Sources of Developmental Change in the Efficiency of Information Search

Azzurra Ruggeri¹,², Tania Lombrozo¹, Thomas L. Griffiths¹, Fei Xu¹

¹Department of Psychology, University of California, Berkeley
²Max Planck Institute for Human Development, Berlin, Germany

Correspondence concerning this article should be addressed to Azzurra Ruggeri, Department of Psychology, University of California, Berkeley, Tolman Hall #1650, Berkeley, CA 94720-1650
Phone: +1 510-642-5292. E-mail: a.ruggeri@berkeley.edu

Acknowledgments

We would like to thank Celina Vicuna, Minh-Thy Nguyen, Meg Bishop, Samika Kumar, Simone Riley, and Justine Hoch for assistance in recruiting, data collection, and coding, as well as Zi Lin Sim and Anita Todd for their helpful comments. We would also like to thank all the students and teachers from the Madera Elementary School, in El Cerrito, and the Laurence Hall of Science, in Berkeley, for providing research space. This research was supported by a Seventh European Community Framework Programme Marie Curie International Outgoing Fellowship awarded to A. Ruggeri. T. Lombrozo was supported by a McDonnell Scholar Award on Understanding Cognition. F. Xu was supported by a China Distinguished Foreign Expert Award.
Abstract

Children are active learners: they learn not only from the information people offer and the evidence they happen to observe, but by actively seeking information. However, children’s information search strategies are typically less efficient than those of adults. In two studies, we isolate potential sources of developmental change in how children (7- and 10-year-olds) and adults search for information. To do so, we develop a hierarchical version of the 20-questions game, in which participants either ask questions (Study 1) or test individual objects (Study 2) to discover which category of objects within a nested structure (e.g., animals, birds, or owls) has a novel property. We also develop a computational model of the task, which allows us to evaluate performance in quantitative terms. As expected, we find developmental improvement in the efficiency of information search. In addition, we show that participants’ performance exceeds random search, but falls short of optimal performance. We find mixed support for the idea that children’s inefficiency stems from difficulty thinking beyond the level of individual objects or hypotheses. Instead, we reveal a previously undocumented source of developmental change: Children are significantly more likely than adults to continue their search for information beyond the point at which a single hypothesis remains, and thus to ask questions and select objects associated with zero information gain. This suggests that one crucial source of developmental change in information search efficiency lies in children’s “stopping rules.”

Keywords: information search, active learning, 20-questions game, cognitive development
Sources of Developmental Change in the Efficiency of Information Search

Children actively engage with their environment from an early age. As soon as they can sit or walk, infants spontaneously grab and manipulate objects and approach or avoid people. As language develops, young children ask about the meaning of words, request the labels of objects, and inquire about the many new and puzzling phenomena they encounter. Piaget (1954) theorized that active engagement with the world is a crucial component of learning. As children explore, they come across information that does not fit within their existing conceptual structures. This cognitive disequilibrium generates a state of uncertainty that is uncomfortable, and that motivates children to devote their mental energy and efforts to adapting or developing new conceptual structures that better accommodate the new information. Piaget advocated the view that children are active learners, and described behaviors of his own children that looked very much like testing hypotheses with self-generated data.

Although Piaget did not develop a full-fledged theory of how active learning works, more recent and systematic research in cognitive development has supported the idea that children are indeed active learners. For instance, there’s evidence that infants, toddlers, and preschoolers control their attention based on how much information can be gained from the environment (e.g., Kidd, Piantadosi, & Aslin, 2012), and that even young children have some rudimentary ability to generate evidence to figure out the causal structure of the world (e.g., Schulz, Gopnik, & Glymour, 2007; Bonawitz, van Schijndel, Friel, & Schulz, 2012; Sim & Xu, 2014).

Beyond direct exploration through observations and interventions, children can actively gather evidence from people by asking questions (see Chouinard, 2007; Callanan & Oakes, 1992; Campos & Stenberg, 1981; Meltzoff, 1988a, 1988b, 1990). When engaged in conversations with adults, preschool-aged children ask an average of over 100 questions per hour (see Chouinard, 2007). According to Tomasello (1999a, 1999b), the ability to seek out information from knowledgeable others gives us a particular evolutionary advantage, allowing us to learn efficiently. Indeed, research to date has shown
that young children ask questions that support the acquisition of relevant and reliable information. For example, preschool-aged children ask domain-appropriate questions (Greif, Kemler Nelson, Keil and Guterrez, 2006), and they preferentially consult reliable informants (Birch, Akmal, & Frampton, 2010; Birch, Vauthier, & Bloom, 2008; Corriveau & Harris, 2009; Gweon, Pelton, Konopka, & Schulz, 2014; Koenig & Harris, 2005; Mills, Legare, Bills, & Mejias, 2010; Mills, Legare, Grant, & Landrum, 2011; Sabbagh & Baldwin, 2001; Sobel & Corriveau, 2010). Even 2- and 3-year-olds have reasonable expectations about which responses count as satisfying answers to their questions: They tend to agree and ask follow-up questions when adults provide explanatory answers, but re-ask their original question, or provide their own explanations, otherwise (Frazier, Gelman, & Wellman, 2009; see also Lombrozo, in press). Finally, we know that 4-year-olds have the capacity to acquire relevant information through strategic questions and that by age 5 they reliably use this information to solve problems (see Chouinard, 2007; Legare, Mills, Souza, Plummer, & Yasskin, 2013; Ruggeri & Feufel, 2015; Ruggeri & Katsikopoulos, 2013).

These findings provide compelling evidence that children’s questions play an important role in learning: They are generated with expectations about the content and reliability of possible answers, and those answers are evaluated and eventually used to guide behavior. At the same time, there’s evidence that children’s questions tend to be inefficient. To investigate the efficiency of children’s questions, studies have used variants of the “20-questions game,” in which participants try to identify an unknown target object by asking as few yes-or-no questions as possible, either generating the questions from scratch (e.g., Chouinard, 2007; Legare et al., 2013; Mosher & Hornsby, 1966; Ruggeri & Feufel, 2015; Ruggeri & Lombrozo, 2015) or selecting them from a list of given alternatives (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri & Feufel, 2015). This research has found that the ability to ask efficient questions undergoes a large developmental change from age four to adulthood. Younger children tend to ask “hypothesis-scanning” questions, which offer tentative solutions (e.g., “Is
it the dog?”) and typically support a less efficient path to the correct solution (Ruggeri & Lombrozo, 2015). In contrast, older children and adults more readily ask “constraint-seeking” questions, which can more efficiently partition the hypothesis space by targeting features or superordinate categories that are shared by multiple hypotheses (e.g., “Does it have a tail?” or “Is it an animal?”).

It thus appears that asking questions is a powerful mechanism for learning, but one that children do not master until later childhood. This raises important questions about the sources of younger children’s inefficiency, and about what changes over the course of development. In the present work, we take a new approach to studying these questions, testing 7-year-olds, 10-year-olds, and adults in a novel information search paradigm. Specifically, we employ two “hierarchical” versions of the 20-questions task in order to isolate distinct hypotheses about potential sources of inefficiency. In this hierarchical task, objects conform to a hierarchical structure (e.g., there are two owls, in a set of four birds, in a set of eight animals), and the solution to the task can occur at any level (e.g., owls, birds, or animals).

Our hierarchical task allows us to explore two hypotheses. The first hypothesis is that children’s inefficiency stems from difficulty going beyond the level of individual objects (e.g., “Is it this dog?”) to ask questions that target higher-order properties of objects, such as shared features (e.g., “Does it have a tail?”) or category membership (e.g., “Is it an animal?”). The second hypothesis is that children experience difficulty going beyond individual hypotheses (whether hypotheses correspond to individual objects or to groups of objects) to generate questions that target multiple hypotheses. Having solutions that can correspond to groups of objects (e.g., dogs or animals), allows us to pull apart the object level from the hypothesis level.

We also analyze our results within a formal computational framework, which allows us to measure performance in absolute terms. By differentiating between non-optimal questions (i.e., those that are informative, but not optimally) and unnecessary questions (i.e., those that are not informative at all), this quantitative analysis allows us to investigate a third hypothesis about children’s inefficiency in asking
questions: that they fail to recognize when enough information has been collected to solve the task, and therefore continue asking questions beyond the point at which they are informative. In the sections that follow, we explain these alternative hypotheses and how our tasks and formal framework address them.

**Young children’s inefficiency: single objects or single hypotheses?**

Why do younger children tend to ask hypothesis-scanning questions in a 20-questions task, despite their typical inefficiency? A dominant explanation is that young children have a hard time going beyond the object level, failing to spontaneously represent, and therefore target, more abstract categories or features. Consistent with this idea, Ruggeri and Feufel (2015) found that scaffolding higher-level representations facilitated children’s ability to ask constraint-seeking questions. In their study, 7- and 10-year-old children, as well as adults, were presented with 20 cards on a computer screen, each of which contained a word label (e.g., “dog” or “sheep”). Participants were randomly assigned to one of two experimental conditions based on the specificity of the label: a basic-level condition (e.g., “dog”) or a subordinate-level condition (e.g., “Dalmatian”). The authors found that participants were more likely to ask constraint-seeking questions in the former condition than in the latter, suggesting that more abstract labels facilitated a shift away from object-based reasoning when generating questions. They also found that the ability to readily generate more abstract features (e.g., “a dog is a mammal” or “a dog has four legs”) is one factor that affects performance and that developed within their age range (see also Herwig, 1982). With even younger children, Legare and colleagues (2013) found that the ability to identify and to flexibly categorize objects based on alternative features (e.g., color and pattern) predicted how well 4- to 6-year-olds could generate effective questions.

The traditional 20-questions task used in prior research has a crucial limitation: the solution to the task is always a single object. As a result, individual *hypotheses* (i.e., candidate solutions) necessarily correspond to individual *objects* (e.g., the dog). A tendency to ask hypothesis-scanning questions could reflect a tendency to represent the problem at the *object level*, a tendency to do so in terms of *individual*
hypotheses, or both. In other words, the traditional 20-questions task cannot differentiate between two possibilities: a) that younger age groups (i.e., 7-year-olds as compared to 10-year-olds, and 10-year-olds as compared to adults) have more difficulty going beyond object-level representations to generate questions that target higher-order features (such as “mammal” for animals, or “brown-haired” for people), and b) that younger age groups have a harder time going beyond individual hypotheses when representing the hypothesis space, and therefore fail to ask questions that can efficiently partition the hypothesis space by targeting multiple hypotheses at once.

To differentiate between these possibilities, we developed a novel, hierarchical version of the 20-questions task in which the solution was a category of objects (e.g., “all animals” or “all birds”) rather than an individual object (e.g., “this blue owl”), with the consequence that hypotheses do not correspond to individual objects. More specifically, our hierarchical 20-questions task differs from traditional versions in two ways: (1) participants receive an array of objects that can be classified into a symmetrical nested structure organized at three category-levels (see Table 1), each including the same number of objects: superordinate (e.g., eight animals and eight plants), basic (e.g., among the animals there are four fish and four birds), or subordinate (e.g., among the birds, there are two different owls and two different parrots); (2) participants are told that some objects share a novel property (e.g., existing on an alien planet, or activating a machine) and are asked to find out which objects do so. Participants’ task is not to identify one object, but to discover at what category-level objects share a novel property (e.g., is it the animals, the birds, or the owls that are found on a planet Apres?), where the solution can occur at any category level. For instance, the solution could be that “birds” are found on the planet, in which case participants would be told “yes” if they asked whether the owls are on planet Apres, and “some” if they asked whether animals are on planet Apres.
Table 1

Materials and Scenarios Used in Studies 1 and 2

<table>
<thead>
<tr>
<th>Category level</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What kind of objects are present on the far-away planet Apres?</td>
</tr>
<tr>
<td>Superordinate</td>
<td>Animals</td>
</tr>
<tr>
<td>Basic</td>
<td>Fish</td>
</tr>
<tr>
<td>Subordinate</td>
<td>Goldfish</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What kind of objects can turn on this machine?</td>
</tr>
<tr>
<td>Superordinate</td>
<td>Clothes</td>
</tr>
<tr>
<td>Basic</td>
<td>Shirts</td>
</tr>
<tr>
<td>Subordinate</td>
<td>Long sleeves</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What kind of objects can enter a magic box?</td>
</tr>
<tr>
<td>Superordinate</td>
<td>Vehicles</td>
</tr>
<tr>
<td>Basic</td>
<td>Cars</td>
</tr>
<tr>
<td>Subordinate</td>
<td>Vans</td>
</tr>
</tbody>
</table>

The studies we report employed a second innovation relevant to differentiating between the theories that younger children are object-bound versus hypothesis-bound. We developed two versions of the hierarchical 20-questions task: one in which participants could ask questions (question-asking) and another in which they could only select individual objects to test (object-selection). Whereas questions can target individual objects (e.g., “is the blue owl on planet Apres?”) or individual hypotheses at different levels (e.g., “are all the animals on planet Apres?” or “are all the birds on planet Apres?”), the object-selection task requires that all participants work at the object level, as they effectively have no choice but to “ask” about individual objects (e.g., the blue owl).

If a key source of developmental change in the efficiency of information search concerns the ability to generate questions that go beyond individual objects by targeting higher-order features, then we would expect a developmental improvement in performance in our question-asking version of the hierarchical 20-questions task. Specifically, we would expect older age groups to be better able to generate questions at higher levels of the hierarchy (e.g., “are all the animals found on the planet?”).
implementing a more efficient search than that available to object-bound younger children (e.g., “is the blue owl on the planet?”). However, we would expect this developmental difference to disappear in the object-selection version of the task, which forces both adults and children to proceed with an object-based search.

If children’s inefficiency instead (or additionally) stems from difficulty reasoning beyond individual hypotheses, we would predict a different profile for performance. Specifically, we would expect older age groups to more readily recognize the relationships between hypotheses and therefore to generate queries “top-down,” capitalizing on the hierarchical structure of the hypothesis space to begin testing superordinate hypotheses first, then basic, and then subordinate. As we show with simulations involving a formal model of the task (detailed below), solutions at higher levels of the hierarchy (e.g., “animals”) can be reached more efficiently than those at lower levels (e.g., “owls”), not only for question-asking, but also for object-selection. If only adults can exploit this feature of the hierarchically-structured hypothesis space, we would expect them to reach solutions more quickly for solutions at superordinate versus subordinate levels for both question-asking and for object-selection. However, we would expect children to be less capable of doing so in either case.

Young children’s inefficiency: non-optimal questions or unnecessary questions?

We suggested above two different reasons why children could fail to generate the most effective questions. An additional possibility is that children don’t (only) ask non-optimal questions (i.e., those that are only moderately informative), but unnecessary questions (i.e., those that provide no information at all). Specifically, younger children, older children, and adults could differ in their “stopping rule,” or in their ability to effectively apply it. A stopping rule establishes when enough information has been collected to terminate information search and make a decision, and it is thought to be a crucial building block for decision making (Gigerenzer, Todd, & the ABC Research Group, 1999). If younger age
groups have an overly conservative stopping rule, they may continue to ask questions beyond the point at which the answers yield information relevant to solving a given problem.

Previous research in decision-making involving other information search paradigms—such as information boards from which children can reveal information prior to making a decision—has found that younger children tend to adopt more exhaustive and inefficient approaches than older children and adults (Davidson, 1991a; 1991b; 1996; Gregan-Paxton & John, 1995, 1997; Howse, Best, & Stone, 2003; Mata, von Helversen, and Rieskamp, 2011; Ruggeri & Katsikopoulos, 2013). These results suggest that younger children’s stopping rules could be more conservative, or simply less accurate, than those of older children and adults, potentially leading them to ask unnecessary questions in a 20-questions paradigm. In fact, Legare and colleagues (2013) found that about 20% of the questions generated by preschoolers were “confirmatory” in the sense that they requested information that had already been provided. However, prior work has not analyzed search performance in quantitative terms, which makes it difficult to differentiate non-optimal queries from unnecessary queries. For our tasks, we developed a model of each participant’s hypothesis space at each time point, allowing us to evaluate whether children ask questions or select objects beyond the point at which a single hypothesis remains.

Before presenting the details of each study, we present a formal analysis of performance on these tasks, which is necessary to explore and disentangle potential sources of developmental change in the efficiency of information search. This framework is built on a Bayesian model of learning, with optimal learning defined in terms of information gain. We explain each of these components in turn.

**Bayesian Model**

Performance on the hierarchical 20-questions task can be modeled within a Bayesian framework originally developed for concept learning and generalization (Tenenbaum & Griffiths, 2001). The learner’s hypothesis space is the set of hypotheses concerning which set of objects has a target property (e.g., activating a machine). In our case, the hypothesis space consists of 14 alternative hypotheses,
corresponding to all the object categories at any hierarchical level: the two superordinate levels, the four basic levels, and the eight subordinate levels.

We did not consider single-object hypotheses (e.g., only the blue owl), as participants were explicitly told that the property applied to more than one object. Moreover, we did not consider disjunctive hypotheses, that is, the combination of objects across categories, such as “a boot or a desk can turn on the machine.” Such hypotheses were never provided by participants as possible solutions. Finally, because participants were told that all categories, at any level, were equally likely to be true, we assumed that participants initially expected all hypotheses to be equally likely, regardless of their level in the hierarchy.

We assume that participants update their beliefs after each observation by evaluating the hypotheses still under consideration according to Bayes’ rule, computing their posterior probability \( p(h|X) \) given all observations \( X \), which is proportional to the product of their prior probabilities \( p(h) \) and likelihoods \( p(X|h) \):

\[
p(h|X) = \frac{p(X|h)p(h)}{\sum_{h'} p(X|h')p(h')}
\]

The prior \( p(h) \) represents participants’ expectations about how likely the candidate hypotheses are. Matching the instructions given to participants, we assume a uniform prior. The likelihood \( p(X|h) \) represents how likely it is that \( X \) will be observed if \( h \) is true. In line with the structure of the task, we assume that for each observation \( x \in X \), \( p(x|h) \) is 1 if the observation is compatible with \( h \) and 0 otherwise, and that observations are independently conditioned on \( h \), so \( p(X|h) \) is just the product of \( p(x|h) \) for each observation \( x \). The posterior \( p(h|X) \) is thus a function of the observations \( X \) and prior beliefs about the probability of each candidate hypothesis considered.
Expected Information Gain

Several possible measures can be used to compute how informative each search option is, whether it’s a question to ask or an object to test. These include probability gain, impact, and expected savings (see Nelson, 2005). Following previous research on the 20-questions task (Nelson et al., 2014; Ruggeri & Lombrozo, 2015; Ruggeri & Feufel, 2015), we measured the informativeness of participants’ queries in terms of their expected information gain (e.g., Oaksford & Chater, 1994). At each step of the search process, an optimal learner evaluates all possible actions in terms of their information gain, IG, computed by subtracting the predicted posterior entropy from the prior entropy:

\[ IG = H_{\text{prior}} - H_{\text{posterior}}. \]

The entropy \( H \) embodies the uncertainty about which of the candidate hypotheses is true. Its computation is based on the probabilities of each of the candidate hypotheses:

\[ H_{\text{prior}} = - \sum_h p(h) \log_2 p(h) \]

The prior entropy \( H_{\text{prior}} \) defines the status of uncertainty preceding every action. The predicted posterior entropy \( H_{\text{posterior}} \) refers to the predicted uncertainty after the action is chosen and the corresponding feedback is observed. The predicted posterior entropy is measured as the sum of the entropies corresponding to each possible future scenario weighted according to the probability of that scenario:

\[ H_{\text{posterior}} = \sum_{x_i} p(x_i | X) H(x_i) \]

where \( x_i \) is a possible observation, \( p(x_i | X) \) is the probability of that observation resulting from taking the candidate action given all the information from previous observations, and \( H(x_i) \) is the entropy of the posterior distribution over hypotheses after observing \( x_i \). More formally,

\[ p(x_i | X) = \sum_h p(x_i | h)p(h | X) \]
$$H(x_i) = -\sum_h p(h|X,x_i) \log_2 p(h|X,x_i)$$

With these formal tools, we can compare children’s and adults’ performance to that of an optimal learner and better identify sources of inefficiency in their information search.

**Overview of studies**

In two studies, we explore the three potential sources of inefficiency in children’s questions identified above. These sources are not mutually exclusive – children could face challenges going beyond the object level, going beyond individual hypotheses, and implementing an effective stopping rule. However, these sources can be isolated and identified with appropriate tasks and analyses. Our two studies use the hierarchical 20-questions task, one involving question-asking (Study 1) and the other object-selection (Study 2). In both studies we tested three age groups: 7-year-olds, 10-year-olds, and adults. These age ranges were motivated by prior research suggesting a strong developmental shift in children’s information search strategies between the ages of 7 and 10 (see Mosher & Hornsby, 1966; Ruggeri & Feufel, 2015; Ruggeri & Lombrozo, 2015).

**Study 1: Question asking**

In Study 1, participants completed the hierarchical 20-questions task by asking questions. We additionally modeled optimal and random performance using our formal analysis. The task and analyses allowed us to test three hypotheses. First, if children’s inefficiency is related to difficulty going beyond the object level to identify higher-order features to target in generating questions, then we expect to replicate prior work in finding a large developmental change in the efficiency of information search, with adults requiring fewer questions to reach the solution than 7-year-olds, and 10-year-olds falling in between. Second, if children’s inefficiency as compared to adults is related to the challenge of going beyond individual hypotheses to ask questions that target multiple hypotheses at once, then we would additionally expect children, but not adults, to fail to identify the solution with fewer questions when it
is at a higher level. Finally, if children’s inefficiency stems from an overly conservative stopping rule, then we would expect to see children continuing to ask questions beyond the point at which only a single hypothesis remains.

**Method**

**Participants.** Participants in Study 1 were 24 children in second grade (10 female, \(M_{age} = 90.5 \) months; \(SD = 5.56\)), and 23 children in fifth grade (8 female, \(M_{age} = 119.4 \) months; \(SD = 12.7\)), recruited from a primary school and a local children’s museum, as well as 23 university students (15 female, \(M_{age} = 21.1 \) years; \(SD = 2.6\)). Although most participants were from White, middle-class backgrounds, a range of ethnicities reflecting the diversity of the population were represented.

**Materials.** We used three scenarios for the hierarchical 20-questions task (see Table 1), each involving 16 distinct objects. In the Machine scenario, the participants had to find out what kind of objects can turn on a machine; in the Planet Apres scenario, participants had to find out what kind of objects can be present on a far-away planet; and in the Magic Box scenario, participants had to find out what kind of objects can enter a magic box. Participants completed all three scenarios in a random order, with the three hierarchical levels of the objects constituting the solution (subordinate, basic, and superordinate) randomly assigned to the three scenarios.

**Procedure.** The experimental session started with a short familiarization phase, aimed at making the children comfortable playing with the tablet on which the objects were displayed and following the experimenter’s instructions. In three trials, participants were presented with an array of six objects and had to identify and select some of the objects by touching them on the screen in response to the experimenter’s instructions. For example, they were asked to touch all the circles or all the stars.
Figure 1. First steps of the step-optimal search path for the question-asking (Study 1, top panel) and object-selection (Study 2, bottom panel) paradigms in the Planet Apres scenario.
After the familiarization, participants began the first of three test trials, in which they had to identify the target category by asking yes-or-no questions. After each question (e.g., “Can birds be found on planet Apres?”), participants received feedback from the experimenter with a response of “yes,” “no,” or “some.” The response “some” was provided when only some members (e.g., parrots) of the category targeted (e.g., birds) shared the novel property (e.g., being found on planet Apres). Participants were prompted to put a red (“no” feedback) or green (“yes” feedback) frame around the object(s) to which their question referred by touching the object(s) on the tablet. For example, if the experimenter said “yes” to whether birds make the machine turn on, the participant was asked to put a green frame around each bird. The experimenter made sure that participants framed all and only the objects each question applied to. This procedure ensured that the participants understood the feedback, and it reduced memory demands, as participants did not have to remember their questions or the feedback received as the task proceeded. Cross-category questions – that is, those that targeted groups of objects belonging to different categories (e.g., “Does it have something brown?” in the Planet Apres scenario, which targets all Trees and the Brown Owl) – were not answered: Children were told that the computer did not know the answer to that question, and they were prompted to ask another question. After receiving feedback, participants could choose whether to ask another question or guess the solution. Participants could ask questions and guess the solution as often as they wanted, but they were told to find the solution with as few questions as possible.

At the end of the three trials constituting the experimental session, participants performed a sorting task to determine whether they understood the hierarchical structure of a set and were able to verbally label categories at each level. They were given 16 cards, representing the objects with which they were presented in the second experimental trial, and had to sort them into two piles (superordinate level) such

---

1 Overall, participants asked 29 cross-category questions, that is, questions that targeted groups of objects belonging to different categories, such as “Does it have something brown?” in the Planet Apres scenario (N_{5-year-olds} = 18; N_{10-year-olds} = 3; N_{adults} = 8). Such questions were not answered (“The computer does not know the answer to this question, can you try another question?”) and were excluded from further analyses.
that the two groups differed from each other but the objects within each group were similar. The second step required sorting the cards within each of these two subgroups into two new piles (basic level), and the third step required sorting the cards within each of these four subgroups into two new piles (subordinate level). The sorting task eventually produced eight groups of two objects each. At each step, participants were asked to name each group they had sorted the objects into. When participants, at any stage of the sorting task, did not organize the objects according to the expected hierarchical categorization (e.g., if they sorted them by color), the experimenter prompted the participant to sort the objects differently (i.e., “is there another way to sort the objects into two groups?”), without suggesting any specific way to do so.

**Simulations**

To model the behavior of an optimal learner, we simulated 10,000 step-optimal search trajectories for the question-asking paradigm. A step-optimal search trajectory is one that selects, at each step, the question with the highest expected information gain. For example, the most efficient first question targets a superordinate category (e.g., “Are animals on planet Apres?”; $IG$ (information gain) = 1.29). Questions that target a basic (e.g., birds) or subordinate (e.g., parrots) category would achieve lower information gain (basic: $IG = 1.16$; subordinate: $IG = 0.75$). If the feedback on this first question is “yes,” the search is over, the solution being “all animals are on planet Apres.” If the feedback is “no,” the most informative follow-up question targets a basic-level plant category (i.e., either trees or flowers, $IG = 1.56$). Targeting the superordinate category plants ($IG = 0.60$) or a subordinate-level plant category (e.g., daisies, $IG = 0.98$) would achieve a lower information gain (see Figure 1, top panel). Targeting any animal category would not provide any information ($IG = 0$). If the feedback is “some,” the most informative follow-up question targets a basic-level animal category (i.e., either fish or birds, $IG = 1.46$). Targeting a subordinate-level animal category (e.g., parrots) would achieve a lower information gain ($IG = .92$). Targeting any plant category would not provide any information ($IG = 0$).
Results

Simulation results for optimal learners. For the question-asking paradigm, the optimal question-asking strategy reached the solution, on average, in 2.85 questions. An analysis of variance (ANOVA) with the number of questions needed to reach the solution as a dependent variable and solution level (3 levels: subordinate, basic, superordinate) as a between-subjects variable found that this number varied with the hierarchical level of the solution (see Figure 2), $F(2, 299) = 60.03$, $p < .001$, $\eta^2 = .288$. A Bonferroni-corrected multiple-comparisons analysis revealed that fewer questions were needed when the solution was at the superordinate level ($M_{\text{superordinate}} = 2.08$ $SE = 0.14$) than at the basic level ($M_{\text{basic}} = 2.80$, $SE = 0.12$, $p < .001$), and that fewer questions were needed when the solution was at the basic level than at the subordinate level ($M_{\text{subordinate}} = 3.35$, $SE = 0.07$, $p < .001$; see Figure 2).

![Figure 2](image)

Figure 2. Study 1: Question-asking. Average number of questions asked before offering the solution, displayed by category level of the solution (i.e., solution condition). Error bars represent 1 SEM in each direction.


Empirical results for human learners. Overall, participants required an average of 4.24 questions to reach the solution. Even adults ($M_{\text{adults}} = 3.36, SE = 0.35$) required significantly more questions than the simulated optimal question asker (2.85), $t(321) = 3.143, p = .002$.

To analyze the overall efficiency of participants’ question asking, we performed a mixed ANOVA with the number of questions asked prior to providing the correct solution as a dependent variable, age group (3: 7-year-olds, 10-year-olds, adults) as a between-subjects variable, and solution level (3: superordinate, basic, subordinate) and trial number (3) as within-subject variables. This analysis revealed a main effect of age group, $F(2, 67) = 5.29, p = .007, \eta^2 = .136$ (see Figure 2). A Bonferroni-corrected multiple-comparisons analysis confirmed that 7-year-olds ($M_{7\text{-year-olds}} = 4.92, SE = 0.34$) asked more questions than adults ($M_{\text{adults}} = 3.36, SE = 0.35, p = .006$). We did not find differences between the number of questions asked by 7- and 10-year-olds ($M_{10\text{-year-olds}} = 4.38, SE = 0.35, p = .807$), nor by 10-year-olds and adults ($p = .126$).

We also found a main effect of solution level, $F(2, 67) = 20.02, p < .001, \eta^2 = .320$. Mirroring the findings for optimal question asking, a Bonferroni-corrected multiple-comparisons analysis found that participants asked fewer questions in the superordinate condition ($M_{\text{superordinate}} = 3.37, SE = 0.26$) than in the basic condition ($M_{\text{basic}} = 4.08, SE = 0.24, p = .038$), and in the basic condition than in the subordinate condition ($M_{\text{subordinate}} = 5.21, SE = 0.31, p < .001$; see Figure 2). We did not find an interaction between age group and solution level ($p = .120$), nor effects of scenario or trial number.

Participants’ initial questions. Although we failed to find a significant interaction between age group and the hierarchical level of the solution, an analysis of participants’ initial questions does suggest that adults were better able to capitalize on hierarchical structure to ask informative questions (see Figure 3). Across the three trials, adults asked a larger number of first questions at the most informative superordinate level ($IG = 1.29; M_{\text{adults}} = 1.87, SE = 0.16$) than did older children ($M_{10\text{-year-olds}} = 1.13, SE = 0.20$), $t(162) = -3.33, p = .001$, who in turn asked a larger number of such questions than younger
children \((M_{7\text{-year-olds}} = 0.41, SE = 0.16), t(145) = -3.02, p = .003\). Symmetrically, adults asked fewer first questions at the least informative subordinate level \((IG = 0.75; M_{\text{adults}} = 0.13, SE = 0.06)\) than older children \((M_{10\text{-year-olds}} = 0.29, SE = 0.10), t(162) = -3.14, p = .002\), who in turn asked fewer such questions than younger children \((M_{7\text{-year-olds}} = 0.69, SE = 0.17), t(145) = -2.14, p = .034\). The number of initial questions at the basic level \((IG = 1.16)\) did not vary significantly across age groups \((ps > .05)\): \(M_{7\text{-year-olds}} = 0.79, SE = 0.18\), \(M_{10\text{-year-olds}} = 0.92, SE = 0.17\), \(M_{\text{adults}} = 0.55, SE = 0.12\).

**Figure 3.** Study 1: Question-asking. Average number of first questions (across the three trials) asked at the superordinate, basic, and subordinate levels by participants in Study 1 (question asking), displayed by age group. Error bars represent 1 SEM in each direction.

**Participants’ performance versus optimal history-matched and random history-matched models.** The analyses presented above demonstrate that children and adults fell short of optimal information search, which represents a potential upper bound on their performance. However, they do not provide a reasonable lower bound for comparison, nor do they reveal how the efficiency of
information search may have varied over the course of inquiry. To assess performance in a more fine-grained way, we compared the expected information gain of participants’ questions at each time point to two additional models: the optimal history-matched model, which provides an upper bound on expected performance, and the random history-matched model, which provides a lower bound.

The optimal history-matched model selects at each step the question that has the highest information gain, considering the current hypothesis space as defined by a given participant’s previous selections. The random history-matched model selects a question at random. This random selection is repeated 10 times at each step, with replacement, to simulate a representative random average, and we considered the average information gain of the 10 randomly selected questions.

A repeated-measures ANOVA with age group (3: 7-year-olds, 10-year-olds, adults) as a between-subjects factor and model (3: participants, optimal history-matched, random history-matched) as a within-subject factor found that participants’ average information gain ($M_{\text{participants}} = 0.83, SE = 0.02$) was higher than the information gain resulting from a random selection ($M_{\text{random}} = 0.50, SE = 0.01$), but lower than that resulting from the optimal history-matched model ($M_{\text{optimal}} = 1.06, SE = 0.02$), $F(2, 134) = 697.97, p < .001, \eta^2 = .91$. This analysis confirms that participants fell short of optimality, but additionally demonstrates that their performance was significantly better than chance.

This approach to quantifying participants’ performance also allowed us to analyze the efficiency of information search at each time point, and to identify unnecessary questions. Figure 4 displays the average information gain as predicted by the optimal history-matched and random history-matched models and the actual information gain of participants’ questions as the task unfolded. The figure suggests two sources of inefficiency. For the first several questions—until the hypothesis space was narrowed down to a single hypothesis—the average information gain associated with participants’ questions fell below that of the optimal history-matched model (see also Table 2) – that is, they asked non-optimal questions. But additionally, some participants continued to ask questions beyond the point
at which only a single hypothesis remained – that is, they asked unnecessary questions. Such questions would necessarily have zero information gain and therefore contributed to overall inefficiency. In fact, 83% of the 7-year-olds (n = 20), 70% of the 10-year-olds (n = 16), and 52% of the adults (n = 12) asked, in at least one trial, more questions than strictly necessary to identify a single hypothesis. A repeated-measures ANOVA revealed an effect of age group on the number of questions asked beyond the point at which a single hypothesis remained, $F(2, 67) = 4.50, p = .015, \eta^2 = .118$. A Bonferroni-corrected multiple-comparisons analysis confirmed that the 7- and 10-year-olds asked on average more additional questions ($M_{7\text{-year-olds}} = 1.28, SE = 0.22$) than adults ($M_{\text{adults}} = 0.32, SE = 0.23, p = .011$). We did not find any differences in the number of additional questions asked by 7- and 10-year-olds ($M_{10\text{-year-olds}} = 0.81, SE = 0.23, p = .448$), nor by 10-year-olds and adults ($p = .396$).

One possibility is that children asked unnecessary questions in the hope of receiving positive feedback, which could itself be rewarding. Counter to this hypothesis, the majority of unnecessary questions in fact received negative feedback. However, positive feedback was more likely for the younger children. A univariate ANOVA revealed an effect of age group on the percentage of unnecessary questions that received positive feedback, $F(2, 55) = 9.18, p < .001, \eta^2 = .272$. A Bonferroni corrected post-hoc analysis confirmed that, on average, a higher percentage of the unnecessary questions asked by 7-year-olds received positive feedback ($M_{7\text{-year-olds}} = 30.00\%, SE = 4.43\%$) as compared to 10-year-olds ($M_{10\text{-year-olds}} = 2.36\%, SE = 7.54\%$) and adults ($M_{\text{adults}} = 1.18\%, SE = 5.78\%)$.

**Sorting task.** Most participants successfully completed the sorting task without help (7-year-olds: 84%; 10-year-olds: 88%; Adults: 97%). With the exception of three younger children (12%), the remaining participants were all able to complete the sorting task with help (i.e., after being prompted at least once by the experimenter to reorganize the objects in a different way). We found no correlation
between performance in the sorting task and performance in the search task, and our results are not affected by excluding from all analyses those participants who could not complete the sorting task.

*Figure 4.* Study 1: Question asking. Average information gain predicted by the optimal history-matched and random history-matched model and actual information gain of participants’ questions across questions asked, displayed by order of question and age group. The numbers over the x-axis represent the number of participants that asked the corresponding number of questions. Error bars represent 1 SEM in each direction.