How the hypothesis space is represented. Evidence from young children's active search and predictions in a multiple-cue inference task

Angela Jones

Max Planck Institute for Human Development, Berlin, Germany

Douglas B. Markant

Department of Psychological Science, University of North Carolina, Charlotte, NC, USA

Thorsten Pachur

Max Planck Institute for Human Development, Berlin, Germany

Alison Gopnik

Department of Psychology, University of California, Berkeley, CA, USA

Azzurra Ruggeri

Max Planck Institute for Human Development, Berlin, Germany

School of Education, Technical University of Munich, Germany

Author Note

Correspondence concerning this article should be addressed to Azzurra Ruggeri, MPRG iSearch | Information Search, Ecological and Active Learning research with Children, Max Planck Institute for Human Development, Berlin, Germany, Lentzeallee 94, Berlin, Germany, Phone: + 49 30 82 406 268. E-mail: ruggeri@mpib-berlin.mpg.de

The authors have no conflict of interest to declare. We would like to thank Costanza De Simone, Theresa Niemann, Sarah Vega, Ari Pascarella and Patricia Sueya Li for assistance in recruiting and data collection. We would also like to thank Federico Meini for the software development and technical support. We thank all the children from the Laurence Hall of Science in Berkeley, the Madera Primary School in El Cerrito, and the Natural History Museum and Labyrinth Museum in Berlin for providing research space.

Abstract

To successfully navigate in an uncertain world, one has to learn the relationship between cues (e.g., symptoms) and an outcome (e.g., disease). During this learning, it is sometimes possible to actively manipulate the cue values, allowing one to test hypotheses about this relationship directly. Across two studies, we investigated how 5- to 7-year-olds select cue configurations when learning cue-outcome relationships and how they are guided by representations of the hypothesis space regarding these relationships. In our task, children selected which monster pairs to see running in a race, allowing them to learn how two cues (the color and shape of monsters) predicted the relative speed of the monsters; subsequently, they made predictions about the speed of new monsters. Using computational modeling, we compared several models in their ability to capture children's responses. We found that young children's search was most consistent with a model that assumed reliance on a hypothesis space represented in terms of the relative speed of individual monsters. However, when memory aids were provided during search, 7-year-olds were best described by a model that assumes reliance on a more efficient, high-level representation that organizes the hypothesis space based on abstracted cue-outcome relationships. Our results highlight the guiding role of hypothesis-space representations for search during learning, suggest that young children already spontaneously abstract hypothesis-space representations, and provide the first evidence for a shift between search and test in terms of the hypothesis-space structures on which children rely when navigating in an uncertain world.

Keywords: multiple-cue inference, active learning, hypothesis-space, representation

How the hypothesis space is represented. Evidence from young children's active search and predictions in a multiple-cue inference task

Introduction

"Stick out your tongue and say *aaaah.* Have you been feeling tired or stressed lately? Any headaches?" Doctors usually ask many such targeted questions about patients' symptoms and run additional tests to gather the information they need to make a diagnosis. Knowing which questions to ask or tests to run and being able to interpret symptoms, answers, and results to identify what illness a patient is most likely to suffer from is a highly specialized skill. However, it is, in principle, an ability that we all rely on in much more mundane situations. Imagine you are attending your first school's sack race. Because you are the youngest racer, you get to choose whether to race against Caren or Jason. You have seen a couple of races already: Mary, a very tall girl, won against Bob, and Mike, also pretty tall, won against Lucy. So, you have learned that Mary is faster than Bob, and that Mike is faster than Lucy. And that might be all you have learned. In this case, you still would have no idea which opponent you are more likely to win against, and observing 10 or 20 more races would not help much–unless you could see Caren and Jason competing against each other, of course. However, you would have gathered more information from those two races you observed if you had tried to understand what makes people faster at this game. Why did Mary and Mike win? Maybe because they are so tall, or because they had a lighter sack? Organizing your hypothesis in this way, that is, identifying salient characteristics which may be related to the outcome, has implications for how to search. In this example, the most information can be gathered by watching a tall kid with a heavy sack race against a small kid with a thinner sack. This information might help you choosing between Caren, who is way shorter than you but has an ultra professional aerodynamic sack, or Jason, who's much taller than you but has a thick black leather sack. This example illustrates that the way we represent the set of hypotheses under consideration, for example by organizing them in groups according to more abstract shared features or by only considering the individual hypotheses, drives our information search and impacts our ability to predict the outcome. In turn, search choices and predictions may indicate how the hypothesis space might have been represented.

In this paper, we explored how 5- to 7-year-olds search for information that would enable them to make accurate predictions about an outcome (the winner of a race) based on multiple cues (e.g., height and weight of the sack). In particular, we implemented for the first time, to our knowledge, an *active learning* paradigm in which children could search for information by selecting the cue configurations for which they wanted to observe the outcome. Moreover, using computational modeling, we investigated what children's behavior revealed about their representation of the hypothesis space during both exploration and prediction.

The Development of Information Search Across Childhood

The ability to search for information emerges at a very early age (Cook et al., 2011; Legare et al., 2013; McCormack et al., 2016; Ruggeri et al., 2019). Preschoolers are more likely to explore when they are presented with confounded evidence—that is, when they are uncertain about the causal mechanism at work (Cook et al., 2011; Schulz and Bonawitz, 2007)—or when they face evidence that violates their prior beliefs (e.g., Bonawitz et al., 2012), and infants already prefer to explore surprising events (Stahl and Feigenson, 2015; Sim and Xu, 2017). However, the ability to search for information *efficiently* is subject to considerable developmental changes. For example, despite being able to select the most informative of two given questions already at age 5 (Ruggeri et al., 2017), children do not start consistently implementing effective question-asking strategies until age 10 (Mosher and Hornsby, 1966; Herwig, 1982; Ruggeri and Lombrozo, 2015; Ruggeri et al., 2016; Ruggeri and Feufel, 2015). This is also supported by process-tracing studies that examined children's information search using information boards, where participants have to look up information about different cues for a set of options (e.g., for a set of bikes, the price, number of gears, and color) to make a decision (e.g., which bike to buy). These studies show consistent developmental improvements in search efficiency between the ages of 7 and 14 years, with younger children searching more exhaustively and in a less systematic manner than older children (Davidson, 1991a,b; Gregan-Paxton and Roeder John, 1995, 1997; Howse et al., 2003).

It is currently unclear, however, what drives the observed developmental differences in search efficiency. Previous studies suggest that older children are more efficient than younger children because they are more systematic—in the sense that they are better able to focus on the dimensions that are more important for making the decision (Davidson, 1991b; Betsch et al., 2014; Ruggeri and Katsikopoulos, 2013; Mata et al., 2011). Recent research also suggests that one crucial source of developmental change in information search efficiency lies in children's stopping rules: Children are considerably more likely than adults to continue their search for information beyond the point at which a decision can be made (e.g., when a single hypothesis remains; Ruggeri et al., 2016). Another proposed explanation for young children's limited efficiency in information search is that they have difficulty going beyond the object level, that is, they fail to spontaneously identify, represent, and therefore reason with more abstract task structures. Consistent with this idea, Ruggeri and Feufel (2015) found that scaffolding more abstract representations of the hypotheses in the 20-questions game helped 7- and 10-year-olds ask more informative questions. In this study, 7- and 10-year-old children as well as adults were presented with 20 cards, each presenting a word label (e.g., "dog" or "sheep"). Participants were randomly assigned to one of two experimental conditions that differed in the level of the abstraction of the label: a basic-level condition (e.g., "dog") or a subordinate-level condition (e.g., "Dalmatian"). Participants were more likely to ask effective questions, ones that targeted the objects' categories (e.g., is it a pet?) rather than individual objects, in the former condition than in the latter. This suggests that providing more abstract labels facilitated a shift away from reasoning based on individual objects when generating questions. The study showed that the ability to generate more abstract features for given objects (e.g., "a dog is a mammal") also improves between age 7 and 10 (see also Herwig, 1982).

These results suggest that developmental differences in the ability to represent task-relevant information, and in particular the hypotheses under consideration, in an abstract way may drive developmental differences in *information search*. Let's go back to the racing example presented above, with a slight twist, that bring us closer to our experimental design: Suppose you have to order four monsters on a podium, from the fastest to the slowest runner. If you could only think about the individual monsters, your hypothesis space would consist of all their possible permutations, that is, all the possible rank orders on the podium. We refer to this as a *permutation-based* hypothesis space. Even if there were only 4 monsters, the hypothesis space would include 24 hypotheses about all the possible rankings. With this representation of the problem, you would have to observe at least 6 carefully selected races to narrow down the hypothesis space to the correct podium rank. Moreover, you would have to memorize and integrate the results of all of the observed races—a rather long, memory-intensive and therefore potentially error-prone process.

However, if you were able to abstract some relevant features, for example their color (blue or green) and shape (square or circle), you can organize the potential hypotheses by abstracting the relationship between these two available cues and the outcome (cf. Einhorn et al., 1979; Juslin et al., 2003a,b). Specifically, this representation of the hypothesis space, that we refer to as *cue–abstraction*, contains only 8 hypotheses (see Table 1). In this case, finding the correct podium order would take only three information search steps: Find out which color and shape makes the monsters faster (the *direction* of each cue), and find out which cue is more important for determining a monster's speed (the *cue order*). Importantly, the way the hypothesis space is organized can affect the way information is *generalized* to situations that have never been encountered before, as only cue abstraction allows to learn relationships that can be reliably extended to novel sets of objects (cf.

Hypothesis	**Color Direction**	**Shape Direction**	**Cue Order**
1	B>G	S>C	CO>SH
2	B>G	S>C	SH>CO
3	B>G	C>S	CO>SH
4	B>G	C>S	SH>CO
5	G>B	S>C	CO>SH
6	G>B	S>C	SH>CO
7	G>B	C>S	CO>SH
8	G>B	C>S	SH>CO

Table 1

Structure of the cue-abstraction model.

B = Blue; G = Green; S = Square; C = Circle; CO = Color; SH = Shape

Pachur and Olsson, 2012; Pachur and Trippas, 2019).

In a study that compared children and young adults in the context of a multiple-cue inference task, von Helversen et al. (2010) found that whereas the majority of adults were best described by a cue-abstraction strategy (see also Juslin et al., 2003a,b; Trippas and Pachur, 2019), 9-11 year old children were more likely to rely on a similarity-based processes for prediction, that is, they made inferences based on how much a new object resembles previously encountered ones. In addition, reliance on cue abstraction was associated with better performance for adults but not for children, suggesting that even those children who approached the task with the appropriate strategy struggled to implement it correctly (cf. Pachur and Olsson, 2012). Cue abstraction may be especially sensitive to developmental shifts in attentional control and working memory, as it requires one to maintain and reason about multiple cue-outcome relations. Thus, even older children may rely on different representations depending on the demands of the task.

Overview of the Studies

In this article, we report two experimental studies in which we investigated the early emergence of the ability to actively learn cue-outcome relationships in order to make accurate predictions about new objects. We implemented an active search paradigm in which 5- to 7-year-old children were presented with four monsters (see top left panel in 1) and tasked to find out which kinds of monsters were faster. In the search phase, they could select which monster pairs to see running in a race in order to learn how two cues (color and shape) predict the monsters' relative speed. At each step, children should select to observe the race they think would be the most informative. In other words, children should pick the pair with the highest *expected information gain* (EIG) at each step of the search. EIG quantifies the usefulness of a query based on how much new information it is expected to provide, that is, how many hypotheses will be ruled out after seeing the outcome of that query, and with what likelihood. Formally, EIG expresses the reduction of entropy, or the uncertainty as to which hypothesis is correct, upon making a query and observing its outcome (Shannon, 1948). Crucially, EIG depends on how the hypothesis space is structured. Indeed, as illustrated with the previous example, different representations of the hypothesis space suggest that different queries are maximally informative. In the test phase of the task, children were asked to predict the winner of races between novel monsters and to construct a “podium" in which they ranked the four monsters seen during learning. The way the hypothesis space is represented also affects the ability to identify the correct ranking from the evidence gathered during search, as well as to generalize what was learned to unfamiliar monsters.

In Study 1, we evaluated whether and to what extent 5- and 6-year-olds' search selections and predictions indicated that they relied on a *cue-abstraction* representation of the hypothesis space, in which hypotheses are organized by abstracting the relationships between cues and the outcome, in this case encoding both the cue direction and order (i.e., which color is faster, which shape is faster, and whether color or shape is more important for predicting a monster's speed; see Table 1). This representation of the hypothesis space requires the ability to abstract the relationship between cues and the outcome from the pairs of monsters encountered; this might be challenging for younger children, who have been shown to struggle with such abstractions (Herwig, 1982; Ruggeri and Feufel, 2015). Abstracting the cue order (i.e., whether color or shape was more important for determining relative speed), a second-order cue, may be especially difficult for them. We therefore generally expected that 5-year-olds would be less likely than 6-year-olds to make the most informative selections and accurate generalizations as predicted from a cue-abstraction representation of the hypothesis space. In Study 2, we extended our developmental analysis to include a sample of 7-year-olds, and additionally considered the *permutation-based* representation of the hypothesis space, in which hypotheses correspond to a specific order of the four monsters (see Table **??**). This alternative representation is similar to what is referred to as exemplar processing (Juslin et al., 2003a,b; Pachur and Trippas, 2019). In addition, we used computational modeling of children's search selections and predictions in order to identify which representation of the hypothesis space they were more likely to have relied on in each phase of the task.

Study 1

Method

Participants. Participants were 51 5- to 6-year-olds (29 female; $M = 72.40$ months, $SD = 6.64$ months), recruited and tested at museums and primary schools in the East Bay of the San Francisco area. An additional 17 participants were excluded due to equipment malfunction ($n = 14$), withdrawal of consent ($n = 1$), lack of fluency in English ($n = 1$), or failure to complete the study ($n = 1$). Ethical approval was obtained by the Institutional Review Board (IRB) of [blind for review] (protocol: [blind for review]), and parents gave informed consent for their children's participation before the experiment. The children were native [blind for review]or fluent in [blind for review], were predominantly white and came from various social classes.

Design and procedure. The task was presented to the children on a laptop computer and consisted of an active learning and a test phase. The task lasted approximately 10 min.

Active learning phase. Children were presented with four monsters: a green square, a green round, a blue square, and a blue round monster (Figure 1). The speed of each monster was determined by its features (i.e., color and shape): The blue monsters were faster than the green ones, the square ones faster than the round ones, and color was the more important cue for predicting monsters' relative speed. The monster order from the fastest to the slowest was therefore: blue square, blue round, green square, green round.

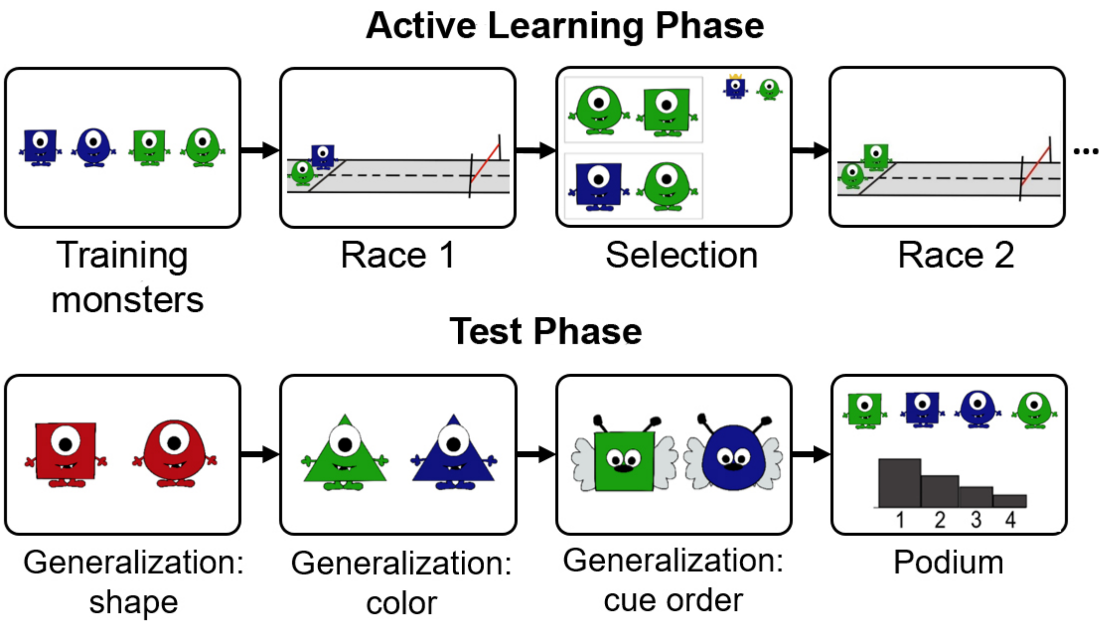


Figure 1. Experimental procedure in Study 1.

Participants were first presented with a short video clip showing the blue square monster winning a race against the green round monster. Verbal instructions prompted participants to pay particular attention to the monsters' color and shape. The active learning phase was carried out in three blocks. In each block, children were presented with

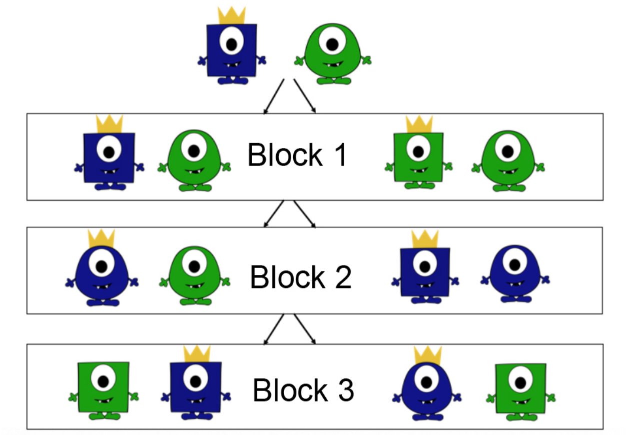


Figure 2. Pairs of monsters presented in the three blocks of the active learning phase of Study 1. The winning monster for each pair is marked with a crown.

two monster pairs on the screen (Figure 1). Children selected which of the two monster pairs they wanted to see racing to find out which kinds of monsters were faster (Figures 1, 2). They then saw a video of the selected monsters racing and observed the outcome of each race. Monster pairs that participants had already seen racing appeared on the right side of the screen, with the winner marked with a crown (Figure 1). These memory aids were provided during the entire learning phase. The learning blocks were designed such that according to the cue-abstraction representation of the hypothesis space, one monster pair provided new information about the cue-outcome relationships, whereas the other provided no new information according (EIG = 0; e.g., selecting a blue square monster and a blue round monster after having seen a green square and green round monster racing, which only confirms that square monsters are faster, without providing any new information).

If, in any block, a child selected the monster pair that did not provide any information (n.b., EIG = 0 according to the cue-abstraction representation of the hypothesis space), the same block was presented again. This ensured that childen had collected all the relevant information they would need, according to the cue-abstraction representation of the hypothesis space, to make the predictions presented in the test phase. Because of this, children who selected uninformative pairs were eventually presented with more learning blocks in total, whereas children who always selected the most informative pair were only presented with three learning blocks.

Test phase. The test phase consisted of a generalization task and a podium task, designed to assess how well children had learned the cue-outcome relationships.

In the *generalization* task, children were shown pairs of unfamiliar monsters and were asked to predict which monster would win the race (Figure 1). Memory aids from the learning phase remained visible on the right side of the screen. The first two trials in the generalization task presented monsters with either known colors but new shapes (blue and green triangles) or known shapes and a new color (a red circle and red square). These two trials aimed to test whether children had learned the direction of the color and shape cues, that is, which color and which shape indicated a higher speed. The order of these trials was counterbalanced across participants. The third trial presented a pair in which one monster had the faster shape but the slower color (a green square) and the other had the slower shape but the faster color (a blue round monster; see Figure 1). This trial tested whether children had learned the order of the cues, that is, that color was more important than shape for predicting the winner of a race.

In the *podium* task, children were presented with the four monsters from the learning phase and were instructed to rank them from the fastest to the slowest by positioning them on a four-step podium. (Figure 1).

Results

Active learning phase. First of all, we found very similar learning patterns across age groups, with comparable mean number of learning blocks ($M_{5years} = 4.50$, $SD = 0.81$; $M_{6years} = 4.16$, $SD = 0.85$; $p = .15$) and cue-abstraction-EIG (i.e., EIG calculated according to the cue-abstraction representation of the hypothesis space; $M_{5years} = 0.58$,

$SD = 0.22$; $M_{6years} = 0.68$, $SD = 0.24$; $p = .92$). Second, we simulated 51 *cue-abstraction-optimal* participants, who always selected the monster pair that was informative according to the cue-abstraction representation of the hypothesis space, and 51 participants who selected one of the two monster pairs at random, but never selected the uninformative pair twice, as none of the children in our sample did so. We then compared our sample to the simulated participants, and found that children's selections were less informative than those generated by cue-abstraction-optimal participants ($t(50) = -11.210$, $p < .001$). This was true for both 5- and 6-year-olds separately ($p < .001$). We also found that their selections were more informative than those of the simulated random participants ($t(50) = 3.30$, $p = .002$). However, this difference was mostly driven by 6-year-olds ($t(24) = 2.86$, $p = .009$), as the cue-abstraction-EIG of 5-year-olds did not statistically differ from random ($p = .094$).

Test phase. If children had a cue-abstraction representation of the hypothesis space, at the end of the active learning phase they should have known that square monsters are faster than round monsters, that blue monsters are faster than green ones, and that color is more important than shape to determine speed. Note that only information encoded at this level of abstraction is generalizable to new monsters. We therefore calculated the percentage of generalization trials in which children selected the correct monster as suggested by the cue-abstraction representation, and therefore encoding, of the hypothesis space (i.e., the square red monster over the round red monster, the blue monster over the green monster, and the round blue monster over the square green monster).

Overall, the mean accuracy in the generalization task slightly increased with age ($\chi 2(3) = 7.61$, $p = .055$). In particular, note that 5-year-olds were at chance in the generalization trial assessing their learning of cue-order (see Figure 3). However, mean accuracy was overall significantly above chance for both age groups ($M_{5years} = 67.97\%$, $SD = 30.52$, exact binomial: $p < .001$; $M_{6years} = 78.67\%$, $SD = 25.24$, exact binomial: $p < .001$). A linear regression with continuous age and children's cue-abstraction-EIG as predictors confirmed the age effect ($p = .012$), but showed no effect of the selections during search (cue-abstraction-EIG) on generalization ($p = .729$).

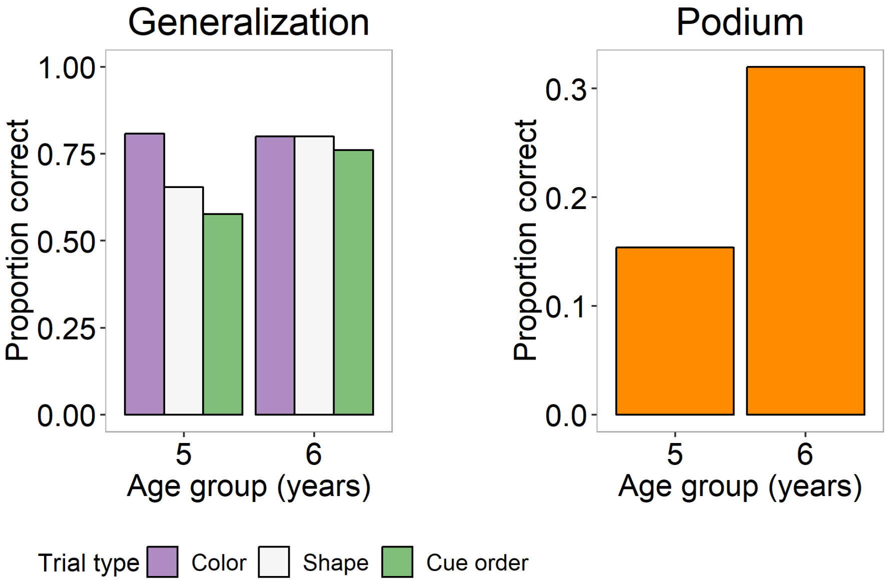


Figure 3. Test phase performance in Study 1, scored according to the predictions of a cue-abstraction representation of the hypothesis space. Left: Proportion of participants who chose the correct monster on each generalization trial, by age group. Right: Proportion of children who ranked all four monsters correctly on the podium test. With 24 possible rankings, chance performance is equal to .042.

If children had a cue-abstraction representation of the hypothesis space, at the end of the active learning phase they would have narrowed down the hypotheses to one, determining a unique rank order of the monsters. Podium accuracy was therefore scored according to whether children assembled the correct podium compatible with a cue-abstraction representation (1=correct, 0=incorrect). Accuracy in the podium task was also above chance for both age groups (Figure 3), with 15.38% of the 5-year-olds (two-tailed binomial test: $p = .02$) and 32% of the 6-year-olds (two-tailed binomial test: $p < .001$) assembling the correct podium. A binary logistic regression with continuous age and children's cue-abstraction-EIG as predictors suggested a small age effect ($OR = 3.57$,

$[.936, 13.62]$), but showed no effect of the learning pattern (cue-abstraction-EIG) on generalization ($OR = 2.87$, $[.162, 50.97]$).

Discussion of Study 1

The results of Study 1 suggest that there might be a difference in how the hypothesis space is represented during information search versus prediction. In particular, the analysis of children's selections provides some evidence that 6-, but not 5-year-olds have organized their hypothesis space by abstracting the cue-outcome relationship. Indeed, 5-year-olds' selections did not differ from those of random participants. However, our results of the generalization and podium tasks showed that, at least to a certain extent, both age groups differ from chance, independently of being more or less driven by a cue-abstraction representation of the hypothesis space during search, with a slight age improvement. This suggests that all children organized the information collected during the active learning phase according to a cue-abstraction representation of the hypothesis space. In other words, we could speculate that even 5-year-olds, who did not seem to rely on a cue-abstraction representation of the hypothesis space when making selections, nevertheless reasoned about the information collected by abstracting the cue-outcome relationships at test, possibly because otherwise they would have nothing to inform their predictions. Note that this reorganization was facilitated by the memory aids, presenting all the information observed during the learning phase.

Study 2

Because in Study 1 we designed the active learning phase to encourage a cue-abstraction representation of the hypothesis space, our results may not reflect how young children spontaneously approach a multiple-cue inference task. We address this limitation in Study 2, by leaving children free to create their own monster pairs and observe as many races as they wanted. In particular, this allows us to analyze whether children's selections and predictions are better described and compatible with a cue-abstraction or a permutation-based representation of the hypothesis space. The two different representation of the hypothesis space make different predictions as to which monster pairs are most informative to observe and therefore should be selected, and also make different predictions for the test phase, because the collected information is encoded differently and thus narrows down the hypothesis space in different ways. For example, knowing that the green square monster is faster than the green round monster will leave open 4 out of 8 hypotheses in the cue-abstraction hypothesis space, and 12 out of 24 hypotheses in the permutation-based hypothesis space. In this case, subsequently observing the blue square monster and the blue round monster racing would not be informative at all in the cue-abstraction space, because it does not rule out any of the 4 remaining hypotheses. However, the same observation would rule out 6 out of 12 remaining hypotheses in the permutation-based hypothesis space, making it the most informative action available.

An additional concern in Study 1 was that rather than focusing on the relevant cues (color and shape) to assess the relative speed of the monsters, the children could have used other visual cues such as timing or the distance between the monsters before they reached the finish line. To address this, we obscured the final half of the race track with trees (Figure 1). Furthermore, to better capture the developmental trajectory, we extended our age range to include 7-year-olds. Finally, we introduced a memory load manipulation which determined whether children were provided with memory aids (a record of the previous selections and outcomes). This would help understanding the impact of memory load on children's hypothesis space representation.

Based on the results of Study 1, we generally expected that older children would be more likely than younger children to rely on a cue-abstraction representation of the hypothesis space during search. We also hypothesized that children will be more likely to represent the hypothesis space in a cue-abstraction fashion when assigned to the high memory load condition—where encoding and updating a smaller hypothesis space might be particularly beneficial. Lastly, we expected that all children would be more likely to rely on a cue-abstraction representation of the hypothesis space during the test phase, compared to the active learning phase, as emerged in Study 1.

Method

Participants. Participants were sixty-four 5-year-olds (33 female; $M = 65.51$ months, $SD = 3.48$ months), sixty-seven 6-year-olds (33 female; $M = 78.25$ months, $SD = 3.55$ months), and sixty-eight 7-year-olds (33 female; $M = 89.99$ months, $SD = 3.52$ months). A further 3 participants were excluded because of withdrawal of consent, and 2 because of technical error. Children were recruited and tested in museums in [blind for review]. They were [blind for review] or fluent in [blind for review], were predominantly white and came from various social classes. Ethics approval was obtained by the IRB of the [blind for review] (protocol: [blind for review]) and parents gave informed consent for their children to participate before the start of the study.

Design and procedure. The experiment was carried out on a tablet. The paradigm was identical to the one used in Study 1, with the following modifications: First, children were presented with the four monsters and could freely select the learning pairs they wanted to see racing, instead of having to choose between pre-selected pairs. We incentivized children to search efficiently by initially providing them with 10 stickers, of which they had to "pay" one every time they wanted to see a monster pair racing. They were told that they could keep the stickers left over at the end of the game as a prize. Children had to observe at least three monster pairs racing before moving on to the test phase, but could see up to 10 monster races, using all the stickers they were given. After the third trial, they were asked after each race whether they wanted to see another monster pair racing, or whether they knew what kinds of monsters were faster and therefore were ready to move on to the test phase. Second, the final half of the racetrack was obscured by trees (Figure 1), to prevent children from basing their predictions on other visual cues other than color or shape, such as the distance between the monsters before reaching the finish line. Third, we manipulated the correct monster order between participants. Note that all the implemented orderings were potentially consistent with a cue-abstraction representation of hte hypothesis space, that is, abstraction of cues was possible. Fourth, we manipulated memory load by randomly assigning children to one of two conditions: Children in the *high load* condition did not receive any memory aids, whereas children in the *low load* condition received memory aids, just like in Study 1 (see Figure 1). The test phase remained unchanged, except for the third generalization pair, which was slightly altered to make it more distinct from the pairs encountered during the learning phase (Figure 1).

Results

Active learning phase. Aggregated across age groups and conditions, children observed an average of 3.29 ($SD = 0.74$) races before moving on to the test phase. The number of learning trials did not differ as a function of continuous age ($IR = 0.97$, $[0.89, 1.05]$, $p = .44$) or memory load ($IR = 1.04$, $[0.89, 1.21]$, $p = .64$).

We calculated the EIG of each child's selections according to both the cue-abstraction and the permutation-based representations of the hypothesis space. Overall, the average EIG of children's selections was lower when calculated according to the the cue-abstraction compared to the permutation-based representation of the hypothesis space (see Table 2), suggesting that children's search was generally better captured by assuming a permutation-based representation of the hypothesis space. This result is supported by a latent mixture-models analysis presented below.

A linear regression of cue-abstraction-EIG with continuous age, memory load, and their interaction as predictors showed that it increased with age ($\beta = 0.10$, $p < .001$) and low memory load ($\beta = 0.59$, $p = .003$). There was also a significant Age × Memory load interaction ($\beta = -0.88$, $p = .004$; $F(3, 196) = 7.597$, $p < .001$; $R^2 = .09$), indicating that the effect of memory load on mean cue-abstraction-EIG decreased with age. A linear regression of permutation-based-EIG also showed a main effect of age ($\beta = 0.04$, $p < .001$)

Study 2				
	High load		**Low load**	
	Cue-abstraction	Permutation	Cue-abstraction	Permutation
5-year-olds	.39 (.15)	.56 (.12)	.45 (.16)	.59 (.10)
6-year-olds	.48 (.12)	.62 (.07)	.48 (.13)	.61 (.10)
7-year-olds	.53 (.12)	.62 (.09)	.48 (.11)	.64 (.07)

Table 2

Mean EIG (and SD) of selections calculated according to the cue-abstraction (cue-abstraction-EIG) and permutation-based (Permutation-based-EIG) representations of the hypothesis space, displayed by age group and condition.

but not of memory load ($\beta = 0.02$, $p = .32$). Adding an Age × Memory load interaction did not significantly improve model fit ($\chi^2(1) = 0.016$, $p = .32$).

Children often stopped the search before they had narrowed down the hypothesis space to a single hypothesis, that is, when there was still information to be gained (according to any of the representations of the hypothesis space considered). Although the prevalence of such early stopping declined with age (5-year-olds: 67.69%; 6-year-olds: 67.16%; 7-year-olds: 54.41%), a logistic regression showed that these differences were not significant ($OR = 0.74$, $[0.53, 1.04]$, $p = .081$) and that rates of early stopping were also unaffected by condition ($OR = 1.01$, $[0.56, 1.80]$, $p = .981$). This indicates that children did not markedly differ in their ability to search exhaustively.

Test phase. As mentioned in the results section of Study 1, children's predictions in the generalization trials would be much more strongly supported if they had encoded the information collected in the active learning phase according to a cue-abstraction, compared to a permutation-based representation of the hypothesis space. Aggregating across conditions, performance in the generalization task was above chance only for 6- and 7-year-olds ($M_{5years} = 57\%$, $SD = 32$, two-tailed binomial test: $p = .06$; $M_{6years} = 69\%$,

$SD = 28$, $p < .001$; $M_{7years} = 76\%$, $SD = 25$, $p < .001$). A binary logistic regression showed that the probability of making a prediction compatible with a cue-abstraction organization of the hypothesis space increased significantly with age (Figure 4; $OR = 1.47$, $[1.19, 1.82]$), but there was no significant effect of memory load ($OR = 1.28$, $[0.89, 1.84]$). A higher number of learning trials was negatively associated with the probability of giving a correct response ($OR = 0.76$, $[0.60, 0.96]$).

Differently from Study 1, we found that a higher cue-abstraction-EIG during search increased the probability of giving a correct response ($OR = 5.13$, $[1.48, 17.75]$) in the generalization trials. As expected, the EIG achieved under the permutation-based hypothesis space was unrelated to on the probability of giving a correct response ($OR = 0.20$, $[0.03, 1.24]$). This suggest that those children who were already searching guided by a cue-abstraction representation of the hypothesis space were also more likely to rely on this representation during the generalization trials.

In the podium task, the proportion of children who indicated the correct composition of the podium increased with age and was above chance at all ages (Figure 4): Overall, 18.46% of 5-year-olds, 16.42% of 6-year-olds, and 36.76% of 7-year-olds identified the correct podium. We used logistic regression to assess to what extent age and memory load condition were related to performance in the podium task. We also tested a regression model that included an Age × Condition interaction, but this did not significantly improve model fit ($\chi^2(1) = 1.41$, $p = .24$). Adding mean EIG achieved during learning according to both hypothesis-space models and the number of learning trials as predictors also failed to improve model fit ($\chi^2(3) = 2.073$, $p = .55$). The probability of assembling a correct podium increased significantly with age ($OR = 1.71$, 95% CI $[1.15, 2.54]$) and in the low load condition ($OR = 2.02$, 95% CI $[1.02, 4.00]$).

Mixture-model comparison across active learning and test phases. The behavioral results suggest that children relied on different representations of the hypothesis space at different points in the task, in that selections during search were more consistent

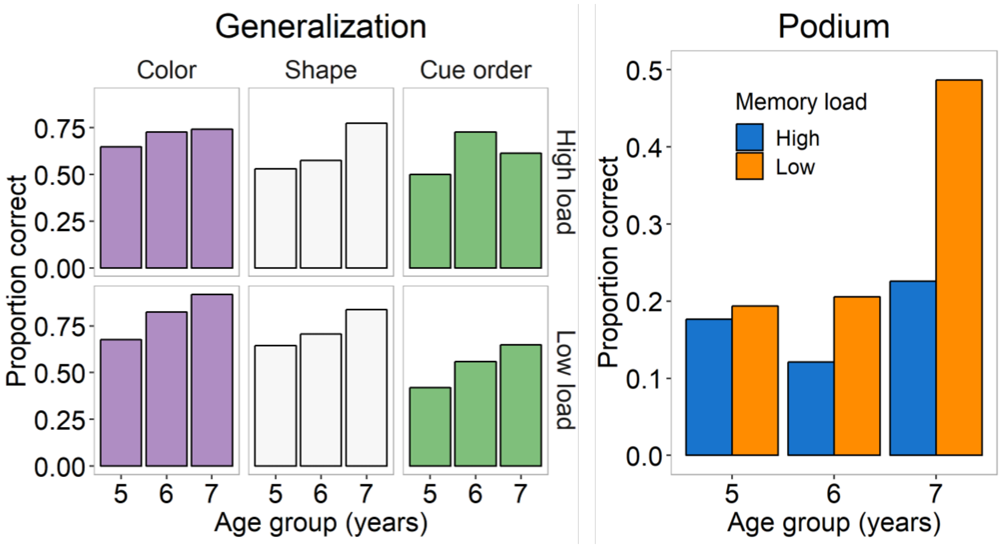


Figure 4. Test performance in Study 2, scored according to the predictions of a cue-abstraction representation of the hypothesis space. Left: Proportion of participants who responded correctly in the generalization task for each trial type (vertical panels) in high and low memory load (horizontal panels). Right: Proportion of participants who got the podium correct in each condition.

with the permutation-based hypothesis space, whereas performance during the test showed evidence of cue abstraction, particularly among older chidren. To directly compare reliance on cue-abstraction versus a permutation-based hypothesis spaces during each phase, we modeled children's selections and predictions using a hierarchical Bayesian latent mixture model (Bramley et al., 2015; Bartlema et al., 2014) (see Supplementary Materials for full details). The mixture model compared three candidate strategies by estimating their probability of having generated participants' choices in each phase: a random strategy (RAND) that corresponded to random selection during the learning phase and guessing in the test phase; the permutation-based strategy (PERM); and the cue-abstraction strategy (CA). Both the PERM and CA strategies predict that selections during the active learning phase are driven by EIG (as calculated using the corresponding hypothesis space; see Supplementary Materials S1) and the resulting outcomes are used to rule out hypotheses that are inconsistent with that evidence. The set of hypotheses which remain at the end of the search phase are then used to predict responses on the generalization and podium tests. The estimated mixture probabilities (θ) indicate the likelihood of each strategy within each group based on how well each strategy fits participants' choices.

This approach complements the behavioral analyses in a number of ways. First, it provides an integrative account of strategy use during the test phase, as all of an individual's responses (generalization trials and podium) are modeled as arising from a common underlying representation of the hypothesis space. Second, it allows us to directly weigh the evidence for each strategy when taking into account the information collected by each individual during learning. This is particularly important because participants selected different sets of observations during the active learning phase, which in turn affects the expected performance on test trials. For example, the PERM and CA strategies both predict low test accuracy among participants who make uninformative selections and end the learning phase without narrowing down either hypothesis space. Similarly, a small number of high-EIG observations may be enough to identify the correct hypothesis under the CA model, but still leave uncertainty under the PERM model. Examining the estimated mixture probabilities therefore provides a clearer indication of a group's reliance on each strategy, and therefore on the different hypothesis space representations, given all observed behaviors in the active learning and the test phase.

Figure 5 shows the posterior means and 95% highest posterior density intervals for the mixture probabilities θ in the learning phase (top row) and the test phase (bottom row). During the active learning phase, echoing the EIG results presented above, all children's selections were better captured by the PERM strategy (see Table S1 for pairwise differences between mixture probabilities), with one exception: For 5-year-olds in the high load condition, the RAND strategy had the highest posterior probability. This suggests that the task demands may have been too high for the youngest age group to implement any systematic search strategy.

In the test phase, which considers together the responses to the generalization and the podium tests, strategy use was less consistent for all age groups (see Table S1 for pairwise comparisons between strategies). The RAND strategy best captured 5-year-olds' predictions in both memory load conditions, as well as 6- and 7-year-olds' predictions in the high load condition. In the low load condition, the PERM strategy was most prominent among for the 6-year-olds, whereas the CA strategy best-captured test performance among 7-year-olds.

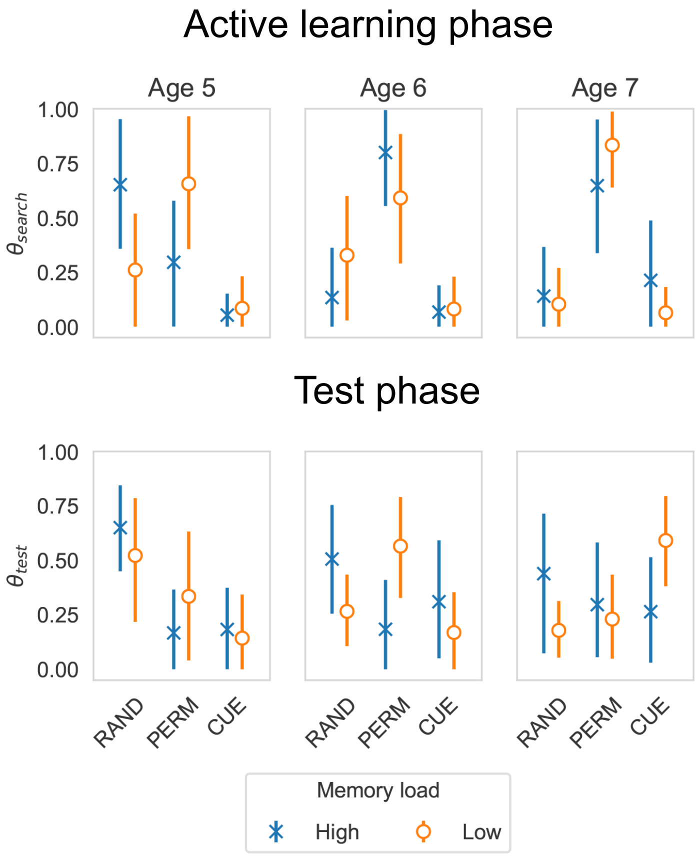


Figure 5. Mixture probabilities of each strategy in Study 2 for the learning phase (top) and test phase (bottom). Error bars represent 95% highest posterior density intervals.

Discussion of Study 2

In Study 2 we compared two possible representations of the hypothesis space—cue-abstraction and permutation-based—in their ability to capture children's selections and predictions in an active learning multiple-cue inference task. We analyzed children's behavior to find evidence for them relying on one or the other when selecting which information to observe and making predictions. We also considered a completely unsystematic strategy that would result in a random selection during search and guessing at test.

The results of Study 2 are in line with those of Study 1, in that they suggest there is a difference in how 6- and 7-year-olds represented the hypothesis space during the active selection versus the test phases. In particular, both the behavioral and computational analyses of older children's selections provide evidence that they are more likely to rely on a permutation-based representation of the hypothesis space during active learning. However, their predictions during the test phase indicate a *shift* in how the hypothesis space is represented. On the one hand, in the test phase, 6- and 7-year-olds in the high memory load condition were now better fit by a random strategy. This suggests that, even when the search was guided by a permutation-based representation of the hypothesis space, integrating the information collected and making informed predictions without memory aids was too challenging. On the other hand, in the low memory load condition, 6-year-olds were still better fit by the permutation-based representation of the hypothesis space, while the predictions of 7-year-olds were better captured by a strategy assuming a cue-abstraction representation of the hypothesis space. In general, 5-year-olds were better captured by a random strategy, indicating that this version of the task might have been too challenging for them even with the benefit of memory aids.

The comparison between the results of Study 1 and 2 also suggest that a high memory load—the absence of the memory aids offered in Study 1—prevents children from being able to abstract the cue-outcome relationship, and in the case of 5-year-olds, to entertain either systematic representation of the hypothesis space. Children in low memory load condition, however, showed a similar developmental trend towards an improved ability to represent the hypothesis space in a cue-abstraction fashion as those in the sample of Study 1, at least in the test phase. It is difficult to compare the results of the active learning phase across the two studies though, as the forced-choice design in Study 1 strongly encouraged a cue-abstraction representation of the hypothesis space and therefore may not reflect how young children spontaneously approach a multiple-cue inference task.

Additionally, in Study 1 children were forced to continue training until they had seen all of the informative races (according to the cue-abstraction representation of the hypothesis space), so that they had all the information needed for abstracting the cue-outcome relationships. Since many children in Study 2 did not search efficiently it is also possible that children did not collect enough information necessary for cue abstraction. This difficulty in *generating* versus *selecting* the most informative actions is in line with developmental findings from question-asking research (Ruggeri et al., 2017, 2016; Ruggeri and Feufel, 2015). This work shows that children younger than 10 years of age have a hard time generating the most informative questions from scratch, although they can already select the most informative questions at age 5.

Although the memory load manipulation did not affect children's search patterns or generalization performance, low memory load was associated with higher accuracy in the podium task (particularly for the 7-year-olds). This indicates that the increased memory load made it more difficult to learn the monster orders seen during learning, but did not prevent older children from abstracting the cue-outcome relationships and make generalizations to new objects. However, all children performed above chance level in the podium task, suggesting that the memory load was not so high as to make learning impossible (even for 5-year-olds). Furthermore, comparing the results of the computational modeling analyses for the 5-year-olds' between conditions suggests that implementing a systematic (i.e., non-random) search strategy may incur lower cognitive load than using the information gathered during search to make inferences.

Conclusion

We presented two studies in which we explored how 5- to 7-year-olds actively search for information to make accurate predictions based on multiple cues. Moreover, using computational modeling, we investigated what children's search patterns and predictions reveal about their representation of the hypothesis space.

The results of Study 1 suggest that 5-year-olds are already able to abstract cue-outcomes relationships and generalize to new exemplars if enough scaffolding is given during the active learning phase. This is a much younger age than typically assumed. Indeed, previous work suggested that children have difficulty identifying relevant cues (Montanelli, 1972; Betsch et al., 2014; Davidson, 1991b) and reasoning about hypotheses in an abstract, hierarchical manner until late childhood (i.e., around age 10; Ruggeri and Feufel, 2015). However, Study 2 showed that even older children struggled to do so without scaffolding, echoing results from studies of question-asking (Ruggeri et al., 2017; Ruggeri and Feufel, 2015). Crucially, both studies provided for the first time evidence that children can *switch* between different representations of the hypothesis space: In some cases, any organization of the hypothesis space that seems to have driven search can dissolve at test, especially when the memory demands are too heavy, with children performing at random at test. In other cases, children seem to be able to spontaneously reorganize the hypothesis space after the search phase, to encode the information collected by abstracting the cue-outcome relationship, in this way finding support for more accurate predictions. For example, although during the learning phase the 5-year-olds in Study 1 and the 7-year-olds in Study 2 seemed to rely on a permutation-based representation of the hypothesis space (when memory aids were provided), they were able to reorganize and shift to a cue-abstraction representation during the test phase (for a related switch in reliance on cue abstraction and exemplar processing between different judgment tasks in adults, see

Pachur and Olsson, 2012; Trippas and Pachur, 2019).

More generally, considering how children may represent the hypothesis space at different ages and at different stages of the learning process could help explain inconsistencies in existing findings. For instance, it is well-established that the efficiency of children's questions increases dramatically between the ages of 5 and 10, with particularly large improvements from age seven (e.g., Ruggeri et al., 2016; Herwig, 1982; Mosher and Hornsby, 1966). However, the factors driving these changes remain poorly understood. While it is obvious that the maturation of children's verbal skills and executive functions are important contributors, another intriguing possibility is that the emergence of the ability to represent relevant information in a hierarchical manner, together with the flexibility to reorganize the hypothesis space in the course of a task to more effectively guide information search and support prediction, may also be crucial.

Similarly, achieving a better understanding of how children's search strategies relate to their representation of the hypothesis space can help explain why interventions aimed at improving children's learning strategies do not always work as intended. For example, attempts to improve the efficiency of children's questions (e.g., Denney and Turner, 1979; Denney, 1972; Courage, 1989) and use of unconfounded experiments in causal learning (e.g., Dean and Kuhn, 2007; Chase and Klahr, 2017) have met with limited success. If students have misunderstandings about the relevant variables or cues, which should be reflected in their hypothesis-space structure, interventions which do not address these misconceptions are unlikely to be effective.

We note that our candidate notions of the hypothesis space were based on research on how adults solve multiple-cue inference tasks (e.g., Juslin et al., 2003a,b; Pachur and Olsson, 2012). It would be beneficial for future research to determine whether the permutation-based and cue-abstraction representations of the hypothesis spaces we have focused on are indeed the most likely representations used by young children in multiple-cue inference tasks. As hinted at by Study 2, children may use other hypothesis-space structures that we did not consider, such as a simpler variant of the cue-abstraction hypothesis space in which only the cue directions are encoded, or a completely different representation of the hypothesis space that requires less cognitive capacity to update and maintain.

References

Bartlema, A., Lee, M., Wetzels, R., and Vanpaemel, W. (2014). A bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, 59:132–150.

Betsch, T., Lang, A., Lehmann, A., and Axmann, J. M. (2014). Utilizing probabilities as decision weights in closed and open information boards: A comparison of children and adults. *Acta Psychologica*, 153:74–86.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., and Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, 64(4):215–234.

Bramley, N. R., Lagnado, D. A., and Speekenbrink, M. (2015). Conservative Forgetful Scholars : How People Learn Causal Structure Through Sequences of Interventions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 41(3):708–731.

Chase, C. C. and Klahr, D. (2017). Invention Versus Direct Instruction: For Some Content, It's a Tie. *Journal of Science Education and Technology*, 26(6):582–596.

Cook, C., Goodman, N. D., and Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 120(3):341–349.

Courage, M. L. (1989). Children's inquiry strategies in referential communication and in the game of twenty questions. *Child Development*, 60(4):877–886.

Davidson, D. (1991a). Children's Decision-Making Examined with an Information-Board Procedure. *Cognitive Development*, 6(1):77–90.

Davidson, D. (1991b). Developmental differences in children's search of predecisional information. *Journal of Experimental Child Psychology*, 52:239–255.

Dean, D. J. and Kuhn, D. (2007). Direct instruction vs. discovery: the long view. *Science Education*, 91(1):36–74.

Denney, D. R. (1972). Modeling and eliciting effects upon conceptual strategies. *Child Development*, 43(3):810–823.

Denney, N. W. and Turner, M. C. (1979). Facilitating cognitive performance in children: A comparison of strategy modeling and strategy modeling with overt self-verbalization. *Journal of Experimental Child Psychology*, 28(1):119–131.

Einhorn, H. J., Kleinmuntz, D. N., and Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, 86(5):465–485.

Gregan-Paxton, J. and Roeder John, D. (1995). Are Young Children Adaptive Decision Makers? A Study of Age Differences in Information Search Behavior. *Journal of Consumer Research*, 21:567âĂŞ–580.

Gregan-Paxton, J. and Roeder John, D. (1997). The Emergence of Adaptive Decision Making in Children. *Journal of Consumer Research*, 24:43–56.

Herwig, J. E. (1982). Effects of age, stimuli, and category recognition factors in children's inquiry behavior. *Journal of Experimental Child Psychology*, 33(2):196–206.

Howse, R. B., Best, D. L., and Stone, E. R. (2003). Children's decision making: the effects of training, reinforcement, and memory aids. *Cognitive Development*, 18:247–268.

Juslin, P., Jones, S., Olsson, H., and Winman, A. (2003a). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):924–941.

Juslin, P., Olsson, H., and Olsson, A.-C. (2003b). Exemplar effects in categorization and multiple-Cue judgment. *Journal of Experimental Psychology: General*, 132(1):133–156.

Legare, C. H., Mills, C. M., Souza, A. L., Plummer, L. E., and Yasskin, R. (2013). The use of questions as problem-solving strategies during early childhood. *Journal of Experimental Child Psychology*, 114(1):63–76.

Mata, R., von Helversen, B., and Rieskamp, J. (2011). When Easy Comes Hard: The Development of Adaptive Strategy Selection. *Child Development*, 82(2):687–700.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., and Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, 141:1–22.

Montanelli, D. S. (1972). Multiple-Cue Learning in Children. *Developmental Psychology*, 7(3):302–312.

Mosher, F. A. and Hornsby, J. R. (1966). On asking questions. *Studies in Cognitive Growth*, pages 86–102.

Pachur, T. and Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, 65(2):207–240.

Pachur, T. and Trippas, D. (2019). Ways to learn from experience. In *Taming uncertainty*, pages 207–222.

Ruggeri, A. and Feufel, M. A. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, 6:918.

Ruggeri, A. and Katsikopoulos, K. V. (2013). Make your own kinds of cues: When children make more accurate inferences than adults. *Journal of Experimental Child Psychology*, 115(3):517–535.

Ruggeri, A. and Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., and Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, 52(12):2159–2173.

Ruggeri, A., Sim, Z. L., and Xu, F. (2017). "Why Is Toma Late to School Again?" Preschoolers Identify the Most Informative Questions. *Developmental Psychology*, 53(9):1620–1632.

Ruggeri, A., Swaboda, N., Sim, Z. L., and Gopnik, A. (2019). Shake it baby, but only when needed: Preschoolers adapt their exploratory strategies to the information structure of the task. *Cognition*, 193:104013.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Schulz, L. E. and Bonawitz, E. B. (2007). Serious Fun: Preschoolers Engage in More Exploratory Play When Evidence Is Confounded. *Developmental Psychology*, 43(4):1045–1050.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Sim, Z. and Xu, F. (2017). Infants preferentially approach and explore the unexpected. *British Journal of Developmental Psychology*, 35:596 – 608.

Stahl, A. E. and Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348:91 – 94.

Trippas, D. and Pachur, T. (2019). Nothing compares: Unraveling learning task effects in judgment and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(12):2239–2266.

von Helversen, B., Mata, R., and Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology*, 46(1):220–229.

Supplementary Materials

S1. Bayesian Models

We modeled behavior in the task as Bayesian belief updating with respect to a hypothesis space $\mathcal{H}$. In the PERM model, $\mathcal{H}$ was comprised of 24 possible orderings of the four monsters shown in Figure 1. In the CA strategy, $\mathcal{H}$ was comprised of 8 possible hierarchical rules involving the two cues (color and shape).

The four monsters were $BS = (blue, square)$, $BC = (blue, circle)$, $GS = (green, square)$, and $GC = (green, circle)$. On each trial of the learning phase a "race" X was observed between two monsters, with a total of six possible unique races. Each hypothesis $h \in \mathcal{H}$ specifies the deterministic likelihood of observing a outcome of a race between two monsters, $p(winner(x) = m|h)$ for $m \in X$, based on their relative rank (PERM model) or the order and direction of the two cues (CA strategy). On trial t, the learner has observed a set of races $\mathcal{D}$ and their joint likelihood under a hypothesis is $p(\mathcal{D}|h)$. The posterior distribution is given by Bayes rule,

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}|h')p(h')}, \tag{1}$$

where the prior $p(h)$ is assumed to be uniform over the hypothesis space. The probability of a new race resulting in the outcome $winner(x)$ is then given by the predictive distribution

$$p(winner(x) = m|\mathcal{D}) = \sum_{h \in H} p(winner(x) = m|h)p(h|\mathcal{D}). \tag{2}$$

Selections during the active learning phase. Selections were modeled using stepwise expected information gain (EIG), which is the expected reduction in Shannon entropy measured over the posterior distribution as a result of observing the outcome of a race. Shannon entropy is given by

$$H(\mathcal{D}) = -\sum_{h \in \mathcal{H}} p(h|\mathcal{D}) \; log \; p(h|\mathcal{D}), \tag{3}$$

where $\mathcal{D}$ is the set of observations thus far. Entropy is maximized when all hypotheses have equal probability according to the posterior distribution and is equal to zero when one hypothesis has $p(h|\mathcal{D}) = 1$. The value of selecting test X is the expected decrease in entropy resulting from each outcome of the test, weighted by its probability of occurring,

$$EIG(X) = \sum_{m \in X} p(\text{winner}(X) = m|\mathcal{D}) \, [H(\mathcal{D}) - H((\text{winner}(X) = m), \mathcal{D})]. \tag{4}$$

Generalization test. There were three trials in the generalization task in which the participant predicted the winner of a race between monsters that were not seen during the learning phase. For the cue-abstraction hypothesis space, the likelihood of each outcome was simply based on the cue order on the known dimension. For example, in the Shape trial participants predicted the outcome of a race between a red square (RS) and red circle (RC). The probability $p(RS > RC)$ was determined by the cue direction for shape under each hypothesis.

For the permutation-based hypothesis space, predictions were based on the relative ranking of monsters that were matched on the feature dimension with an unfamiliar value. For example, in the Shape trial, the probability $p(RS > RC)$ was the proportion of races between monsters of the same color in which the square was higher ranked than the circle. If a given hypothesis specified that a square was faster than a circle for both values of the color dimension (i.e., $BS > BC$ and $GS > GC$), the probability of a red square being faster than a red circle was 1. In contrast to the CA strategy, under the PERM strategy it was possible to have reversed orderings (e.g., the square being faster when the shapes were blue, but slower when the shapes were green, as in the ordering $BS > BC > GC > GS$) which would lead to more uncertainty about the outcome of the novel generalization pair.

Podium test. In the podium task, the participant is asked to identify the correct ordering of the four monsters seen during the learning phase. Under the Bayesian model, the probability of each response is simply the posterior probability of the corresponding hypothesis.

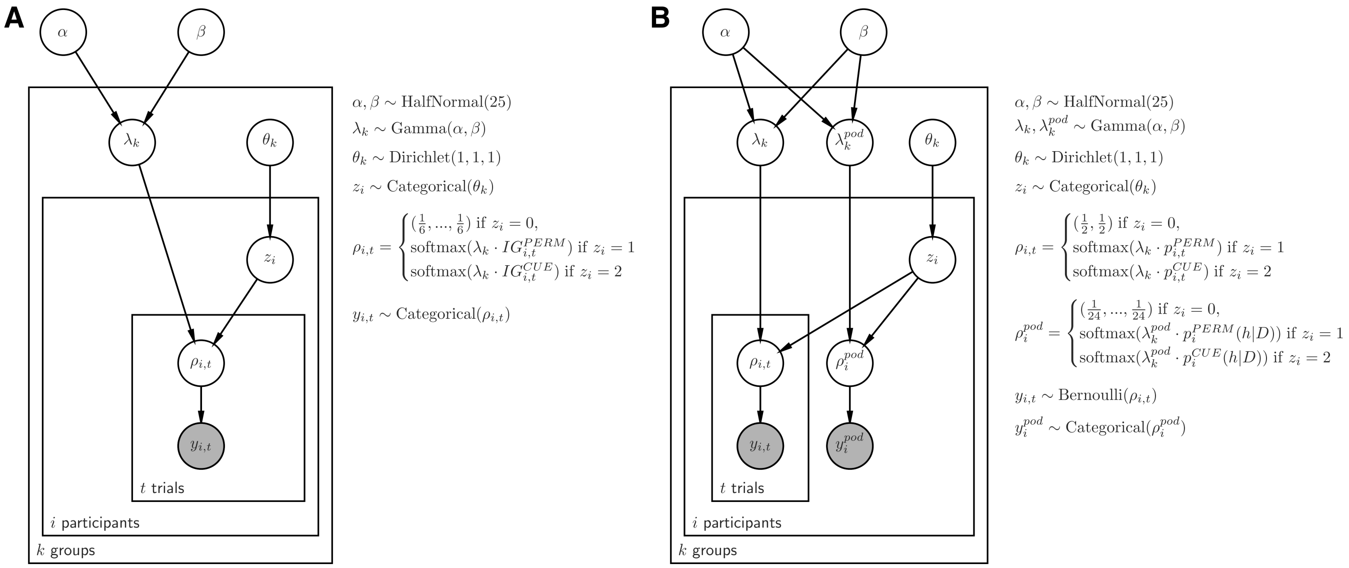
S2. Latent Mixture Modeling in Study 2

We used hierarchical Bayesian latent mixture modeling to estimate the relative probabilities of cue-abstraction and permutation-based reasoning during the learning and test phases. Performance in each phase was modeled as a mixture of three strategies: a random strategy (RAND) that corresponded to random search during learning and guessing at test; the permutation-based model (PERM); and the cue-abstraction model (CA).

The model specifications are shown in Figure S1 for each phase. There were six groups in Study 2 (3 age groups × 2 conditions). The probability of each strategy in group k was denoted by the mixture probability θ_k, with each strategy assigned an equal prior probability. The mixture probabilities determine the probability of an individual adopting a particular strategy, where the chosen strategy for participant i is denoted by z_i. A value of $z_i = 0$ corresponds to the RAND strategy, under which all search and test choices have equal probability. If an individual adopts the PERM ($z_i = 1$) or CA ($z_i = 2$) strategies, the corresponding hypothesis space was used to evaluate the likelihood of their selections (learning phase) or test responses (test phase), based on the Bayesian updating model described in Section S1. Choices in each phase were modeled using softmax functions with group-specific inverse-temperature parameters λ_k (see below for details of each phase). Inverse-temperature parameters were assigned a prior distribution of $Gamma(\alpha, \beta)$, where α and β were hyperparameters that were common to all groups.

Learning phase. On each trial in the learning phase participants could select one of six possible observations (pairwise match-ups of monsters). Under the random strategy, all six observations were equally likely on every trial. Under the PERM and CA strategies, the probability $\rho_{i,t}$ of choosing observation i on trial t was modeled with a softmax choice function over the EIG of all possible observations on that trial. The strategy adopted by an individual determined whether EIG was calculated according to the PERM or CA strategy.

Test phase. During the test phase, each participant made four responses: three generalization responses (shape, color, and order) and one podium response. For the



$$\alpha, \beta \sim \text{HalfNormal}(25)$$

$$\lambda_k \sim \text{Gamma}(\alpha, \beta)$$

$$\theta_k \sim \text{Dirichlet}(1, 1, 1)$$

$$z_i \sim \text{Categorical}(\theta_k)$$

$$\rho_{i,t} = \begin{cases} (\frac{1}{6}, ..., \frac{1}{6}) \text{ if } z_i = 0, \\ \text{softmax}(\lambda_k \cdot IG_{i,t}^{PERM}) \text{ if } z_i = 1 \\ \text{softmax}(\lambda_k \cdot IG_{i,t}^{CUE}) \text{ if } z_i = 2 \end{cases}$$

$$y_{i,t} \sim \text{Categorical}(\rho_{i,t})$$

$$\alpha, \beta \sim \text{HalfNormal}(25)$$

$$\lambda_k, \lambda_k^{pod} \sim \text{Gamma}(\alpha, \beta)$$

$$\theta_k \sim \text{Dirichlet}(1, 1, 1)$$

$$z_i \sim \text{Categorical}(\theta_k)$$

$$\rho_{i,t} = \begin{cases} (\frac{1}{2}, \frac{1}{2}) \text{ if } z_i = 0, \\ \text{softmax}(\lambda_k \cdot p_{i,t}^{PERM}) \text{ if } z_i = 1 \\ \text{softmax}(\lambda_k \cdot p_{i,t}^{CUE}) \text{ if } z_i = 2 \end{cases}$$

$$\rho_i^{pod} = \begin{cases} (\frac{1}{24}, ..., \frac{1}{24}) \text{ if } z_i = 0, \\ \text{softmax}(\lambda_k^{pod} \cdot p_i^{PERM}(h|D)) \text{ if } z_i = 1 \\ \text{softmax}(\lambda_k^{pod} \cdot p_i^{CUE}(h|D)) \text{ if } z_i = 2 \end{cases}$$

$$y_{i,t} \sim \text{Bernoulli}(\rho_{i,t})$$

$$y_i^{pod} \sim \text{Categorical}(\rho_i^{pod})$$

Figure S1. Plate diagrams for hierarchical mixture models in the learning phase (A) and test phase (B) for Study 2.

random strategy, all choices were equally likely. For the PERM and CA strategies, the probability of possible responses were based on the respective hypothesis space and the evidence an individual observed during the learning phase. Choices are again modeled using a softmax function with an inverse-temperature parameter λ_k^{test}.

For responses in the podium task, the probability of selecting a podium was similarly defined via a softmax choice function over the posterior distribution for the corresponding model, with a separate inverse-temperature parameter λ_k^{pod}.

Parameter estimation. Parameters were estimated using Markov Chain Monte Carlo (MCMC) with the PyMC Python library (Salvatier et al., 2016). Parameters were estimated separately for the learning and test phases. For both models we used four MCMC chains with 20,000 samples and 2,000 burn-in iterations. All estimates converged as indicated by Gelmin-Rubin ratios under 1.05.

			Mean difference [95% HDI]	
Age	Memory load	Comparison	Active learning phase	Test phase
5	High	PERM - RAND	-.36 [-.93, .23]	-.48 [-.82, -.13] *
		CA - RAND	-.60 [-.92, -.26] *	-.47 [-.79, -.12] *
		CA - PERM	-.24 [-.60, .06]	.02 [-.34, .37]
	Low	PERM - RAND	.40 [-.16, .96]	-.19 [-.70, .38]
		CA - RAND	-.18 [-.52, .13]	-.38 [-.75, .03]
		CA - PERM	-.57 [-.96, -.18] *	-.19 [-.61, .26]
6	High	PERM - RAND	.67 [.20, .99] *	-.32 [-.70, .09]
		CA - RAND	-.07 [-.35, .18]	-.20 [-.66, .28]
		CA - PERM	-.73 [-.99, -.42] *	.13 [-.32, .58]
	Low	PERM - RAND	.26 [-.32, .81]	.30 [-.06, .65]
		CA - RAND	-.25 [-.58, .11]	-.10 [-.38, .18]
		CA - PERM	-.51 [-.86, -.14] *	-.40 [-.76, .01]
7	High	PERM - RAND	.51 [.04, .95] *	-.14 [-.61, .45]
		CA - RAND	.07 [-.32, .48]	-.18 [-.61, .38]
		CA - PERM	-.43 [-.94, .13]	-.03 [-.48, .36]
	Low	PERM - RAND	.73 [.39, .99] *	.05 [-.21, .34]
		CA - RAND	-.04 [-.28, .17]	.41 [.12, .70] *
		CA - PERM	-.77 [-.99, -.50] *	.36 [-.03, .73]

Table S1

Pairwise differences between mixture probabilities θ for each strategy in the active learning and test phases. Differences where 95% HDIs do not overlap with 0 are marked with an asterisk (*).