Desirable difficulties during the development of active inquiry skills

George Kachergis, Marjorie Rhodes, & Todd Gureckis

{george.kachergis, marjorie.rhodes, todd.gureckis}@nyu.edu

Department of Psychology, New York University

New York, NY

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. Four- to ten-year-old the methods say 5 to 10 yo 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. Although this assistance improved some aspects of their active inquiry behavior, 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 may actually improve children’s active inquiry skills, thus illustrating a type of “desirable difficulty" for reasoning.

Keywords: question asking, information search, active inquiry, hypothesis testing, scientific reasoning

Desirable difficulties during the development of active inquiry skills

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 this early evidence, 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, active inquiry unfolds as a sequence of mental steps (see Figure 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 (Russell, Stefik, Pirolli, & Card, 1993; Klein, Moon, & Hoffman, 2006a, 2006b). 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 elements are properly functioning and aligned the bike moves forward. However, misalignment of even one component can be catastrophic.

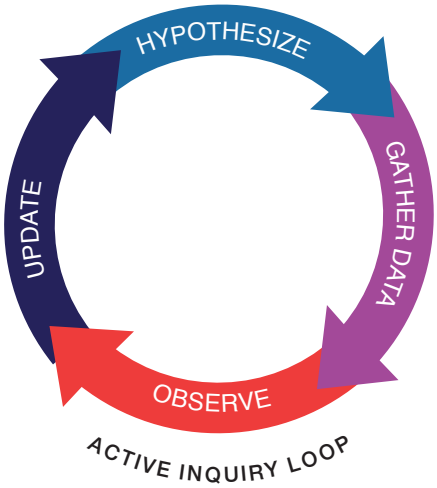

Figure 1. The active loop depicts the successive cognitive process that are engaged when attempting to derive a meaningful understanding of an initially ambiguous situation. The stages of the loop closely mirror the process of scientific reasoning engaged by scientists. However, a similar set of inductive 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).

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). However, understanding limitations in children's active inqury 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 four to ten-year old children the methods say 5 to 10 yo 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 children had to use the new evidence 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 suggests an important refinement to the information processing model summarized in Figure 1. i’m not sure about this last phrase... but i want to suggest some implication of our finding

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 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). Such results are somewhat consistent with the model described above because having the right set of hypotheses, or at least the right types of category information, in mind seems to drive more effective information search.

Maybe we can say a little more about EIG here. the context-sensitive analysis of question quality is really important. I think we want to prime the reader for "moving beyond" analyses of constraint vs. scanning and focus on context sensitive information measures. Keep this comment in mind when approaching the rest of the paper because there might be a few places to make adjustments in the way we discuss things. for example some of the longer text on EIG of the next subsection might be more appropriate here if we discuss context-sensitivty

Belief updating and active inquiry

While it is clear 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 intwined. The active inquiry loop in Figure 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 inquiry will be less effective (c.f., research on the hot stove effect, Denrell & March, 2001; Rich & Gureckis, 2015).

Coenen & 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). This model has been widely used in both the adult and developmental literature to understand how people decide between different queries (Oaksford & Chater, 1994; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Coenen, Rehder, & Gureckis, 2014; J. Nelson, 2005; Gureckis & Markant, 2009; Markant & Gureckis, 2012; J. D. Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2015), and will be used as a normative benchmark for question quality in the present study. 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 20-questions game a child might ask “Does your character have a hat?" or “Is your character a male?" To decide between these two queries EIG considers each possible answer (“yes" or “no" for both) and how much each answer would alter the learner’s current beliefs given the question. If all the remaining characters in the game are wearing hats then the answerer would never respond “no" to the hat question, but answering “yes" does not normatively alter the learner’s beliefs. In contrast, if roughly half the remaining characters were male, then either answer (and therefore the average) to the second question would strongly shift what the learner knows about the true character. On this basis 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 & 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 show showed clear evidence of biased belief updating (e.g., incorrectly interpretting 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 the ability to detect errors in belief updating). The other references here were dropped because the summary was seemingly unclear. See Latex beneath here for my comments

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 Figure 1 above are sequential, they likely rely on a common pool of cognitive and attentional resources and are thus not completely independent. In this case 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 is that they produce evidence that applies to a single hypothesis. If instead children ask constaint-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, 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 object standard of informatively given by the EIG model described in more detailed below. We additionaly analyze our data specifically in terms of constraint-seeking and hypothesis scanning question. 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 question s. In addition, we tested children in two key age groups?? and expected that younger children would benefit more from the assistance than would older children. We need to mention age as a hypothesized factor otherwise there is not reason to run multiple ages as far as i can tell

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. To increase translational impacts of the project, it 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.

Methods

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, 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. Participants were assigned in deterministically? counterbalanced order to one of two between-subjects conditions: the automatic-update or manual-update condition. how many were assigned to each condition?

Stimuli. On each round, children were presented with a display containing sixteen insects. One of the insects was randomly selected to be 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 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". Figure 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).

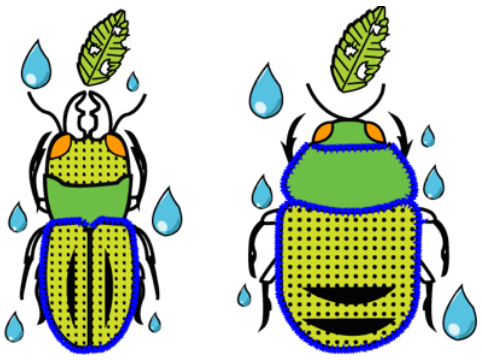

Figure 2. Examples of two insect body types with all 9 of the binary features present. Each round used one of 16 possible body shapes. Should we show the other body types?

Design. Across the sixteen items, some features were more frequent than others (e.g., one was relevant to eight of the insects), while some were very infrequent (e.g., two were relevant to only two insects), with an abstract structure shown in Figure 3 This should be a table. Also it is

not described very well. Exemplar should maybe be insect. Maybe A-P could be explained? What about the meaning of the 1 versus 0?. The abstract structure in Figure 3 was randomly assigned to the visual features for each child (i.e., can you unpack this), and then remained consistent across rounds. This gave child the possibility to learn the quality of different features across the rounds. The design of the abstract sturcture 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.

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

Figure 3. The abstract feature structure of the 16 exemplars used in each round. Each child had these abstract features randomly assigned to the visual features, but had a consistent assignment used round-to-round. Should explain what the numbers mean. Also the letters.

Each of these features was visually represented on a buttonRefer to a figure that shows the layout? i'd maybe give an overview of the display here, available for children to tap with their finger. An additional feature button depicted 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, chidren 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 elements 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.

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 an iPad game asking them to identify which one of 16 insects was hidden under a cartoon rug (see Figure 4). 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 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 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 becomes grayed out (i.e., eliminated).

After a feature query (i.e., constraint-seeking query), the insect under the rug gives feedback, saying "Yes!" (it has the feature; narrated by the experimenter), or "No!" (it does not have the feature). This is followed by the elimination phase, during which insect that are inconsistent with the feedback are eliminated, and the hypothesis space is 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 and all the no longer relevant insects are eliminated (grayed out), and the game returned to the guessing phase Above it says there are two phases - query and elimination. guessing phase is now introduced so maybe get the wording consistent. I'll leave it to you George because I'm not sure what is what. I think you mean query phase. In the manual-update condition, after a subject made a feature query

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

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 Figure 4. 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. Not sure what "at least" means here. also there was not explanation of all the features? some might be hard to understand based on their symbols. I moved this from someplace else but it feels a little out of place both in original location and here. In the manual-update it was possible for mistakes to be made during the elimination phase?. Insects that should have been eliminated but were kept (a 'miss') continued to be valid 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 direct 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?"

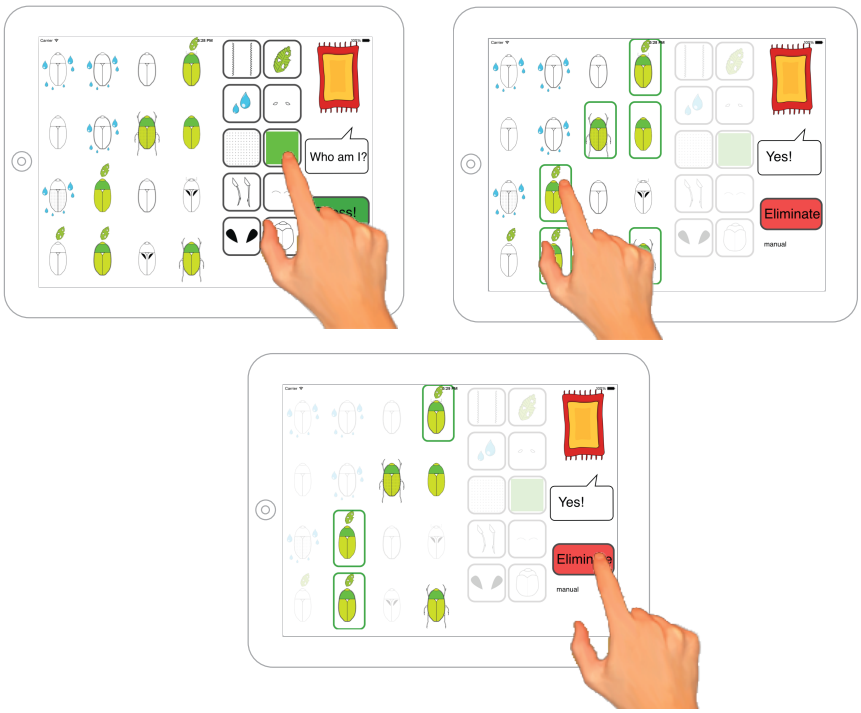

Figure 4. Task overview: in the upper left, a feature button is used, asking if the insect hidden under the rug is green. Given feedback ("Yes!"), participants in the manual update condition select the insects that are consistent with this new information (upper right), whereas in the automatic condition 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.

Results

Overall. We analyzed only the first 10 rounds of each game (only 8 children played more than 10 rounds, including one who played 51 rounds). This covers 722 rounds from 121 children. The mean number of total queries (feature and exemplar) taken to complete a round was 6.5 in the automatic-update condition, and 7.6 in the manual-update condition. Although the median queries to complete a round in each condition was 6, the distributions were significantly different What distributions? Different from what? From 6 or from each other? (Kolmogorov-Smirnov test, $D = 0.13$, $p < .01$). For comparison, we simulated 700 rounds of the game with an agent that

clicked randomly in the task, choosing uniformly at random on the first click from 16 exemplars and 10 feature buttons, and continuing with whatever stimuli (and feature queries) remain after each click, while making no update errors. This random agent took on average 8.9 queries (median: 9) to complete a round—more queries than participants in either condition. This suggest at least minimal structure in children's active inquiry behavior.

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. This analysis indicated significant main effects of condition (F(1,229) = 4.60, $p < .05$) and age group (F(1,229) = 12.20, $p < .001$), and no significant main effect of query type (F(1,229) = 0.10, $p = .75$). Overall, older children required fewer queries of either type to complete a round ($M_{5-7} = 4.2$, $M_{8-10} = 3.3$), also evidenced by a significant negative correlation between what and age? mean clicks? with age ($t(119) = 3.24$, $p = .001$, $r = -.28$). There were significant interactions of condition and query type (F(1,229) = 22.18, $p < .001$), and age group and query type (F(1,229) = 12.25, $p < .001$), 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 ($M_{man} = 5.0$, $M_{auto} = 3.2$, $t(103.5) = 4.1$, $p < .001$), while there were more feature queries in the automatic condition ($M_{auto} = 3.8$) ($M_{man} = 3.3$, $t(102.9) = 2.1$, $p < .05$). These query rates are all lower than the simulated random rounds' mean number of feature queries (6.5) and exemplar queries (5.3), but above the optimal.[2]

Figure 5 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 ($M_{5-7} = 4.8$, $M_{8-10} = 3.0$, $t(119.0) = 4.00$, $p < .001$). Older participants used a greater proportion of feature queries than younger participants in both the automatic ($M_{5-7} = .50$

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

vs. $M_{8-10} = .66$, $t(57.2) = 3.12$, $p < .01$) and manual conditions ($M_{5-7} = .39$ vs. $M_{8-10} = .50$, $t(50.3) = 2.30$, $p < .05$). Thus, both conditions replicate the Mosher and Hornsby (1966) finding that older children use a greater proportion of constraint-seeking questions.

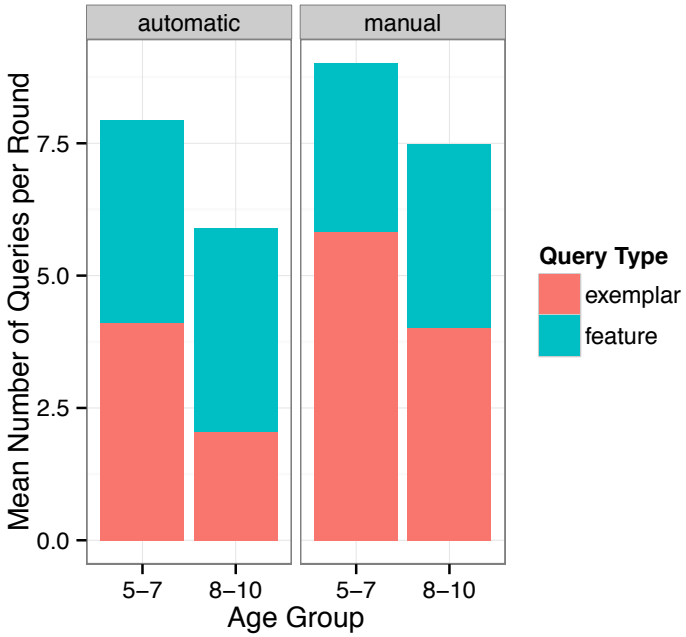

Figure 5. Mean number of queries of each type per round by age and condition. Older children use fewer exemplar queries than younger children. Manual-update participants used more exemplar queries and fewer feature queries than automatic-update participants.

like the insect/bug issue the feature query/constraint seeking thing is a terminology confusion. personally i like feature/exemplar but in order to connect with the developmental literature maybe we should use the other? The finding of more feature queries in the automatic condition and more exemplar queries in the manual condition raises a number of questions. Are participants reluctant to use feature queries in the manual condition because of the difficulty of updating the hypothesis space? When manual-update participants do use a feature query, do they think more carefully about which feature they choose? We investigate response times in each condition to reveal how much care and thought participants are putting into making each type of query. The transitions between results sections are always phrased in terms of rhetorical questions which aren't my favorite. they imply a type of exploratory data analysis that undermines some of

the statistical tests used here. Furthermore some of these specific questions would be hard to answer in reality, right... I might just present the reaction time results and then discuss the possible interpretation of them afterwards rather than making this claim out front.

Response Times. Participants' median 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 ($F(1,229) = 42.52$, $p < .001$) and condition ($F(1,229) = 4.14$, $p < .05$), but not a significant main effect of age group ($F(1,229) = 0.73$). On average, participants took longer to make queries in the manual condition (4800 ms) than in the automatic condition (4000 ms). Overall, participants took much longer to make feature queries (7,470 ms) than to press an exemplar button (2,680 ms), perhaps indicating more thought before making more complex queries. There was also a significant interaction effect of query type and condition ($F(1,234) = 11.85$, $p < .001$). Figure 6 shows the mean of subjects' median RTs for each query type, split by condition. Feature queries were slower in the manual-update condition (7900 ms vs. 5430 ms in automatic), 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 were faster in the manual-update condition (1850 ms vs. automatic: 2570), which could be greater readiness to use the simpler strategy. 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.[3] Therefore we proceed to investigate errors in manual updating.

[3] If the correct answer is mistakenly eliminated, exemplar queries are needed to find it and finish the round. These additional exemplar queries were excluded from analysis.

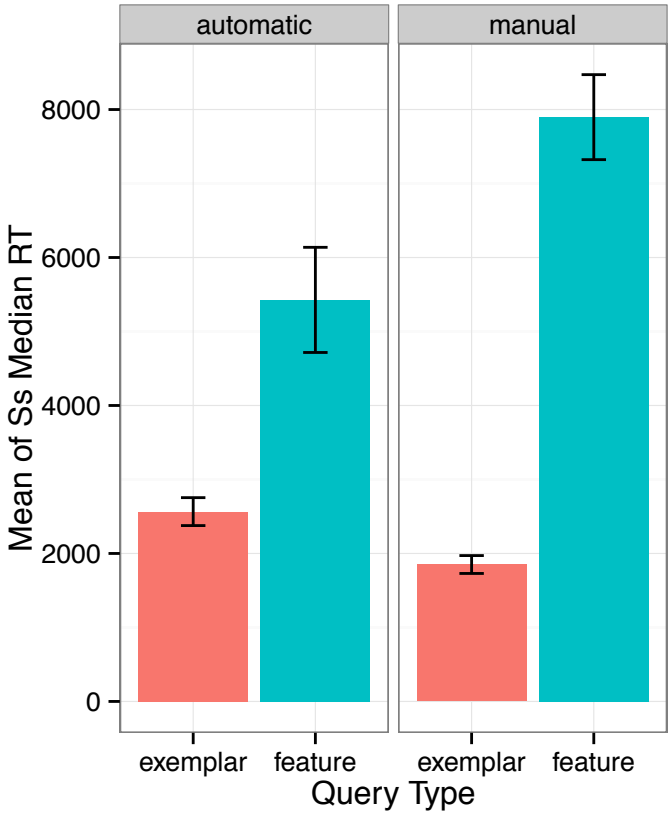

Figure 6. Mean of participants' median RT for each condition and query type. Exemplar queries were faster than feature queries, which represent a more complex strategy and thus likely required more thought. Feature queries were particularly slower in the manual-update condition. Error bars show +/-1SE.

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 resource of 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 children's feature queries which provides a context-sensitive measure of query informativeness.

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 in not absolute (either could be better in different circumstances). As describe 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).

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

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?

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

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. 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, and due to the partially-nested feature structure used, maximal EIG queries are often available at other stages of the round.

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

In our study, the EIG for each participants' feature queries[4] were computed, and their mean 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) = 55.0, $p < .001$) and age group (F(1,115) = 12.42, $p < .001$), with no significant interaction effect (F(1,115) = 0.2, $p > .05$).[5] Figure 7 shows mean EIG per feature query by age group and condition, along with a baseline simulation showing the mean EIG of all the feature queries (i.e., as if each subject had chosen randomly from the feature queries). Note that although randomly-chosen features for the manual-update subjects have a higher EIG than for automatic-update subjects (driven in part by update errors quickly reducing the hypothesis space), the simulated random EIGs are far below the corresponding human data. Mean EIG of feature queries for each subject was marginally significantly correlated with age ($t(116) = 1.77$, $p = .08$, $r = .16$), 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 ($M_{auto} = .60$, $M_{man} = .74$, $t(116) = 5.49$, $p < .001$). Thus, although manual-update participants

[4]Exemplar query EIGs are less interesting, as they are a simple function of how many remaining hypotheses there are. 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.But the context sensitivity also means choosing the feature query when it is better than any available exemplar query and not necessarily when they are matched.

[5]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).

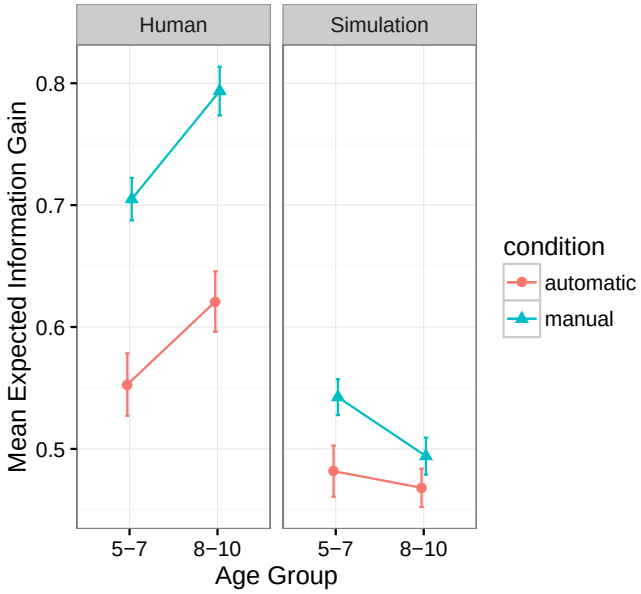

Figure 7. Mean expected information gain for feature queries by age group and condition, with a simulation making random feature queries–in the same situations as subjects (not the earlier random agents)–for comparison. Manual-update subjects had higher EIG than automatic-update subjects, and both were better than random–but suboptimal (1). Older children had higher EIG than younger children. Bars show +/-1SE.

used fewer feature queries overall, and tended to make 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 in fact performed better. 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.

Click-by-click behavior. Figure 8 shows the mean proportion of feature vs. exemplar queries by click 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?

Figure 9 shows mean EIG of feature queries by click index (ignoring exemplar clicks), 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 feature query had been chosen at random in each instance. Figure 9 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.

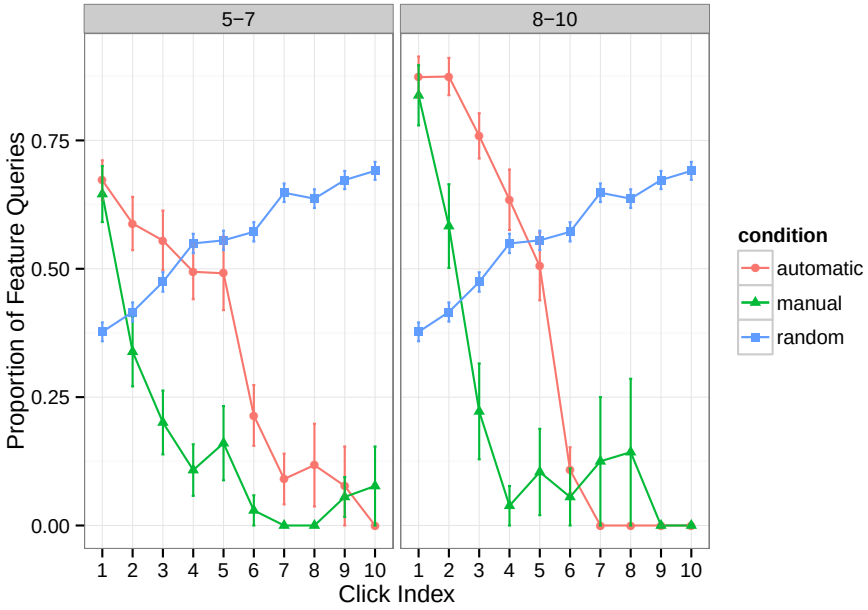

Figure 8. Proportion of feature vs. exemplar queries by click for each update condition, with a randomly-clicking agent for comparison. People in both conditions are more likely to make feature queries rather than exemplar queries in the first three clicks of a round, but manual-update participants move 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 clicks of both conditions.

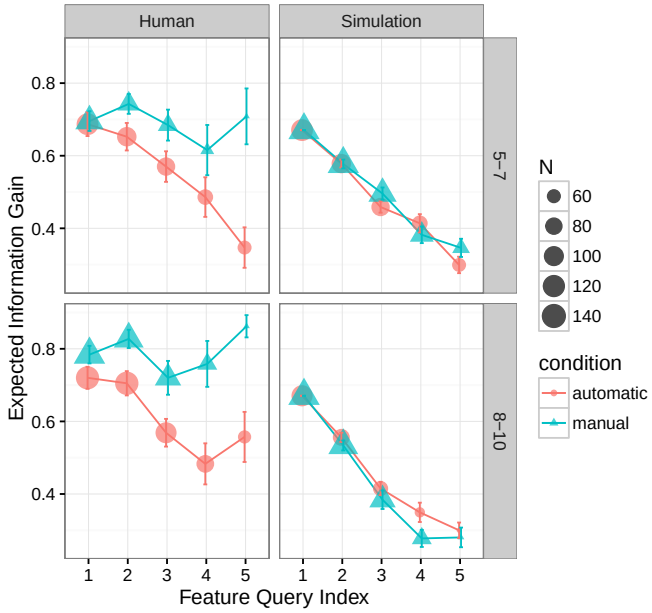

Figure 9. Subjects' mean expected information gain (EIG) of feature queries by click index (ignoring all exemplar clicks) by condition, with symbol size indicating frequency of occurrence. Although the first feature query–made when the full hypothesis space is possible–had nearly the same mean EIG in 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. Bars show +/-1SE.

General Discussion

The present study asked children 5-10 years of age to learn feature distributions in an unfamiliar hypothesis space, and examined both their qualitative questioning strategies, and how efficiently they were able to search that space. Previous studies have examined question asking in somewhat familiar hypothesis spaces (Herwig, 1982; Mosher & Hornsby, 1966; J. D. Nelson et al., 2014; Ruggeri & Lombrozo, 2014, 2015), but no study we are aware of has considered the step of updating the hypothesis space when new evidence is received–let alone in a novel hypothesis space. In many previous developmental studies, experimenters help children update the remaining hypothesis, quite reasonably: this step is potentially quite challenging, especially for younger children. Importantly, we manipulated the support children were given while updating the hypothesis space: after a feature query, participants in the automatic update condition were shown which insects were eliminated at the press of a button, whereas manual update participants were required to select the insects that were consistent with the feedback.

In line with previous research (Mosher & Hornsby, 1966; Ruggeri & Lombrozo, 2014), the present study found older children (ages 8-10) asked a 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. From this qualitative analysis alone, then, it is tempting to conclude that automatic updating leads to a better querying strategy than manual updating.

However, in terms of expected information gain, a context-sensitive measure of how well a chosen feature bisects the remaining hypothesis space, it turned out that children in the automatic-update condition made less informative feature queries. We suggest that the greater mental effort required by manual updating actually lead to more careful consideration of which feature query to use, and ultimately a better choice. Indeed, response times for feature queries were slower under manual updating, indicating that greater thought went into making those

choices, corroborated by the fact that slower feature query RTs were correlated with more informative queries. Within-round analysis found that automatic-update participants were likely to make feature queries for the first few clicks, while manual-updaters switched often switched to exemplar queries after one (5-7 year-olds) or two feature queries (8-10 year-olds). In terms of quality, feature queries in both update conditions were similar for the first query in a round–and better than the simulation. However, manual-update subjects made more contextually-sensitive feature queries after the first query, whereas automatic-update participants looked much like they were choosing random feature queries, without regard for the current hypothesis space's features. In both conditions, older children made more informative feature queries, but even 5-7 year-olds asked far more informative questions than a random simulation, showing some efficiency in navigating an unfamiliar domain even after only a few minutes of experience.

In summary, this study provides evidence that hypothesis updating is a difficult, error-prone step in the active inquiry process. Moreover, children of all ages tested here (5- to 10-years-old) are sensitive to the difficulty of this step: when aided in hypothesis updating, asked more constraint-seeking questions than when they had to manually update the space. However, we also uncovered evidence of a desirable difficulty in this step: manual updating resulted in more informative, contextually-sensitive constraint-seeking questions than the supported update process. Surprisingly, even younger children (5- to 7-year-olds) in the manual-update condition made feature queries with higher expected information gain than their peers in the automatic-update condition. Future work will aim to reduce errors in manual hypothesis updating via intervention, and will aim to uncover other bottlenecks–or desirable difficulties–in active inquiry.

Finally, it is worth noting that this partially self-guided iPad study was conducted in the relaxed learning environment of the American Museum of Natural History's Discovery Room. Children's developing abilities to engage in the basic steps of scientific thinking are on full display in such informal science learning environments, such as at science and children's museums. Indeed, a central goal of these environments is to provide children with hands-on

opportunities to learn from their own explorations, to enable them to gain active experience with the steps of scientific investigation, to discover new knowledge, and to develop enthusiasm for and interest in science (Bell, Lewenstein, Shouse, & Fender, 2009; Fenichel & Schweingruber, 2010). As a result, informal science environments may provide an excellent domain in which to investigate the development of understanding. This study demonstrates that such environs can be suitable and fruitful venues for knowledge discovery not only for children, but for scientists, too.

Acknowledgments

This work was supported by the John Templeton Foundation “Varieties of Understanding" grant to TMG and MR. We are grateful to Kathryn Yee, Aja Blanco, and Christina Chu for data collection.

References

Bell, P., Lewenstein, B., Shouse, A., & Fender, M. (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, D.C.: National Academies.

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

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.

Fenichel, M., & Schweingruber, H. (2010). *Surrounded by science: Learning science in informal environments*. National Academies.

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* (pp. 570–574). Austin, TX.

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.

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 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.