Desirable difficulties in 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. We hypothesized an intrinsic link between the ability to update beliefs in light of evidence and the ability to ask informative questions. Four- to ten-year-old children played an iPad game asking them to identify a hidden bug. Learners could either ask about individual bugs, or make a series of feature queries (e.g., "Does the hidden bug 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. Although we found that helping children update their beliefs 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.

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

Introduction

A central aim of science education is to teach students how to approach the task of understanding their environment. Rather than teaching only a catalogue of facts about the biological and physical worlds, current standards emphasize teaching the conceptual and analytic skills that underlie science: detecting patterns in environments that initially appear chaotic, abstracting the general principles that can be used to understand and predict events, and importantly, learning how to ask informative questions to reveal these patterns and principles when they are not immediately obvious (Bransford, Brown, & Cocking, 2000; Donovan & Bransford, 2005; Duschl, Schweingruber, & Shouse, 2007).

Many of the cognitive skills required for active scientific inquiry 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 show may be that active inquiry depends on the coordination of a variety of component cognitive processes (belief updating, decision making, hypothesis generation, etc.). 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 is like a bicycle: when all the elements are properly functioning and aligned the bike moves forward. However, misalignment of any one component can be catastrophic.

The present study joins with some recent work which attempts to decompose the component processes involved in active inquiry (e.g., Bonawitz & Griffiths, 2010). In particular, we tasked four to ten-year olds to identify a hidden bug in a simple iPad variant of the classic "Guess Who?" game. Children asked questions to try to identify the hidden bug. Across conditions, 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 update 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 more effective question-asking. Interestingly, our results opposed our main hypothesis in that elements which ostensibly made our task more difficult actually improved the quality of children's inquiry behavior.

How the ability to ask revealing questions develops

Experimental tasks based on the "Guess Who?" game have often been used to study question asking and active inquiry with both children and adults. In the game, the asker (participant) tries to determine a hidden object known only to the the answerer (experimenter) (e.g., "What animal am I thinking of?") 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 (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 several equally probable solutions remaining (e.g., Ruggeri & Lombrozo, 2015; 2015).

Whereas previous work has focused on developmental changes in when children generate informative, hypothesis-scanning questions, less prior work has considered possible developmental changes in how children make use of the new evidence that their questions reveal. As described above, effective active inquiry involves the coordination of multiple cognitive processes–representing the hypothesis space, generating an informative query, updating one's representation of the hypothesis space in light of the data produced by the query, and so on. As suggested by prior work, hypothesis scanning questions might be easier for young children to generate because they do not require abstracting informative higher level features (features to query that group classes of hypotheses together and might allow them to be eliminated at once). Yet, another reason why hypothesis scanning questions might be easier for young children is that they produce evidence that is easier for them to process. As a hypothesis scanning question is answered, children are told directly whether the item they queried is correct or not. If instead children ask about a feature (as in a constraint-seeking question), additional cognitive processing is required–children have to take that new information (e.g., that a hidden animal has antennae) and consider each remaining possible exemplar in light of this information (e.g., check if each one has the antennae) and eliminate from the hypothesis space any 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 address these issues, in the present study we manipulated whether children received assistance in updating their hypothesis space or had to undertake this process on their own, following the receipt of new evidence obtained by their queries.

Experiment

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.

Stimuli On each round, sixteen bugs with the same body shape but with varying features were used as stimuli. Bugs 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 1 shows an example of two of the body shapes used, each with all of the binary features present. One of the sixteen possible bugs was chosen as the "hidden bug" on each trial which children attempted to identify by asking questions. The hidden bug was randomly selected on each round, and each round had differently-shaped bug bodies, selected from a pool of 16 unique body shapes. The bug task was used to fit thematically with the content of the AMNH Discovery Room activities.

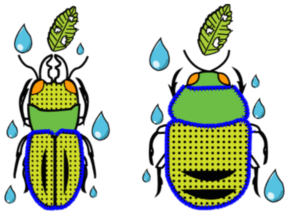

Figure 1: Examples of two bug types with all 9 of the binary features present. Each round used one of the 16 body shapes.

Design Across the sixteen items, some feature were more frequent than others (relevant to eight of the possible bugs), while some were very infrequent (relevant to only 1 bug), with an abstract structure shown in Figure 2. This introduced strong differences in the informational utility of each feature. For example, given no other information it would be informative to ask about feature F1 because is it shared with half the possible bugs. In contrast, feature F9 is less informative on the first trial because most of the bugs do no have this feature. The abstract features in Figure 2 were randomly assigned to the visual features for each participant, and then remained consistent across rounds. This gave participants the opportunity to learn the structure across rounds, for example to perhaps figure out which visual features are most relevant to ask about first.

Each of these features was represented on a button, available for participants to query. Before participants were allowed to begin, the experimenter explained at least three of these buttons, randomly selected. An additional feature button depicted a particular body shape that was not relevant to the bugs on display. Instead of choosing a feature button, participants could at any time query an exemplar to determine if it was the hidden bug or not. This paradigm thus enabled us to investigate both the qualitative strategies used by participants (constraint-seeking feature queries or hypothesis-scanning exemplar queries) and to quantify how efficiently participants searched the hypothesis space, within and across rounds as they learn a novel structured stimulus space. Moreover, we introduced a novel manipulation: after making a feature query, participants in the manual-update condition had to select the hypotheses that were consistent with the feed-

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 2: The abstract feature structure of the 16 exemplars used in each round. Each participant had these abstract features randomly assigned to the visual features, but had a consistent assignment used round-to-round.

back, whereas participants in the automatic-update condition had the hypothesis space automatically updated. This manipulated the ease of updating the hypothesis space: a difficult step in the cycle of active inquiry that has not been well-studied.

Procedure After being trained by an experimenter on a simpler version of the task with unrelated stimuli (a dog searching dog houses), participants played 5 or more rounds of an iPad game asking them to identify which one of 16 bugs is hidden under a rug (see Figure 3). The task alternated between the query phase and the elimination phase. In the query phase, players either queried individual bugs, or used feature queries (e.g., "Is the hidden bug green?") to find out whether the hidden bug had a particular feature. If a single exemplar was queried by tapping on it, feedback was immediate: if it happened to be the hidden bug, a smiley face appears and the round was done, whereas if the tapped exemplar was not the hidden bug, a red "X" appeared on top of the tapped bug and the bug becomes grayed out (i.e., eliminated). After a feature query, the bug gave feedback, saying "Yes!" (it has the feature; narrated by the experimenter), or "No!" (it does not have the feature). This was followed by the elimination phase, during which bugs that are inconsistent with the feedback were eliminated, thus narrowing the hypothesis space.

Participants were assigned in counterbalanced order to one of two hypothesis-updating conditions. In the automatic-update condition, after the feedback from a feature query, subjects merely pressed the "Eliminate" button and all the irrelevant bugs are eliminated (grayed out), and the game returns to the guessing phase. In the manual-update condition, after a subject made a feature query and saw feedback, they had to select each bug that was consistent with the feedback for that feature, as shown in the top right of Figure 3. Bugs were selected (denoted by a green box) by tapping, and could be deselected by tapping again. Only when participants were done selecting bugs did the experimenter press the "Eliminate" button, which eliminated any bugs that were not selected. Although manual-update participants received training for the manual elimination in the dog house training task, as well as gentle reminders in the first round of the bug game, it should be noted that it was possible for mistakes to be made during manual updating–unlike in the automatic condition.

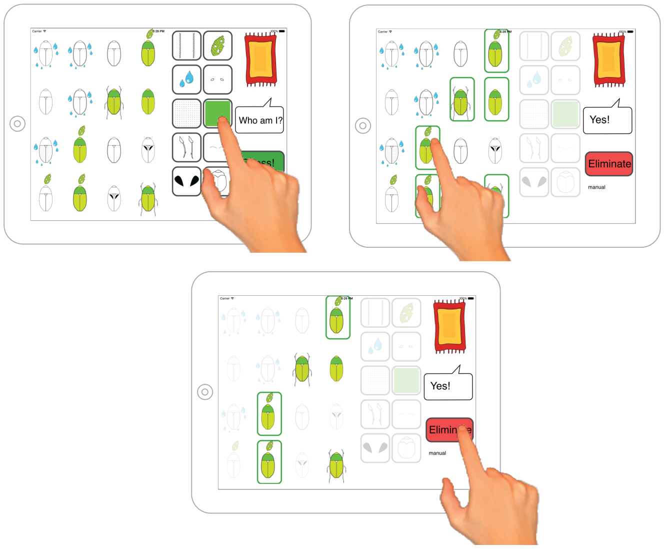

Figure 3: Task overview: in the upper left, a feature button is used, asking if the bug hidden under the rug is green. Given feedback ("Yes!"), participants in the manual update condition select the bugs that are consistent with this new information (upper right), whereas in the automatic condition the consistent bugs 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 bug.

Results

Overall We analyzed the first 10 rounds from each child (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 (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. This agent took on average 8.9 queries (median: 9) to complete a round.

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 button type and condition (F(1,229) = 12.89, $p < .001$). Figure 4 shows the mean of subjects' median RTs for each button 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 ms), which could be greater readiness to use the simpler strategy.

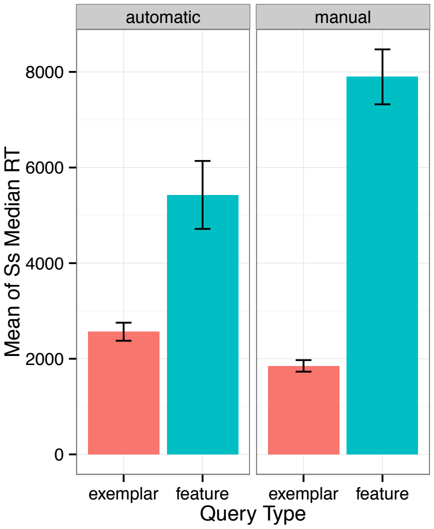

Figure 4: Mean of participants' median RT for each condition and query type. Exemplar queries were faster than feature queries, which represent a more complex strategy and thus likely required more thought. Feature queries were slower in the manual-update condition: it seems the difficulty of updating in this condition made participants think even more carefully about using feature queries. Error bars show +/-1SE.

Querying Behavior Participants' mean number of queries per round were subjected to an ANOVA with condition 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 total queries to complete a round ($M_{5-7} = 4.2$, $M_{8-10} = 3.3$), also evidenced by a significant negative correlation 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, $p < .001$). No other interactions were significant (all Fs < 1).

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. 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 were all lower than the simulated random rounds' mean number of feature queries (6.5) and exemplar queries (5.3), but above the optimal.[1] Older participants used a greater proportion of feature queries than younger participants in both the automatic ($M_{5-7} = .50$ 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.

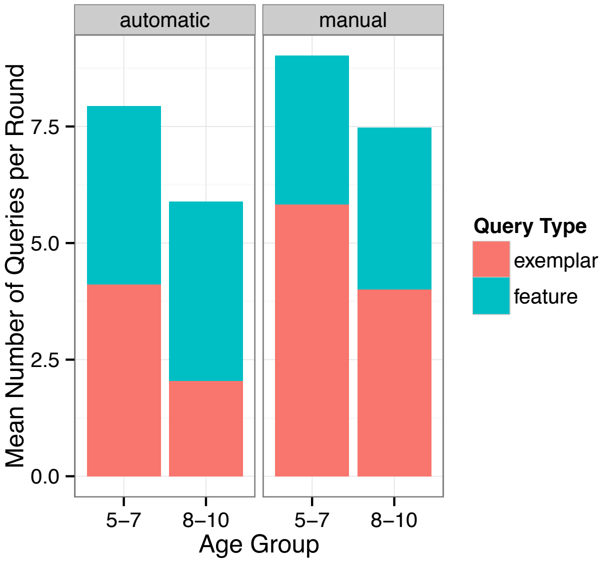

Figure 5: Mean number of queries of each type per round by age and condition. Error bars show +/-1SE.

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 loathe 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.[2] Therefore we proceed to investigate errors in man-

[1] Although there were at first more exemplars (16) than feature buttons (10), after the first 1-2 clicks there would likely be few exemplars remaining, thus the expected number of exemplar queries is lower than the expected number of feature queries in the simulation.

[2] If the correct answer is mistakenly eliminated, additional clicks on the grayed-out bugs were needed to find it and finish the round.

ual updating, as well as information theoretic analyses that will indicate whether the quality of feature queries varied in the two update conditions. Although the qualitative analyses have thus far revealed interesting effects that build on the previous literature, as the game unfolds the utility of different query types (and specific queries) changes, and can be best quantified using a more sophisticated model-based approach to understanding the quality of children's question asking.

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 bug, 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 bug had to be tapped to be kept–whereas a false alarm is an error of omission (i.e., failing to tap a bug), 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, $p > .05$). Given the fairly high rate of errors in manual updating, it is perhaps unsurprising that fewer feature queries and more exemplar queries were made in this condition than under automatic updating of the hypothesis space. However, RT analyses indicated that feature queries took longer under manual updating: is this just reluctance, or could it be that feature queries were more carefully considered in this condition than under the ease of automatic updating? The expected information gain of children's feature queries provides a measure of their sensitivity to the information structure in the stimuli.

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. The appropriate way to analyze the contextual sensitivity (i.e., are they choosing a feature that is present for half of the remaining exemplars, thus quickly reducing the hypothesis space?) 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. 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 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.

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

EIG has often been proposed as a model of how children might evaluate the quality of possible queries. 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. Likewise, Ruggeri, Lombrozo, Griffiths, and Xu (2015) found evidence that children's patterns of search decisions were well-explained in terms of EIG.

In our study, the EIG for each participants' feature queries[3] 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, $p > .05$).[4] Figure 6 shows mean EIG per feature query by age group and condition. Mean EIG of feature queries for each subject was marginally correlated with age ($t(116) = 1.77$, $p = .08$, $r = .16$), suggesting 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 used fewer feature queries overall, and tended to make mistakes during hypothesis updating, the greater amount of time they spent when choosing a feature query tended to pay off: manual-update participants queried features with higher expected information gain than automatic-

This occurred rarely, happening in $< 10\%$ of rounds.

[3] 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, as it is based on the remaining bugs' features.

[4] 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).

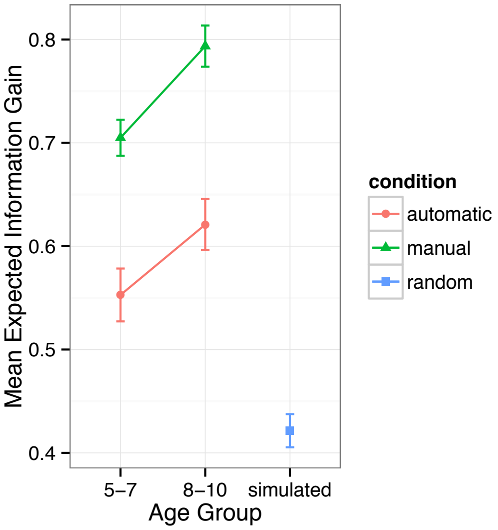

Figure 6: Mean expected information gain for feature queries by age group and condition, with simulated subjects making random feature queries for comparison. Manual- update subjects had higher EIG than automatic-update subjects, and both were better than random–but suboptimal (1). Older children had higher EIG than younger children. Bars show +/-1SE.

update participants. 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.

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. 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 bugs were eliminated at the press of a button, whereas manual update participants were required to select the bugs that were consistent with the feedback.

In line with previous research (Mosher & Hornsby, 1966; Ruggeri & Lombrozo, 2014), 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.

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, perhaps indicating that greater thought went into making those choices. Indeed, slower feature query RTs were correlated with higher EIG. In both conditions, older children made more informative feature queries, but even 5-7 year-olds asked far more informative questions than a simulation that chose a random sequence of queries, 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 are sensitive to the difficulty of this step: if aided in hypothesis updating, they will ask more constraint-seeking questions than if they must manually update the space. However, we also uncovered evidence of a desirable difficulty in this step, for manual updating resulted in more informative, context-sensitive constraint-seeking questions than the supported update process. Future work will aim to reduce errors in hypothesis updating and discover other bottlenecks–or desirable difficulties–in active inquiry.

Acknowledgments

This work was supported by the John Templeton Foundation "Varieties of Understanding" grant to TMG and MR.

References

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.

Bransford, J., Brown, A., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* National Research Council.

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.

Donovan, M., & Bransford, J. (Eds.). (2005). *How students learn: Science in the classroom.* Nat'l Research Council.

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades k-8.* Washington, D.C.: National Research Council.

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.

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

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

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.

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 CogSci 36.* 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.*