Cognition 166 (2017) 407–417

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Desirable difficulties during the development of active inquiry skills

George Kachergis [a,*], Marjorie Rhodes [b], Todd Gureckis [b]

[a] *Artificial Intelligence Dept./Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, Netherlands*
[b] *Department of Psychology, New York University, New York, NY, United States*

ARTICLE INFO

Article history:
Received 18 July 2016
Revised 12 May 2017
Accepted 14 May 2017
Available online 13 June 2017

Keywords:
Question asking
Information search
Active inquiry
Hypothesis testing
Scientific reasoning

ABSTRACT

This study explores developmental changes in the ability to ask informative questions, hypothesizing a link between the ability to update beliefs in light of evidence and the ability to ask informative questions. Five- to ten-year-old children played an iPad game asking them to identify a hidden insect. Learners could either ask about individual insects, or make a series of feature queries (e.g., "Does the hidden insect have antenna?") that could more efficiently narrow the hypothesis space. Critically, the task display either helped children integrate evidence with the hypothesis space or required them to perform this operation themselves. Our prediction was that assisting children with belief updating would help them formulate more informative queries. This assistance improved some aspects of children's active inquiry behavior; however, despite making some updating mistakes, children required to update their own beliefs asked questions that were more context-sensitive and thus informative. The results show how making a task more difficult can improve some aspects of children's active inquiry skills, thus illustrating a type of "desirable difficulty" for reasoning.

1. Introduction

A skill of central importance during development is learning how to ask informative questions in order to make sense of the world. The roots of these abilities are observable even in the early preschool years. For example, in simple causal reasoning tasks, preschool-aged children can distinguish confounded from unconfounded evidence to draw causal inferences (Gopnik, Sobel, Schulz, & Glymour, 2001; Kushnir & Gopnik, 2005, 2007; Schulz & Gopnik, 2004). Preschool-aged children also selectively explore confounded evidence in their own exploratory play (Cook, Goodman, & Schulz, 2011; Gweon & Schulz, 2008; Schulz & Bonawitz, 2007). Despite these early emerging abilities, many of the cognitive skills required for self-guided, active inquiry seem to follow protracted developmental trajectories. For example, in tasks designed to assess scientific reasoning abilities, children in the older elementary school years (ages 8–10) often have difficulty adopting systematic strategies, such as testing the effects of one variable at a time or selecting interventions that will lead to determinate evidence (Chen & Klahr, 1999). Although children in the older elementary school years can be taught to engage in these strategies via direct instruction (Klahr & Nigam, 2004; Kuhn & Dean, 2005), it is notable how difficult it is for them to discover and implement them on their own.

One reason for the difficulties children exhibit in these types of inquiry tasks may be that active inquiry depends on the coordination of a variety of component cognitive processes (Bonawitz & Griffiths, 2010; Coenen & Gureckis, 2015). For example, according to one popular view (Klein, Moon, & Hoffman, 2006a, 2006b; Russell, Stefik, Pirolli, & Card, 1993), active inquiry unfolds as a sequence of mental steps (see Fig. 1). Learners must generate possible hypotheses to explain their environment. They then must engage in decision making to ask questions or gather additional information to decide which of these hypotheses is most likely. They then must understand the results of these inquiry behaviors and update their beliefs accordingly, and so on. The various stages of this loop closely mirror the process of scientific reasoning engaged by scientists (Klein et al., 2006a, Klein, Moon, & Hoffman, 2006b; Russell et al., 1993). Inefficiencies in any or all of these interrelated processes may serve as developmental limitations. For example, young learners may be able to search efficiently for information given a particular set of hypotheses but have trouble updating their beliefs correctly given new evidence. In this sense active inquiry behavior is like a bicycle: when all the

* Corresponding author.
E-mail addresses: g.kachergis@donders.ru.nl (G. Kachergis), marjorie.rhodes@nyu.edu (M. Rhodes), todd.gureckis@nyu.edu (T. Gureckis).

http://dx.doi.org/10.1016/j.cognition.2017.05.021
0010-0277/

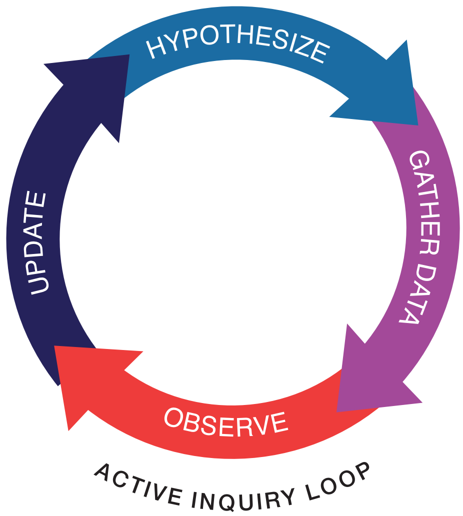


Fig. 1. The active sensemaking loop depicts the successive cognitive process that are engaged when attempting to derive a meaningful understanding of an initially ambiguous situation. The stages of the loop closely mirror the process of scientific reasoning engaged by scientists. However, a similar set of inductive processes are at play in many real-world situations (e.g., working an unfamiliar ATM machine, reading a complex nutrition label). Aspects of the loop are directly related to Bayesian models of learning and information gathering (Bonawitz & Griffiths, 2010; Gureckis & Markant, 2009).

elements are properly functioning and aligned the bike moves forward. However, misalignment of even one component can be catastrophic.

Understanding the integrated nature of these cognitive processes is important not just for our scientific understanding of the development of the human mind, but also because of broader educational implications. For example, many educational philosophies emphasize relatively unstructured, self-guided learning environments (Bruner, 1961; Kolb, 1984; Steffe & Gale, 1995). Understanding limitations in children's active inquiry abilities and how each component of such abilities evolves across age can be used to design more effective learning environments for children of various ages. For example, evidence that younger children benefit from assistance in updating their beliefs in response to new evidence would suggest that learning environments for younger children need to provide support for this component of their learning.

The present study attempts to decompose the component processes involved in active inquiry, specifically focusing on the role of belief updating. We tasked five- to ten-year old children to identify a hidden insect in a simple iPad variant of the classic "Guess Who?" game. Children sequentially asked questions to try to identify the hidden target and received truthful answers. Based on prior work reviewed below (e.g., Mosher & Hornsby, 1966), we expected younger children to have difficulty formulating informative queries and thus sought to explore what types of automated assistance might aid children's reasoning strategies. Specifically, we manipulated whether the computer program helped children to use the new evidence that resulted from their queries to narrow down the hypothesis space, or whether they had to reconcile the revealed evidence and the hypothesis space on their own. Our expectation was that helping children to update their beliefs accurately following the receipt of new information would free up cognitive resources and lead to higher quality question-asking. Interestingly, our results opposed this initial hypothesis in that elements which ostensibly made our task more difficult actually improved the quality of children's inquiry behavior and suggest an important refinement of the information processing model summarized in Fig. 1.

1.1. Developmental change in the ability to ask revealing questions

Active inquiry fundamentally depends on the ability of learners to construct actions or queries which gain information (e.g., asking a question of a knowledgeable adult). A now classic way to study this behavior is through experimental tasks based on the 20-questions or 'Guess Who?" game. In the game, the asker (participant) tries to determine a hidden object known only to the answerer (experimenter) by asking a series of yes-or-no questions. Mosher and Hornsby (1966) identified two broad question types commonly used in the game: *hypothesis-scanning* questions test a single hypothesis or specific instance (e.g., "Is it a monkey?"), whereas *constraint-seeking* questions attempt to constrain the hypothesis space faster by querying features that are present or absent in multiple objects (e.g., "Is it soft?"), but that do not directly identify the answer except by virtue of elimination.

A classic finding in this literature is that younger children (e.g., aged 6) tend to ask more hypothesis-scanning questions, while older children (e.g., aged 11) use more constraint-seeking questions, and also tend to find the answer after fewer questions (Mosher & Hornsby, 1966). One explanation is that only older children have developed the ability to focus on the high-level features that group the hypotheses, whereas younger children focus on individual stimuli. Consistent with this viewpoint, manipulations that help children focus on these higher-level features, such as cuing them with basic level category labels instead of exemplar names (Ruggeri & Feufel, 2015), increase the likelihood that young children will generate constraint-seeking questions (see also Herwig, 1982). Further, although young children are often relatively less likely than older children to ask constraint-seeking questions, even younger children (ages 7–9) are more likely to do so when such questions are particularly informative, such as when the hypothesis space is large and there are several equally probable solutions remaining (Ruggeri & Lombrozo, 2014, 2015). These results reinforce the viewpoint described above: having the right set of hypotheses in mind, or being primed with the right level of category information seems to drive more efficient information search.

The behavioral distinction between constraint-seeking and hypothesis-scanning questions can also be studied from the perspective of normative models (Oaksford & Chater, 1994; Nelson, 2005; Tsividis, Gershman, Tenenbaum, & Schulz, 2013). These models attempt to objectively define the "quality" of a question and to see how people's choices compare (see below for a larger discussion). A number of recent studies have explored how children's question asking compared to such models. For example, Nelson, Divjak, Gudmundsdottir, Martignon, and Meder (2014) found that 8–10 year-old children can search a familiar structured domain (people with varying gender, hair color, etc.) fairly efficiently, tending to ask about frequent real-world features that roughly bisected the search space (e.g., gender first). Likewise, Ruggeri, Lombrozo, Griffiths, and Xu (2015) found that children's patterns of search decisions were well-explained in terms of expected information gain (EIG), one popular model from this class which is described below. Perhaps most importantly, these models are highly context sensitive. Rather than arguing that either constraint-seeking or hypothesis-scanning questions are universally "better," these models take into account the current context including the learner's prior belief and the past evidence that has been revealed. This allows much more fine grained predictions. For example, on a given trial a hypothesis-scanning question might be equally informative compared to a constraint-seeking question (e.g., when only two hypotheses remain). In our study we will analyze children's question asking with respect to these models to allow an objective measurement of the quality of their information seeking behavior.

1.2. Belief updating and active inquiry

While it is clear that there are developmental changes in how children formulate questions, less work has considered developmental changes in how children make use of the new evidence that their questions reveal (but see Denison, Reed, & Xu, 2013). However, there are many reasons to think that these two behaviors might be deeply entwined. The active inquiry loop in Fig. 1 suggests one obvious interaction because if questions or information gathering actions are made on the basis of current beliefs, and those beliefs are wrong, then a query may not have the expected effects (c.f., research on the hot stove effect, Denrell & March, 2001; Rich & Gureckis, 2015). There are certainly many examples where scientific progress has been derailed by incorrect interpretation of evidence, as in the case of experiments thought to support the theory of spontaneous generation of life (Needham, 1745).

Coenen and Gureckis (2015) describe a more fundamental reason for why belief updating and information search might be related. In particular, they focus on a popular computational model of active inquiry called Expected Information Gain (EIG). As mentioned above, this model has been widely used in both the adult and developmental literature to understand how people decide between different queries (Oaksford & Chater, 1994; Coenen, Rehder, & Gureckis, 2014; Gureckis & Markant, 2009; Nelson, 2005; Nelson et al., 2014; Markant & Gureckis, 2012; Ruggeri et al., 2015; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Intuitively, EIG evaluates the quality of a question by considering how much is expected to be learned from each possible answer to that question. For example, in the constrained 20-questions game "Guess Who?", a child might ask "Does your character have a hat?" or "Is your character male?". To decide between these two queries EIG considers each possible answer ("yes" or "no" for each) and how much each answer would alter the learner's current beliefs given the question. If all the remaining characters in the game were wearing hats then the answerer would never respond "no" to the hat question, and the received "yes" would not normatively alter the learner's beliefs; no information would be gained by asking about hats. Even if one of the dozen remaining characters had a hat, asking about hats would have low EIG, since it would be unsurprising that the answer is "no"–only in one of twelve possible worlds does the hidden character happen to be wearing a hat, while in 11 of 12 worlds the character is not. In contrast, if half the remaining characters were male and half were female, then either answer to the gender question would strongly shift what the learner knows, eliminating half of the candidates (either the males, or the females). Thus, the more valuable question according to EIG would be "Is your character male?". In this model, belief updating is fundamental to judging the information quality of a possible query: it is only by imagining how one's beliefs would change given different answers that a question derives meaning and value. On the basis of this observation, Coenen and Gureckis (2015) reported a study aiming to relate individual differences in belief updating during a causal reasoning task to patterns of information seeking behaviors. Subjects that showed clear evidence of biased belief updating (e.g., incorrectly interpreting ambiguous evidence as unambiguous) also showed biased patterns of information gathering in a causal intervention learning task. This study highlights the strongly interactive nature of belief-updating and information seeking behaviors.

Interestingly, past work on the development of question asking abilities in children has tended not to emphasize belief updating as a dependent measure, or precluded studying updating beliefs by the design of the study. For example, Herwig (1982) presented children with a series of two-alternative forced choice decisions between hypothesis-scanning or constraint-seeking question but did not actually give feedback (and therefore could not detect errors in belief updating). In the 20-questions task of Nelson et al. (2014), 8- to 10-year-olds were asked to identify which of 18 people was the hidden target, and played the game to completion several times for different targets. Children eliminated hypotheses (flipping over cards) based on acquired evidence, but were given help by the experimenter if needed, which presumably means they were not allowed to make errors. In Ruggeri and Lombrozo (2014), the experimenters did not explicitly represent the hypothesis space for participants in Experiment 1's causal reasoning task (e.g., "Why was a man late to work yesterday?"), and when ten explicit reasons for being late were given in Experiment 2, they remained in view. That is to say, the process of hypothesis updating was not scrutinized in these prior studies.

In the present study, we hypothesize that biases in the way children search for information (e.g., by favoring hypothesis scanning questions over constraint seeking questions) may stem from difficulties in coordinating the belief updating and search process. There are a variety of specific reasons for this prediction. First, although the components of the sensemaking model described in Fig. 1 above are sequential, they likely rely on a common pool of cognitive and attentional resources, and are thus not completely independent. At a minimum, learners have constant and limited capacities for working memory and reasoning during the task, and may come to avoid strategies that tax these resources if they run into difficulty during the course of the experiment. In this case we hypothesize that the cognitive load from planning questions, or from updating beliefs, may impair performance on either task. Second, hypothesis scanning questions might be easier for young children in that they produce evidence that applies to a single hypothesis. If instead children ask constraint-seeking questions, they must eliminate from the hypothesis space any possibilities that are ruled out by the new information. This process could be cognitively taxing, and also prone to errors. Thus, although constraint-seeking questions are often more informative in theory, we posit that they might not always be so to young children, particularly if children have difficulty using the obtained information to update their representation of the hypothesis space accurately.

To test this hypothesis, in the present study we manipulated whether children received assistance in integrating evidence with the hypothesis space or had to undertake this process on their own. Our expectation was that aiding children in coordinating evidence and beliefs would enable more sophisticated, and informative, inquiry behavior. To evaluate this prediction we evaluated the quality of children's question asking ability against an objective standard of informativeness given by the EIG model described in more detail below. We additionally analyze our data specifically in terms of constraint-seeking and hypothesis-scanning questions. Our central prediction was that assistance in belief updating should increase the relative EIG of children's questions and the relative utilization of constraint-seeking questions. Given that older children (8–10 years) have previously been found to use more constraint-seeking questions than younger children (5–7 years), we tested across these two age groups, expecting that younger children would benefit more from the assistance in hypothesis updating than would older children.

2. Experiment

The purpose of the experiment is to investigate how children utilize hypothesis- scanning and constraint-seeking questions when trying to discover a hidden object. To that end we created a tablet-based game based on the popular "Guess Who?" paradigm. The study was conducted in the context of a children's science museum and the materials and design of the study were selected to integrate with museum content. Our hope was that insights from the study might be used to help museum curators design more effective educational exhibits that target children of different ages. For example, if updating the hypothesis space is difficult for younger children, exhibits for this age group assist them in updating, and perhaps even attempt to teach them the process.

2.1. Methods

2.1.1. Participants

Participants in this experiment were 134 children between the ages of 5 and 10 years old who were recruited at the American Museum of Natural History's Discovery Room. Of the 134 children recruited (67 per condition), we analyze the data from 121 children (21 5-year-olds, 20 6-year-olds, 22 7-year-olds, 20 8-year-olds, 20 9-year-olds, and 18 10-year-olds) who completed 5 or more rounds of the game, understood the instructions, and were not distracted (e.g., by other children or their parents). Participants were assigned in counterbalanced order to either the automatic-update condition or the manual-update condition (automatic: 32 5–7 year-olds and 29 8–10 year-olds; manual: 31 5–7 year-olds and 29 8–10 year-olds).

2.1.2. Stimuli

On each round, children were presented with a display containing sixteen insects. One of the insects was randomly selected to be the target which children attempted to identify by asking questions. The sixteen insects within a round shared the same body shape but were composed of varying perceptual features. In particular, insects were defined by the presence or absence of 9 features: green body, orange eyes, antennae, big spots, tiny spots, legs, leaves, water droplets, and blue "fur". Fig. 2 shows an example of two of the body shapes used, each with all of the binary features present. Across rounds the body shapes (selected from a pool of 16 unique body shapes) varied randomly but within a round the body shape was shared between all sixteen items. The insect task was designed to fit thematically with the content of the AMNH Discovery Room activities which emphasize the often subtle differences between species of animals (specifically, many interactive exhibits involve insects).

2.1.3. Design

Across the sixteen exemplar insects (A-P in Table 1) some perceptual features were more frequent than others (e.g., F1 one was present on eight of the insects while feature F9 was present on only two insects). The features (F1-F9) in this semi-hierarchy were randomly assigned to the visual features for each child (i.e., F1 might be mapped to color (1 = green, 0 = white) for one child, while F1 is mapped to leaf-eating (1 = leaf, 0 = no leaf) for another). This mapping remained consistent across rounds, meaning that even if children did not immediately discern the distribution of features, they may have been able to learn it gradually across the rounds. The design of the abstract structure introduced strong differences in the informational utility of each feature (F1-F9). For example, given no other information, it would be quite informative to ask about feature F1 because it is shared with half of the possible insects. In contrast, feature F9 is less informative on the first trial of each round because most of the insects do no have this feature. Note that feature F10 depicted a particular body shape, and was not relevant to query since body shape was constant across all insects in a round.

Each of these features was visually represented on a button (see Fig. 3), available for children to tap with their finger. An additional feature button (F10) depicting a particular body shape was always present but not relevant to the insects on display since they always shared the same body shape. A tap on a feature button is effectively a "constraint-seeking" question. Instead of choosing a feature button, children could at any time query an exemplar by tapping it to determine if it was the hidden insect or not. This choice is equivalent to a "hypothesis-scanning" query. The interactive dynamics of the display varied across conditions. After making a feature query in the manual-update condition, children must select which insects (i.e., hypotheses) are consistent with the feedback. In contrast, in the automatic-update condition the hypothesis space automatically updated to be consistent with the feedback received.

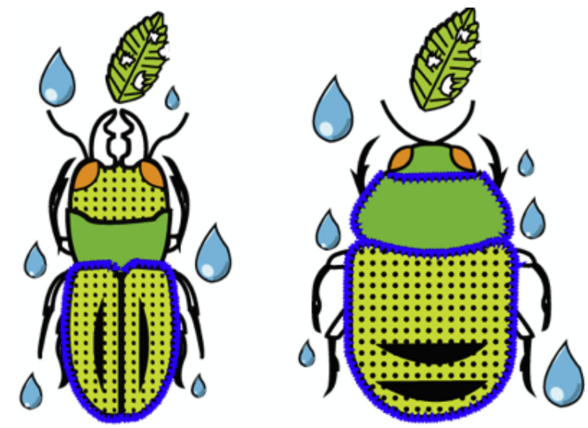


Fig. 2. Examples of two insect body types with all 9 of the binary features present. Each round used one of 16 possible body shapes.

Table 1
The abstract feature structure of the 16 insects (labeled A-P) used in each round. A value of 1 means the feature was present for this insect while a value of 0 means the feature was absent. Abstract features F1-F9 were randomly assigned to the binary (present or absent) visual features for each participant, with a consistent assignment used from round to round. For example, if feature F9 was color (green versus white), then all the 16 bugs might be white except items C and I which would be filled in with a green body. Both the identity of the features and the meaning of the 0s and 1s in the table were randomly determined for each child. F10 was always assigned to a button for body shape, which was shared by all exemplars.

Exemplar	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
A	**1**	0	0	**1**	0	0	0	0	0	0
B	**1**	**1**	0	**1**	0	0	0	0	0	0
C	0	**1**	0	**1**	0	0	0	0	**1**	0
D	0	**1**	0	**1**	0	0	0	0	0	0
E	**1**	0	0	0	0	0	0	0	0	0
F	**1**	**1**	0	0	0	0	0	0	0	0
G	0	**1**	0	0	0	0	**1**	**1**	0	0
H	0	**1**	0	0	0	0	**1**	0	0	0
I	**1**	0	**1**	0	0	**1**	0	0	**1**	0
J	**1**	0	**1**	0	0	**1**	0	0	0	0
K	0	0	**1**	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0
M	**1**	0	**1**	0	**1**	0	0	**1**	0	0
N	**1**	0	**1**	0	**1**	0	0	0	0	0
O	0	0	**1**	0	**1**	0	0	0	0	0
P	0	0	0	0	**1**	0	0	0	0	0

2.1.4. Procedure

After being trained by an experimenter on a simpler version of the task with unrelated stimuli[1] (a dog searching dog houses) so that they understood how to query exemplars and features, and how to eliminate hypotheses, children played 5 or more rounds of the iPad game asking them to identify which one of 16 insects was

[1] Download full task code, data, instruction, and analysis scripts: https://github.com/kachergis/bugguess.

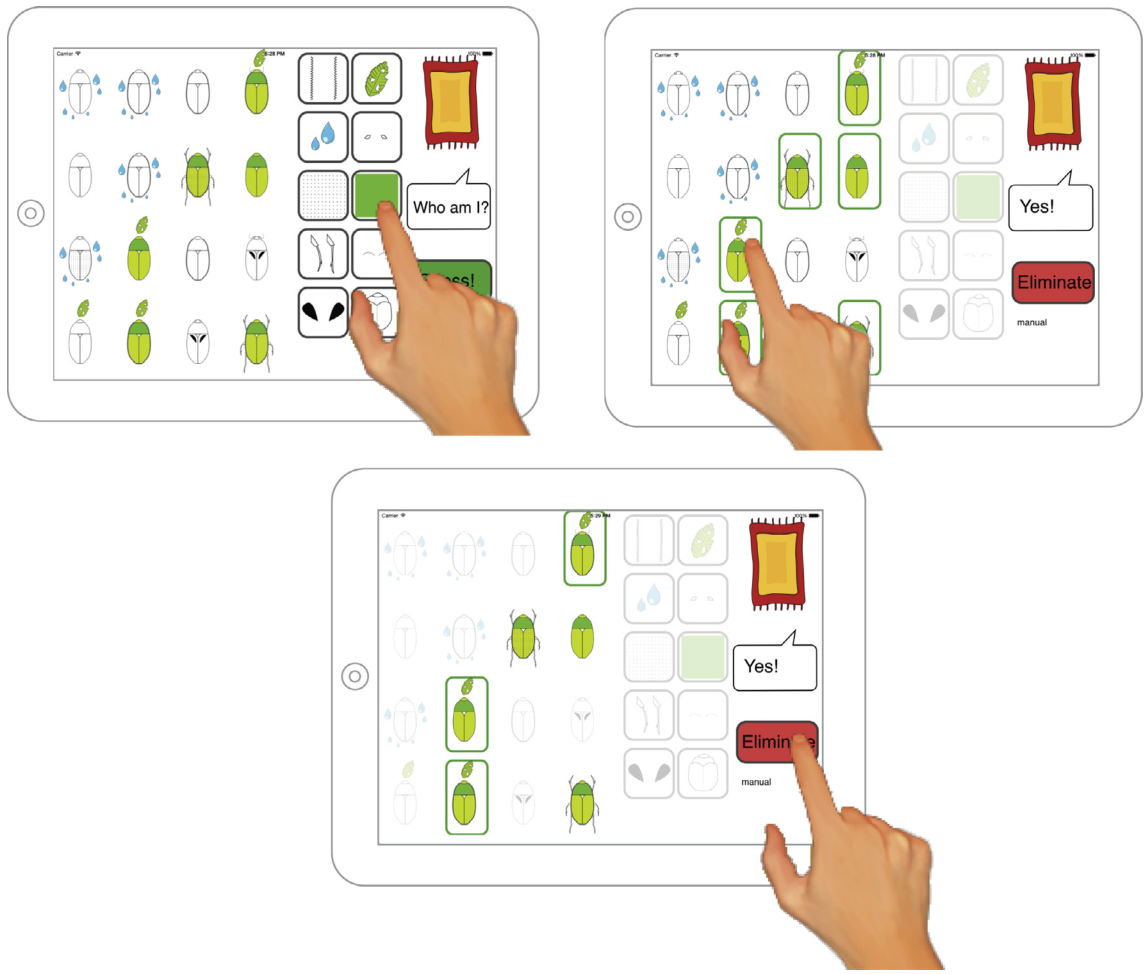


Fig. 3. Task overview: in the upper left, a feature button is used, asking if the insect hidden under the rug is green. Given feedback ("Yes!"), participants in the manual update condition select the insects that are consistent with this new information (upper right), whereas in the automatic condition the consistent insects are selected by the game. Players in both conditions press the red button to return to the button phase, and again either choose a feature button or query a single insect. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hidden under a cartoon rug (see Fig. 3). The task alternated between the query phase and the elimination phase. In the query phase, players could either query an individual insect by tapping one (equivalent to asking, "Is this the hidden bug?"), or choose to use a feature query button (e.g., the green button asks "Is the hidden bug green?") to find out whether the hidden insect had a particular feature.

If a single exemplar was tapped on (i.e., a hypothesis-scanning query), and the item was the experimenter-determined hidden insect, a smiley face appeared and the round was completed. If the tapped exemplar was not the hidden insect, a red "X" was shown on top of the tapped insect and the insect became grayed out (i.e., eliminated).

After a feature query (i.e., constraint-seeking query), the insect under the rug gave feedback, saying "Yes!" (indicating it had the feature; narrated by the experimenter), or "No!" (if it did not have the feature). This was followed by the elimination phase, during which insects that were inconsistent with the feedback were eliminated, and the hypothesis space was thus narrowed. The elimination phase varied based on condition. In the automatic-update condition, after the feedback from a feature query, subjects merely pressed the "Eliminate" button, all the no longer relevant insects were eliminated (grayed out), and the game returned to the query phase. In the manual-update condition, after a subject made a feature query and saw feedback, they had to select each insect that was consistent with the feedback for that feature, as shown in the top right of Fig. 3. Insects were selected (denoted by a green box) by tapping, and could be deselected by tapping again. Only when children verified they were done selecting insects did the experimenter press the "Eliminate" button, which eliminated any insects that were not selected.

Before children were allowed to begin, the experimenter explained a random selection of at least three of the feature buttons (more if the child asked), and asked children to point to an exemplar exhibiting each of the explained features. In the manual-update condition it was possible for mistakes to be made during the elimination phase, as the software did not aid in updating the hypothesis space. Insects that should have been eliminated but were kept (a 'miss') continued to be visible options. Insects that were consistent with the query but wrongly eliminated (a 'false alarm') were grayed out. Our analyses below take into account the role that such errors may have played in the manual-update condition. In the event that the hidden insect was wrongly eliminated during a manual-update error, the round was played out until all of the insect/hypotheses were grayed out. The experimenter would then indicate that the insect must have been mistakenly eliminated (but not at what point), and would end the round by clicking the grayed-out exemplars until the hidden one was found. These final clicks (beyond when all hypotheses were eliminated) were not included in the analysis.

At the beginning of each round, the experimenter would say, "Let's try to find which insect is hiding pretty quickly, so we can do more!" Thus, the task mostly relied on intrinsic motivation to solve the puzzle quickly, providing no explicit cost incentive to be efficient. This was chosen primarily due to the difficulty of rewarding children in the museum. Children were welcome to complete more than five rounds, if they desired to: after the fifth and each successive round, they were asked, "Do you want to play again?".

2.2. Results

2.2.1. Overall

We analyzed only the first 10 rounds[2] from each child (only 8 children played more than 10 rounds, including one who played 51 rounds). This covers 722 rounds from the 121 children who understood the instructions and completed a minimum of five rounds (61 in the automatic condition and 60 in manual). The mean number of total queries (feature and exemplar) taken to complete a round was 6.07 in the automatic-update condition, and 5.08 in the manual-update condition. Based on bootstrapped means, the 95% confidence intervals (CIs) for these distributions did not overlap (bias-corrected and accelerated (BCA) 95% CIs for manual condition: (4.75, 5.44), for automatic condition: (5.77, 6.37)), with the manual condition taking fewer queries. However, we note that in the manual condition, before removing the queries following mistakes in which the correct answer had been eliminated, rounds took an average of 7.36 queries (95% CIs: (6.91, 7.85)) to complete–more than the automatic condition. For comparison, we simulated 700 rounds of the game with an agent that queried randomly in the task, choosing uniformly at random on the first query from 16 exemplars and 10 feature buttons, and continuing with whatever stimuli (and feature queries) remain after each query, while making no update errors. This random agent took on average 8.89 queries (median: 9) to complete a round, with bootstrapped 95% BCA CIs (8.58, 9.20) higher than the CIs of either condition. This simple baseline model provides an important baseline for chance in this complex task, since each feature query can eliminate a variable quantity of exemplars, depending on the features of the remaining hypotheses. Thus, each successive (random) click can have a large effect on both the number of remaining exemplars, and on the relevance of the remaining feature queries. The fact that the random agent takes significantly more queries than participants in either condition suggests at least some structure in children's active inquiry behavior, further investigated below.

In the remaining analyses, following Mosher and Hornsby's (1966) finding of a transition from hypothesis-scanning to constraint-seeking questions from 7- to 8-year-olds, we group the younger (5–7 year-olds) and older (8–10 year-olds) children. The same analyses have been performed using age as a continuous variable and the same significant effects were found in all cases.

2.2.2. Qualitative querying behavior

Participants' mean number of queries per round were subjected to an ANOVA with update condition (automatic vs. manual) and age group (5–7 vs. 8–10) as between-subjects factors and query type as a within-subject factor.[3] This analysis indicated a significant main effect of age group ($F(1,223) = 7.82$, $p < .01$), and no significant main effect of condition ($F(1,223) = 0.29$, $p = .59$) or query type ($F(1,229) = 0.93$, $p = .33$). Overall, older children required fewer queries of either type to complete a round, also evidenced by a significant negative correlation between participants' mean queries to complete a round and age (in years: 5–10; $t(119) = 2.39, p = .02, r = -.21$). There were significant interactions of condition and query type ($F(1,223) = 12.72$, $p < .001$), and age group and query type ($F(1,223) = 9.75$, $p < .001$), detailed below. No other interactions were significant (all F-values < 1). In comparison to the manual condition, there were fewer exemplar queries in the automatic condition ($M_{man} = 4.16$, $M_{auto} = 2.99$, $t(100.5) = 2.97, p < .01$), while there were fewer feature queries in the manual condition ($M_{auto} = 3.88$) ($M_{man} = 3.01, t(85.8) = 2.41$, $p < .05$). The participants' feature query rates in both conditions were lower than the simulated random rounds' mean number of feature queries (5.41, bootstrapped 95% CIs = (5.19, 5.64)), but above the optimal. The participants' number of exemplar queries in the manual round were similar to the simulated agents (4.02, bootstrapped 95% CIs = (3.86, 4.19)), but lower in the automatic condition.[4]

Fig. 4a shows the average number of query types used per round for participants by age group. Both age groups in the manual-update condition used more exemplar queries than feature queries, and older participants in both conditions use fewer exemplar queries than younger participants ($M_{5-7} = 4.28$, $M_{8-10} = 2.64, t(113) = 3.64, p < .001$). Older participants used a greater proportion of feature queries than younger participants in the automatic condition ($M_{5-7} = .49$ vs. $M_{8-10} = .66$, $t(57.8) = 3.31, p < .01$), but there was no significant difference in the manual condition ($M_{5-7} = .44$ vs. $M_{8-10} = .53, t(49.3) = 1.31$, $p = .20$). Thus, the automatic condition replicates the Mosher and Hornsby (1966) finding that older children use a greater proportion of constraint-seeking questions, but this finding is not reliably found in the manual condition alone.

The finding of more feature queries in the automatic condition and more exemplar queries in the manual condition raises a number of questions about when and why participants are choosing particular queries in each condition. We next investigate response times to reveal how much thought participants are putting into making each type of query.

2.2.3. Response times

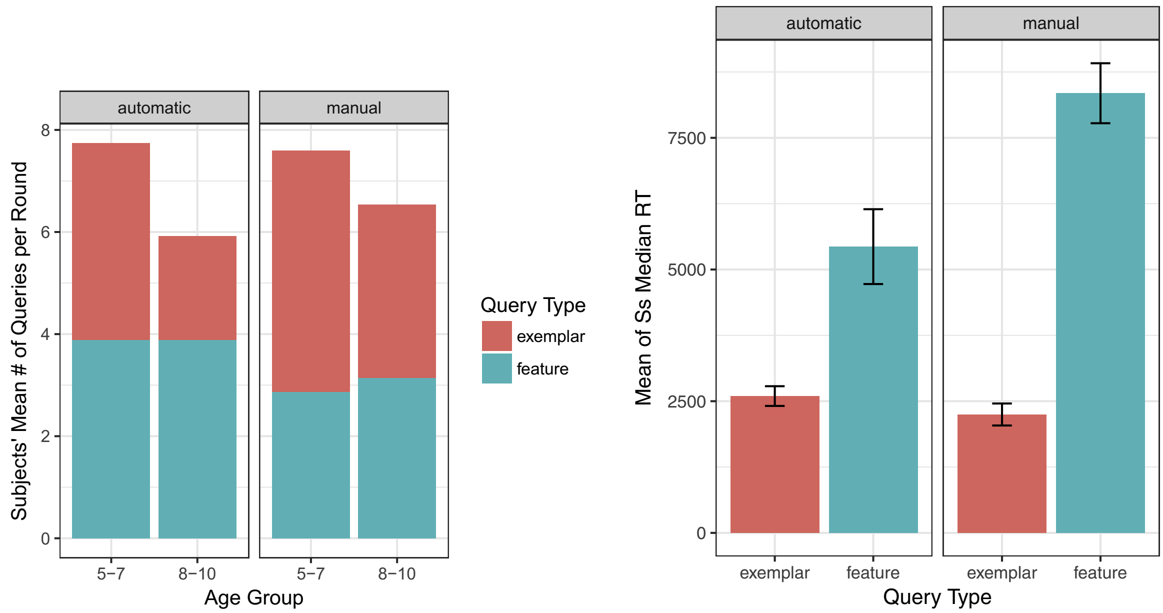
Participants' median[5] RT for each button type (feature and exemplar) was computed and these data were subjected to an ANOVA with condition (automatic, manual) and age group (5–7, 8–10) as between-subjects factors and button type as a within-subject factor. There were significant main effects of button type ($F(1,229) = 26.33$, $p < .001$) and condition ($F(1,229) = 7.36$, $p < .01$), but not a significant main effect of age group ($F(1,229) = 1.89$, $p = 0.17$). On average, participants took longer to make queries in the manual condition (5219 ms) than in the automatic condition (4016 ms). Overall, participants took much longer to make feature queries (6841 ms) than to make an exemplar query (2424 ms), perhaps indicating more thought before making the more complex queries (i.e., feature queries, as they may pertain to multiple exemplars). There was also a significant interaction effect of query type and condition ($F(1,229) = 11.81$, $p < .001$). Fig. 4b shows the mean of subjects' median RTs for each query type, split by condition. Feature queries were slower in the manual-update condition (8347 ms vs. 5435 ms in automatic),

[2] This threshold was chosen after seeing the distribution of rounds played to limit undue influence on by-round analysis of the few children who chose to play more than 10 rounds (whose behavior may be expected to change in later rounds, and who may be different than the majority of children who played < 10 rounds), without throwing out too much of the data (92% remains).

[3] Behavior across rounds was investigated for evidence of learning, but no consistent changes in behavior were evident.

[4] Note that although there are at first more exemplars (16) than feature buttons (10), after the first query or two there will likely be few exemplars remaining to query, which is why the expected number of exemplar queries is lower than the expected number of feature queries in the simulation.

[5] Response times are right-skewed, so medians are a less biased measure of central tendency.



(a) Mean number of queries of each type per round split by age group and condition.

(b) Mean of participants' median RT by condition and query type. Bars show +/-1SE.

Fig. 4. Mean query usage per round (a) shows that older children use fewer exemplar queries than younger children. Manual-update participants used more exemplar queries and fewer feature queries than automatic-update participants. Response times (b) show that exemplar queries were faster than feature queries, which represent a more complex strategy and thus likely required more thought. Feature queries were particularly slower in the manual-update condition.

which could indicate (1) more careful thought given to features in this condition, and/or (2) general hesitance to use feature queries, perhaps because it is time-consuming (even difficult) to manually update hypotheses. Exemplar queries, on the other hand, were at least as fast in the manual-update condition as in the automatic-update condition (2248 ms vs. automatic: 2597 ms). Other interactions were not significant (all F-values < 1).

In summary, it is clear that the manual-update condition results in fewer feature queries and more reliance on exemplar queries. Manual-update participants may be reluctant to use feature queries for at least two reasons: (1) it demands more time and cognitive effort to manually update the hypothesis space after a feature query than in the automatic-update condition, and (2) the manual update process is error-prone, and any mistakes may in turn lead to more exemplar queries in order to recover.[6] Therefore we proceed to investigate errors in manual updating.

2.2.4. Manual update mistakes

The manual-update condition allows participants to commit two types of error during hypothesis updating: a miss is defined as a failure to eliminate a insect, and a false alarm is a failure to keep a hypothesis that was consistent with the query. Note that a miss is an error of commission–i.e., the insect had to be tapped to be kept–whereas a false alarm is an error of omission (i.e., failing to tap a insect), and thus we expect more of the latter. Comparing the manual-update subjects' mean number of errors of each type per round, indeed there were more false alarms ($M = 6.9$, sd = 1.9) than misses ($M = 1.8$, sd = 1.3; paired $t(58) = 19.8, p < .001$). A MANCOVA to determine if error rates were related to age did not find a significant effect for either misses ($F(1,56) = 0.77$, $p > .05$) or false alarms ($F(1,56) = 0.23$, $p > .05$). Consistent with our hypothesis that manual updating increases cognitive load and reduces information seeking behavior, fewer feature queries and more exemplar queries were made in the manual condition. However, RT analyses also indicated that feature queries took longer under manual updating. One possibility is that feature queries were more carefully considered in this condition than under the ease of automatic updating. To evaluate this idea, we conducted a model-based analysis of children's feature queries which provides a context-sensitive measure of query informativeness.

2.2.5. Expected information gain

Each successive query reduces the size of the remaining hypothesis space to some degree: on the first move, querying the appropriate feature (F1) can cut the space in half. When two hypotheses remain, even an exemplar query will cut the space in half. As a result, the distinction between constraint-seeking and hypothesis scanning queries is not absolute (either could be better in different circumstances). As described in the Introduction, one way to analyze the contextual sensitivity of participants' queries is to calculate the Expected Information Gain (EIG) of the query they made.

We first introduce key terms used to define EIG. Entropy measures uncertainty about the outcome of a random variable X and is denoted $H(X)$. Entropy is 0 when there is only one possible outcome, and maximal when all possible outcomes are equiprobable (i.e., a uniform distribution).

$$H(X) = -\sum_{x} p(x) \cdot \log_2(p(x)) \tag{1}$$

Mutual information gain, $I(X;Y)$, measures the change in entropy as we receive a new piece of information Y, i.e., how much does our uncertainty about X change given that we know Y?

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

[6] If the correct answer is mistakenly eliminated, exemplar queries are needed to find it and finish the round. These additional exemplar queries (tapping on grayed-out bugs) were excluded from analysis, and were necessary in 86 out of 364 manual rounds.

The Expected Information Gain (EIG) of a query Q is the weighted average of the information possible from each possible answer to the query, weighted by the current probability of receiving that answer.

$$EIG(Q) = -\sum_{Y} p(Y|Q)I(X;Y) \quad (3)$$

This will be 0 (or near-0) for queries that can be expected to eliminate none or just one or two hypotheses in a large space, and more positive for queries that are likely to eliminate a larger number of hypotheses. In this task, EIG is maximal (1) for a feature query that will eliminate half the remaining hypotheses. Such a query is always available at the beginning of any round (feature F1), and due to the partially-nested feature structure used (see Table 1), maximal EIG queries are often available at other stages of the round. Note that maximizing EIG would result in the same choices as maximizing the expected number of deleted hypotheses, taking into account the number eliminated by both possible outcomes of the query, and the likelihood of each outcome. Due to the semi-hierarchical distribution of features, there is often a single feature with near-maximal EIG, while once a feature query is made, some other feature query will now have near-minimal EIG.

We analyze the EIG for each participant's feature queries separately, as well as in aggregate with the exemplar queries. [7] Participants' mean feature query EIG was subjected to an ANOVA with condition and age group (5–7 vs. 8–10) as between-subjects factors. This ANOVA indicated significant main effects of condition (F(1,115) = 55.03, $p < .001$) and age group (F(1,115) = 12.42, $p < .001$), with no significant interaction (F(1,115) = 0.20, $p = .66$).[8] The same ANOVA applied to participants' mean EIG of all queries indicated a significant main effect of condition (F(1,116) = 25.11, $p < .001$) and a significant main effect of age group (F(1,116) = 21.43, $p < .001$), with no significant interaction (F(1,116) = 1.02, $p = .31$). Fig. 5 shows mean EIG per feature query (a) and for all queries (b) by age group and condition, along with a baseline showing the mean EIG of all the remaining feature queries (i.e., as if each subject had chosen randomly from the feature queries available at any given point). Note that although randomly-chosen features for the manual-update subjects have a slightly higher EIG than for automatic-update subjects (driven in part by update errors quickly reducing the hypothesis space), the baseline random EIGs are far below the corresponding human data. Feature queries made by 8–10 year-olds had significantly higher EIG than those made by 5–7 year-olds ($M_{8-10} = .71$, $M_{5-7} = .63$, $t(117) = 3.06$, $p = .003$), showing that older children tended to use more relevant feature queries. The feature queries made by participants in the automatic condition had significantly lower EIG than those made in the manual condition ($M_{auto} = .59$, $M_{man} = .75$, $t(112.7) = 7.15$, $p < .001$). To verify this finding, we examined in what proportion of feature queries participants in each condition chose the most informative feature query, in terms of the actual EIG for the current hypothesis space. Automatic-update participants queried the most informative feature in 30.3% of the situations, while manual-update participants chose the most informative feature in 37.6% of the situations. Thus, although manual-update participants used fewer feature queries overall, and did make some mistakes during hypothesis updating, they queried features with higher expected information gain than automatic-update participants. Along with the reaction time results described above, this suggests that these children thought more before making their choices and managed to choose more informative feature queries. Indeed, there was a weak but significant correlation of participants' mean feature query RT and EIG ($r = .20, t(116) = 2.17, p < .05$), verifying that longer RTs are associated with more informative feature queries.

2.2.6. Query-by-query behavior

Fig. 6 shows the mean proportion of feature vs. exemplar queries by query index within a round for each update condition split by age group, contrasted with simulated agents choosing any available buttons uniformly at random throughout the game. Older children show a much higher proportion of feature queries in the first three clicks of the automatic condition, and the first two of the manual condition. In both update conditions, the first three clicks are more likely to be feature than exemplar queries, and automatic-update subjects often make a fourth feature query before likely moving to exemplar queries. Both human conditions are quite different than the simulated random agent. Rather, the response profile of human participants looks generally like the optimal sequence: 3 feature queries and then one (sometimes two) exemplar queries. However, as was shown earlier, participants rarely chose the most informative feature to query at any given time, and manual participants made a number of updating errors. Where does the higher EIG for manual-update feature queries come from? Are they choosing the best feature query from the start, or are they simply better at testing more contextually-relevant features later in the round?

Fig. 7 shows the mean EIG of feature queries by feature query index (left), and for all queries (right), with a simulation based on the participants' data for comparison: although following the same sequence of situations as participants, this simulation shows the EIG if a query (just feature at left, or feature and exemplar at right) had been chosen at random in each instance. Fig. 7 reveals that people in the two update conditions had similarly informative first queries–especially for the 5–7 year-olds, who were not much better than random, but that manual subjects' subsequent few feature queries were more informative than automatic subjects' or the random choices. That is to say, manual-update participants chose feature queries that were more contextually appropriate for the particular set of remaining hypotheses, in contrast to automatic-update participants who–despite finding an informative feature for the first query–paid less attention to the unfolding situation. In fact, after the initial high-quality query, the younger automatic-update participants chose queries with nearly the same EIG as the random simulation, implying that they more or less ignored the features of the remaining hypotheses. For older automatic-update participants, feature query EIG was better than random after the first query, although it remained below manual-update EIG across feature queries.

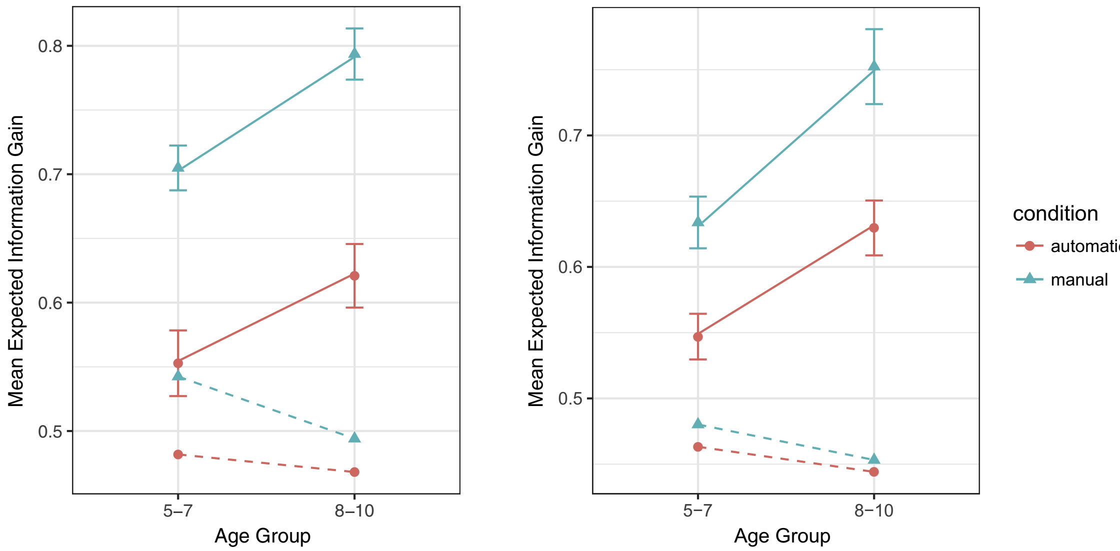
3. General discussion

In the present study, we manipulated the support children were given while updating a hypothesis space during a self-directed learning task. After making a feature (or constraint-seeking) query, participants in the automatic update condition were shown which insects were effectively ruled out at the press of a button, whereas manual update participants were required to select the insects that were consistent with the feedback themselves.

In line with previous research (Mosher & Hornsby, 1966; Ruggeri & Lombrozo, 2014), older children (ages 8–10) asked a

[7] Exemplar query EIGs alone are less interesting, as they are a simple function of how many hypotheses remain. Participants' choice of feature query, on the other hand, indicates how sensitive they are to the relevance of each feature–and to the context of their current situation. However, as the space shrinks, it is interesting to see whether participants' persist in making (now less informative) feature queries, or switch to (increasingly informative) exemplar queries.

[8] The same significant effects and similar mean EIG values were obtained when analyzing only the first two feature queries per round, when manual- and automatic-update participants were on more equal footing (i.e., before further manual errors–which could raise or lower the EIG of the remaining feature queries).



(a) Mean EIG for feature queries.

(b) Mean EIG for all queries.

Fig. 5. Mean expected information gain (EIG) for feature queries by age group and condition, with dotted lines showing the mean EIG of all queries available in the same situations as subjects (not the earlier random agents)–for comparison. Manual-update subjects had higher EIG than automatic-update subjects, and both were better than random–but suboptimal (1). Older children had higher EIG than younger children. Bars show ±1SE.

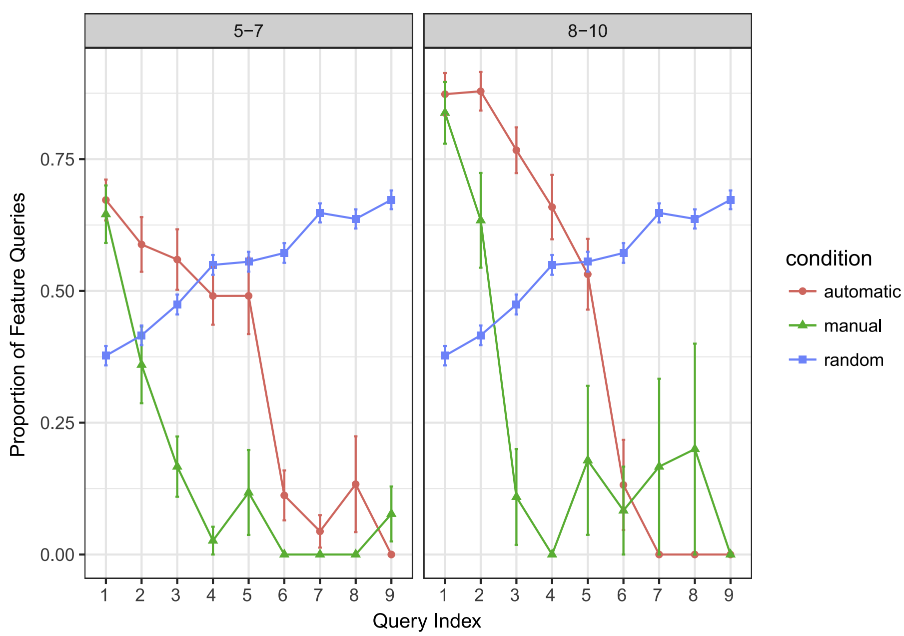
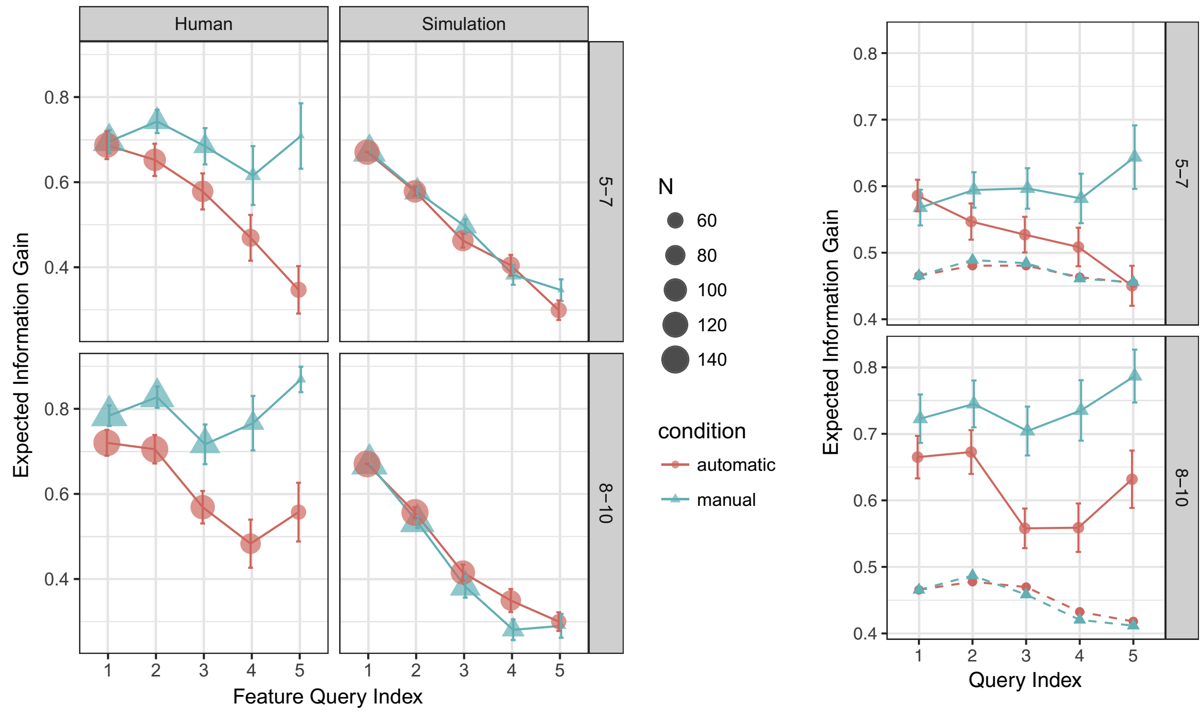


Fig. 6. Proportion of feature vs. exemplar queries by query for each update condition, with a randomly-querying agent for comparison. People in both conditions are more likely to make feature queries rather than exemplar queries in the first query of a round, at higher rates than random chance, but manual-update participants move more quickly to exemplar queries, and are overall more likely to make exemplar queries. Older children make a higher proportion of feature queries in the first few queries of both conditions.

higher proportion of constraint-seeking questions than younger children (ages 5–7), who relied more on hypothesis-scanning (i.e., exemplar queries), in both conditions. These qualitative analyses also found that children use more constraint-seeking questions (i.e., feature queries) in the automatic-update condition. On the surface then, these children were using a more efficient strategy than the manual-update children.

However, in terms of expected information gain, a context-sensitive measure of how well a chosen feature bisects the remaining hypothesis space, children in the automatic-update condition made less informative feature queries. We suggest that the greater mental effort required by manual updating actually led to more careful consideration of which feature query to use, and ultimately a better choice. This is a type of desirable difficulty in the sense that aspects that made the learning task ostensibly more difficult led to more sophisticated question asking behavior. Indeed, response times for feature queries were slower under manual updating, indicating that greater thought went into making those choices.

Our results provide important nuance to recent findings showing that children's question asking behavior conforms to the predictions of normative models such as EIG (Ruggeri & Lombrozo,



(a) Mean EIG by feature query. Error bars show +/-1SE.

(b) Mean EIG by all queries.

Fig. 7. Subjects' mean expected information gain (EIG) of feature queries (a) and all queries (b) by query index and condition, with simulated random choices (mean EIG of all remaining options from the same situations) for comparison (right panels of (a) and dotted lines of (b)). At left, although subjects' first feature query had nearly the same mean EIG in both conditions, the next few feature queries in the manual-update condition had higher mean EIG than the automatic condition. This suggests that manual-update subjects paid more attention to the remaining hypotheses after the first query, and made subsequent feature queries that were sensitive to the current context. Making four or more feature queries in a given round was quite rare, as most participants mostly switched to exemplar queries after the second or third feature query. EIG for all queries (right) show a similar pattern (and a stronger advantage over the simulation), meaning that manual participants are better at switching to exemplar queries at the right time, whereas automatic participants may be persisting in querying (now uninformative) features.

2015). Although even the youngest children asked more informative questions than a random guesser, the quality of children's questions varied widely and depended on the overall nature of the learning task and environment. This type of finding follows from the sensemaking loop in Fig. 1, which argues for a more interactive and integrated reasoning process.

Prior work has found that although quite young children show some of the requisite skills for successful active inquiry, such as the ability to distinguish confounded from unconfounded evidence to draw causal inferences (Gopnik et al., 2001; Kushnir & Gopnik, 2005, 2007; Schulz & Gopnik, 2004), the capacity to make use of these skills to engage in efficient self-directed explorations in complex tasks follows a protracted developmental trajectory (Chen & Klahr, 1999). The present findings provide a new perspective on why this might be the case. Although young children have the capacity to generate informative questions in certain circumstances, as shown here and, for example, by Ruggeri and Lombrozo (2015), children's abilities to ask informative questions and to benefit from the information yielded by their questions depends on the nature of the learning environment. Further, effective active inquiry involves the coordination of multiple cognitive processes–the ability to ask and learn from an effective question depends not only on children's capacity to recognize the most informative question given a particular context, but also to properly integrate the information that the question yields with the current hypothesis space (and then to realize what question will be informative to ask next). Children's actual capacity to engage in effective active inquiry to navigate a new learning environment thus depends on more than the ability to generate the right question, but also the coordination of this skill with other somewhat demanding cognitive processes.

We found evidence for a type of "desirable difficulty" in children's abilities to ask informative questions–children asked more informative questions when they had to update the hypothesis space on their own. It is, however, important to put this "desirable difficulty" finding into perspective. Although children overall seemed to ask more sophisticated questions in the manual update condition, they also made more mistakes. As a result they took more time to identify the bug and often failed at the task. These results speak to the complex interplay of component processes in self-directed learning. The interconnection of information-driven and motivational components makes it difficult to even identify what makes a task "easier" for a young child without first defining which aspect of behavior one wants to influence. At the very least, this study provides evidence that hypothesis updating is a difficult, error-prone step in the active inquiry process (which has often been under-appreciated in past work). From both theoretical and practical perspectives, it would be useful in future research to identify exactly what accounts for the benefits observed in the manual-update condition (e.g., increased motivation to avoid uninformative questions, deeper processing of the obtained evidence and so on) so that learning environments could be designed that maintain these benefits while also helping children to avoid some of the associated costs, such as errors in the updating phase. Moreover, it is worth noting that the hypothesis space in this task was explicitly represented–both in full, and during updating–unlike the mental hypothesis space in a verbal game of 20 questions, or in the realm of science. Although representing the hypothesis space explicitly made it possible to use a novel domain in which we could manipulate feature informativeness, and in which we could observe and support hypothesis updating and observe errors, this departure from a purely mental hypothesis space could mean that children's errors were due to visual attention, and may not apply in the same way in mental hypothesis spaces.[9] Future work might examine the behavioral effects of dropping the external representation of the hypothesis space after familiarization. However, using an external representation of the hypothesis space, and manipulating in this study allowed us to unveil an interaction between belief updating and question asking, two nonadjacent steps of the sensemaking loop. We hope this study will serve as a reminder that task design can have effects further downstream than expected in theory, for learners are sensitive to difficulty across task stages, and may choose strategies to ease their burden.

Acknowledgments

This paper is an extended and modified version of a paper that was presented at the Annual Conference of the Cognitive Science Society in 2016. This work was supported by the John Templeton Foundation "Varieties of Understanding" grant to TMG and MR and Grant No. BCS-1255538 from the National Science Foundation to TMG. We are grateful to Kathryn Yee, Aja Blanco, Christina Chu, and Talena Smith for data collection, and to Daniel Zeiger and the staff of the Discovery Room at the American Museum of Natural History for their support of this project.

References

Bonawitz, E., & Griffiths, T. (2010). Deconfounding hypothesis generation and evaluation in bayesian models. In S. Ohlsson & R. Catrambone (Eds.). *Proceedings of CogSci, Austin, TX* (Vol. 32).

Bruner, J. (1961). The act of discovery. *Harvard Educational Review, 31*(21–32).

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In R. Dale et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society. Austin, TX*.

Coenen, A., Rehder, B., & Gureckis, T. M. (2014). Decisions to intervene on causal systems are adaptively selected. *Cognitive Psychology, 79*, 102–133.

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

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5 and 6-month-olds. *Developmental Psychology, 49*(2), 243–249.

Denrell, J., & March, J. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science, 12*(5), 523–538.

Gopnik, A., Sobel, D., Schulz, L., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

Gureckis, T. M., & Markant, D. B. (2009). Active learning strategies in a spatial concept learning game. In *Proceedings of the 31st annual conference of the cognitive science society*.

Gweon, H., & Schulz, L. (2008). Stretching to learning: Ambiguous evidence and variability in preschoolers' exploratory play. In B. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 31st annual meeting of the cognitive science society, Austin, TX* (pp. 570–574).

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

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–667.

Klein, G., Moon, B., & Hoffman, R. (2006a). Making sense of sensemaking i: A macrocognitive model. *IEEE Intelligent Systems, 21*(5), 88–92.

Klein, G., Moon, B., & Hoffman, R. (2006b). Making sense of sensemaking ii: Alternative perspectives. *IEEE Intelligent Systems, 21*(4), 70–73.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development*. Financial Times/Prentice Hall.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866–870.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*(9), 678–683.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*(1), 186–196.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society. Austin, TX*.

Mosher, F. A., & Hornsby, J. R. (1966). Studies in cognitive growth. In *(chap. On asking questions)*. New York, NY: Wiley.

Needham, F. (1745). An account of some new microscopical discoveries founded on an examination of the calamary and its wonderful milt–vessels. London.

Nelson, J. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*(4), 979–999.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition, 130*, 74–80.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*(4), 608–631.

Rich, A., & Gureckis, T. M. (2015). The attentional learning trap and how to avoid it. In R. Dale et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society. Austin, TX*.

Ruggeri, A., & Feufel, M. A. (2015). How basic-level objects facilitate asking efficient questions in a categorization task. *Frontiers in Psychology, 6*(918), 1–13.

Ruggeri, A., & Lombrozo, T. (2014). Learning by asking: How children ask questions to achieve efficient search. In *Proceedings of the 36th annual conference of the cognitive science society*. Cognitive Science Society.

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

Ruggeri, A., Lombrozo, T., Griffiths, T., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In D. C. Noelle et al. (Eds.). *Proceedings of cogsci* (Vol. 37).

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Acm/ifips interchi conference on human factors in software* (pp. 269–276). New York, NY: ACM.

Schulz, L., & Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*(4), 1045–1050.

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176.

Steffe, L., & Gale, J. (1995). *Constructivism in education*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489.

Tsividis, P., Gershman, S. J., Tenenbaum, J. B., & Schulz, L. (2013). Information selection in noisy environments with large action spaces. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

[9] We thank an anonymous reviewer for this point.