no features with the other prototype. Oher exemplar types differed from their respective prototype by 1, 2, or 3 features, making them increasingly similar to the opposite prototype and less similar to their own. ***(B)*** Four training exemplars per species were selected to differ from their prototype by two features and from each other by four features. The training exemplars for one species were the reverse of those for the other species. ***(C)*** A schematic depiction of the Categorization Task. (A) During training, children repeatedly categorized exemplars that belonged to either of the two species. On each trial, they saw one fish on the screen and, after indicating the species to which that fish belonged, they received corrective verbal feedback and a clapping sound if correct and no sound if incorrect. Each of the eight training exemplars were shown six times per block resulting in 48 trials per block. (B) At test, children categorized both training and new exemplars, including both species' prototypes and 1-, 2-, 3-off exemplars. There was no corrective feedback at test.