Where Questions Come From: Reusing Old Questions in New Situations

Emily G. Liquin (emily.liquin@nyu.edu)
Todd M. Gureckis (todd.gureckis@nyu.edu)

Department of Psychology, New York University
New York, NY 10003 USA

Abstract

Question asking is a powerful means by which humans learn. However, asking a question requires searching through a massive space of possible questions to find a single question that is relevant and informative. How do humans efficiently accomplish this task? Drawing on prior research on other decision problems, we propose that the search for new questions is constrained by those encountered in the past, so that people frequently reuse questions (or parts of questions) rather than generating new questions "from scratch." We find empirical support for this prediction, and we find that this "question reuse" has consequences for the informational value of people's questions. Taken together, this research sheds new light on the mechanisms behind human question asking abilities and, more generally, how we narrow down a large space of possibilities to find a single solution.

Keywords: question asking; active learning; information search; expected information gain

Introduction

The ability to ask questions provides humans with a powerful way to learn about the complex world around us. By asking questions, we can rapidly access information that cannot be observed directly (e.g., "How does the COVID-19 virus infect a cell?") or can only be inferred through repeated experience (e.g., "What's your favorite restaurant?"). Unsurprisingly, then, question asking drives cognitive development (Chouinard, 2007; Ronfard et al., 2018) and predicts learning in educational contexts (Rosenshine et al., 1996).

Though questions are ubiquitous and useful, question asking poses a difficult cognitive and computational challenge. The space of possible questions is vast, and the best question to ask varies widely from situation to situation (Coenen et al., 2019). Nonetheless, even young children ask sophisticated questions (Callanan & Oakes, 1992; Chouinard, 2007), and adults readily adapt their questions to different situations and goals (Rothe et al., 2018, 2019). In contrast, advanced computational models can generate sensible questions about images or texts (Du et al., 2017; Jain et al., 2017), but they usually cannot achieve human-like question asking. How do humans achieve what these models cannot?

In the present research, we explore the cognitive mechanisms that enable humans to ask informative questions. Drawing on prior research on reasoning and decision making (e.g., Morris et al., 2021), we hypothesize that prior experience with questions in a particular context can constrain the search for questions in new situations. Then, we present an experiment that tests and finds support for this hypothesis, and therefore advances our understanding of how humans solve the challenge of asking informative questions. Taken together, this research provides new constraints on computational models of question asking, and it has implications for how informative questions might be encouraged in educational settings (Good et al., 1987; Graesser & Person, 1994). In the following section, we review prior research before turning to our empirical investigation.

Question Asking and Other Search Problems

The process of asking a question bears resemblance to other decision making tasks: deciding what to cook for dinner, figuring out how to quickly stop a leak, coming up with a name for a new pet. To solve these problems, an individual must (1) search a large space of possibilities to generate a modest number of candidate solutions, (2) evaluate these candidates according to some measure of quality, and (3) select the best.[1]

The second and third steps of this process have been studied extensively, both for question asking (for a review, see Coenen et al., 2019) and for other decision problems (e.g., Rangel et al., 2008). For example, researchers have proposed several metrics for a question's quality (Crupi et al., 2018; Nelson, 2005; Nelson et al., 2010), and people's evaluation and selection of candidate questions and information-seeking actions are well described by these metrics (e.g., Coenen et al., 2015; Rothe et al., 2018; Ruggeri et al., 2016; Steyvers et al., 2003).

But where do these candidate questions come from? Research on other decision problems has proposed that people generate candidate solutions according to their past frequency and quality (Bear et al., 2020; Morris et al., 2021; Phillips et al., 2019; see also Dasgupta et al., 2018). For example, the dinner recipes that come to mind are those that frequently produced a delicious meal in the past. Indeed, students' question asking and problem solving improve after teachers model good questions (Birbili & Karagiorgou, 2009; King, 1990, 1991)—and this could be because students reuse the modeled questions. Relatedly, people selectively explore tasks that have previously resulted in learning progress (Ten et al.,

[1] These processes may not necessarily occur in this order. For example, one might generate a single candidate, evaluate it, then only proceed to further generation if the initial candidate does not reach a certain threshold of quality.

2021). However, it has not been directly tested whether people selectively draw from a cache of previously encountered questions when generating questions in new situations.

In contrast, a recent model of question asking (Rothe et al., 2017) generates candidate questions by randomly searching through a compositional *question grammar,* made up of "primitives" that can be composed to produce questions of arbitrary complexity (see also Tian et al., 2020). Rothe et al.'s model has successfully predicted important aspects of human question asking (Rothe et al., 2017, 2018, 2019). However, in its current form, it does not predict any effect of prior experience on later question asking—in particular, because it searches the question grammar at random, starting its search "from scratch" every time a new question is needed.

The Present Research

In the present research, we test whether people draw on a cache of previously encountered questions when generating questions in new situations. If this is the case, people should selectively ask questions similar to previous questions. For example, after encountering the questions "How big is a dog?" and "How big is a cat?", an individual might later reuse this *question template* ("How big is *animal*?") to ask "How big is a hamster?" Furthermore, people might recombine components of previously encountered questions into completely novel questions ("How many animals are bigger than a raccoon?"). In contrast, if questions are generated by random search through a question grammar, as predicted by the computational model proposed by Rothe et al. (2017), then the questions people ask in a new context should not be determined by previously encountered questions.

Differentiating these hypotheses is important because question reuse is likely to have informational consequences. In particular, if an individual continually reuses previously encountered questions in a context where these questions are no longer informative, this individual will ask many uninformative questions—preventing efficient learning. Therefore, if we find evidence that people do reuse previously encountered questions in new situations, it is also important to test (1) whether people selectively reuse questions in situations where these questions are informative and (2) how informative these reused questions actually are.

In the following section, we present an experiment that tests these predictions. We manipulate whether participants are exposed to particular questions, then we investigate what questions participants ask in a subsequent question asking task. Taken together, our results suggest that question asking is biased towards previously encountered question templates and question components, and people reuse question templates in ways that are reasonably informative. Nonetheless, question reuse can be detrimental for question informativeness in certain situations. In sum, this research sheds new light on the cognitive mechanisms behind question asking and other difficult search problems, and it provides new constraints on computational models of question asking.

Table 1: Target question templates and example "repeat questions" generated by participants.

Question Set	Question Template	Participant-Generated Example
1	At what location is the bottom right part of the *color* ship?	What is the bottom right most coordinate for the blue ship?
1	How many ships are *number* tiles long?	How many ships of 2 tiles are there?
1	Are any of the ships touching?	Do any of the ships touch?
2	How many tiles in row *number* are occupied by ships?	How many colored squares are in row 1?
2	How many ships are horizontal?	How many ships are placed horizontally?
2	Are the $color_1$ ship and the $color_2$ ship parallel?	Is the red ship parallel to purple?

Experiment

Methods

Participants We collected data from 107 adult participants from Amazon Mechanical Turk. An additional 3 participants completed the task but requested that their data be excluded. Participants were required to reside in the United States and have a minimum 95% approval rating on previous Mechanical Turk tasks. Participants were randomly assigned to one of three conditions: Question Set 1 (QS 1, n = 34), Question Set 2 (QS 2, n = 37), or Baseline (n = 36). The target sample size of 105 was determined by power analysis (see below).

Procedure The experiment involves a single-player version of the Battleship task used in past work (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014), and is adapted from the method used by Rothe et al. (2018). The goal of this task is to learn the configuration of three rectangular ships (red, blue, and purple), hidden on a 6 x 6 grid of tiles. Each ship is 2, 3, or 4 tiles in length and 1 tile in width, and is placed horizontally or vertically. Participants' task in each round is to figure out the location and size of the three ships.

First, to introduce the task, participants played five rounds of the traditional Battleship task, in which clicking on a tile reveals its contents (red, blue, purple, or water).

Then, we manipulated in the *sorting task* whether participants were exposed to particular sets of "question templates"—questions with identical form but some interchangeable content (see Table 1). These target question templates were generated with moderate frequency by participants in Rothe et al. (2018). In each round of the six-round sorting task, participants were shown a partly revealed board (see Fig. 1 for one example board). Then, participants were shown three questions—one for each question template from either Question Set 1 (QS 1 condition) or Question Set 2 (QS 2 condition)—which they sorted according to the questions'

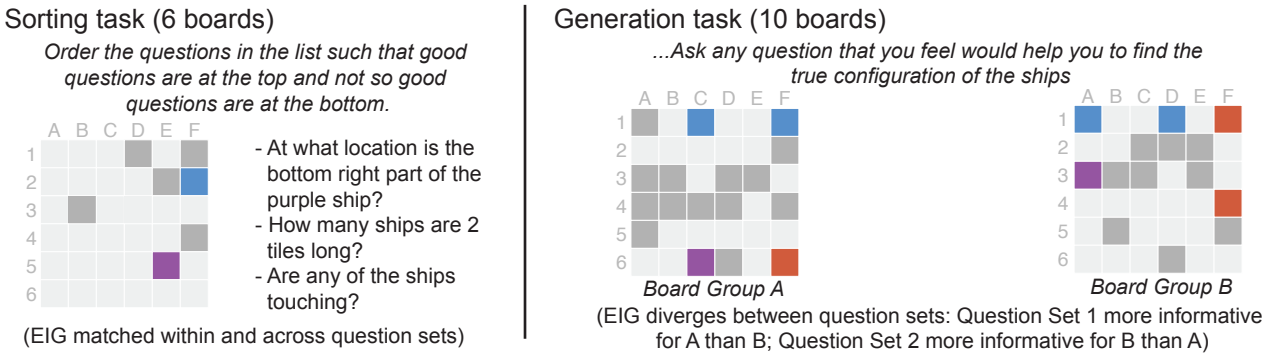

Figure 1: Experiment method. Tile color indicates a ship (blue, red, purple), water (dark gray), or a covered tile (light gray). Participants in the QS 1 and QS 2 conditions completed the sorting task (with Question Set 1 or Question Set 2, respectively) then the generation task; participants in the Baseline condition completed the generation task then the sorting task.

potential to reveal the ships' locations and sizes. In fact, the partly revealed boards in the sorting task were selected so that each question template was similarly informative when averaged across the six boards (but varied within each board). Participants were provided with the answer to their top-ranked question, then they guessed the color of any remaining hidden tiles. Participants received a bonus based on these guesses.

We then elicited questions from participants in the 10-round *generation task*. In each round, participants were again shown a partly revealed board (see Fig. 1) and were prompted to generate a question that would help them find the ships' locations and sizes. Participants were required to ask a single question that had a single answer for each round. Participants received a bonus based on how many questions followed these rules.

Participants in the Baseline condition completed the generation task followed by the sorting task, and therefore had no question exposure prior to the generation task.

Our main analyses concern whether the questions asked in the generation task differ as a function of previous question exposure (Question Set 1 in the QS 1 condition, Question Set 2 in the QS 2 condition, or no exposure in the Baseline condition). In addition, the boards in the generation task were comprised of two "board groups": Board Groups A and B. The boards were selected so that the most informative variant of each question template in Question Set 1 was higher for Board Group A than Board Group B, while the most informative variant of each question template in Question Set 2 was higher for Board Group B than Board Group A. Therefore, we can also test how questions asked in the generation task differ between these board groups.

Computational Model

We quantify question informativeness using a model adapted from Rothe et al. (2018). Formally, participants' goal is to identify a single hypothesis h that describes the true configuration of the ships, from the space of possible configurations H. Following Rothe et al. (2018), the prior is uniform over ship sizes: the size of each ship is uniformly distributed, then each configuration is uniformly distributed given those sizes. The participant can ask a question q (e.g., "Is the red ship horizontal?") and receives a response d (e.g., "yes"). The posterior distribution is then computed using Bayes' rule,

$$p(h|d;q) = \frac{p(d|q;h)p(h)}{\sum_{h' \in H} p(d|q;h')p(h')}. \quad (1)$$

Intuitively, a good question is one that, when answered, will reduce our uncertainty about the world. Formally, the Information Gain (IG) associated with an answer to a particular question is defined

$$IG(d;q) = I[p(h|d;q)] - I[p(h)], \quad (2)$$

where $I[p(h|d;q)]$ is the Shannon entropy (i.e., uncertainty; Shannon, 1948) of the posterior distribution and $I[p(h)]$ is the Shannon entropy of the prior distribution. Thus, the informativeness of an answer is the degree to which it reduces uncertainty about the true configuration of the ships.

To define the informativeness of a *question*, we must account for all its possible answers. Therefore, Expected Information Gain (EIG; Lindley, 1956; Oaksford & Chater, 1994) of a question is determined by the information gain of each answer, weighted by the answers' probability:

$$EIG(q) = \sum_{d \in A_q} p(d|q) IG(d;q) \quad (3)$$

The probability of each answer $p(d|q)$ is determined by the weighted average probability of the answer over all hypotheses, $p(d|q) = \sum_{h \in H} p(d|h;q)p(h)$.

To efficiently learn the true configuration of the ships, the optimal question is one that has the highest EIG. In our open-ended generation task, it is difficult to quantify the *most* informative question, as the space of possible questions is vast. However, we can quantify the informativeness of participants' questions using EIG, then compare question informativeness across different board groups and conditions.

Results

Of the 1070 questions asked in the generation task, we discarded 549 responses that were not questions, were off topic,

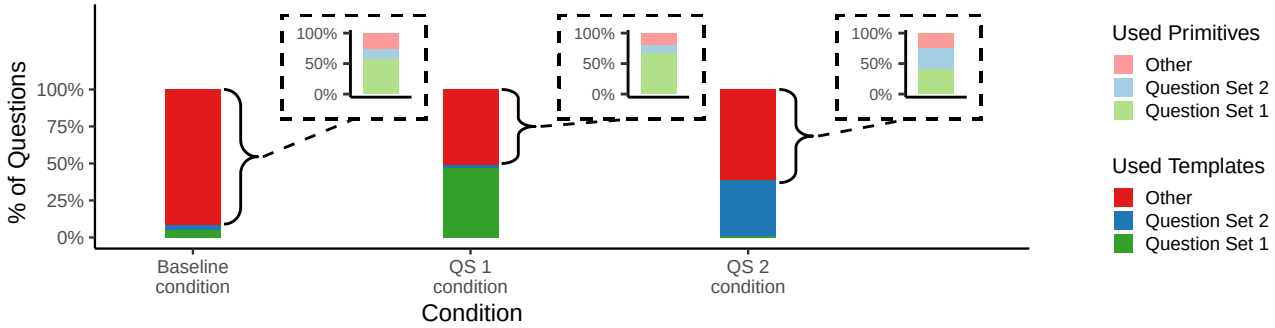

Figure 2: Percentage of questions in each condition that used question templates from Question Set 1, Question Set 2, or neither (primary plot). The questions that did not belong to either question set included questions that used primitive question components from Question Set 1, Question Set 2, or neither (inset plots).

or did not conform to the rules (e.g., "How to add shapes") and 55 questions that were ambiguous (e.g., "Row 5 contains tiles?"). An additional 16 questions were difficult to model because they referred to properties of the partly revealed board, such as which tiles had already been revealed. We excluded an additional four participants who did not ask a valid question for at least one board in each board group. This left a final sample of 445 questions, asked by 56 participants—168 questions (20 participants) in the QS 1 condition, 172 questions (22 participants) in the QS 2 condition, and 105 questions (14 participants) in the Baseline condition.

The number of excluded questions was higher than expected (see recent discussions about changes in data quality on Amazon Mechanical Turk; e.g., Kennedy et al., 2020), but we were unable to collect additional data due to practical limitations. Notably, our initial power analysis indicated 80-90% power based on the smallest effect from pilot data (the difference in EIG between Board Group 1 and Board Group 2 within the QS 2 condition), but additional power analyses using pilot data reveal at least 80% power to detect most effects of interest with our final sample size.

The 445 valid questions were coded by the first author into a question grammar developed by Rothe et al. (2017, 2018). This method allows us to (1) represent questions with different phrasing but identical meaning as the same question (see Table 1 for examples), and (2) efficiently compute EIG for each question using a Python package developed in prior research (`https://github.com/anselmrothe/EIG`; Rothe et al., 2017; Wang & Lake, 2021).

For all of the following analyses, we used mixed-effects regression models including by-participant and by-board random intercepts. To determine statistical significance, we used likelihood ratio tests.

Do people reuse questions? First, we tested whether participants asked questions during the generation task similar to those to which they were exposed. Indeed, 47% of questions asked in the QS 1 condition used Question Set 1 templates, and 38% of questions asked in the QS 2 condition used Question Set 2 templates (see Table 1 and Fig. 2; main plot). We call these "repeat questions."

We used logistic regression to test whether the use of templates from Question Set 1 and Question Set 2 varied by condition. Indeed, there was an overall effect of condition on the use of Question Set 1 templates, $\chi^2 = 28.77, p < .001$. Questions from Question Set 1 were more likely in the QS 1 condition compared to both the Baseline condition, $OR = 0.005$, 95% CI $[0.0003, 0.08]$ and the QS 2 condition, $OR = 0.002$, 95% CI $[0.0002, 0.03]$. There was also an effect of condition on the use of Question Set 2 templates, $\chi^2 = 34.24, p < .001$. Questions from Question Set 2 were more likely in the QS 2 condition compared to both the Baseline condition, $OR = 0.03$, 95% CI $[0.005, 0.17]$, and the QS 1 condition, $OR = 0.02$, 95% CI $[0.003, 0.09]$. To summarize, exposure to certain question templates increases the later use of those question templates. Notably, only 17% of repeat questions were *exact* repeats (in meaning and phrasing) of the questions seen in the sorting task, so it is unlikely that these results reflect "copy-paste" behavior.

In addition, we asked whether exposure affected the use of smaller "question primitives." For each question template, we identified one target primitive (e.g., referring to ship size for the question "How many ships are N tiles long?"), with the constraint that each target primitive was present in only one of the two question sets. This resulted in three unique primitives for Question Set 1 (bottom right corner, ship size, ships touching) and two unique primitives for Question Set 2 (row contents, ship orientation). We coded whether each question used at least one of the target primitives in each question set, and we tested whether the use of these primitives depended on condition. We excluded from this analysis any questions that matched the target question templates, so this analysis is independent of the previous analysis.

Indeed, there was a significant effect of condition on use of Question Set 1 primitives, $\chi^2 = 7.54, p = .02$, and on use of Question Set 2 primitives, $\chi^2 = 9.19, p = .01$ (see Fig. 2, inset plots). Use of Question Set 1 primitives was more likely in the QS 1 condition compared to both the Baseline condition, $OR = 0.37$, 95% CI $[0.07, 1.99]$ and the QS 2 condition, $OR = 0.11$, 95% CI $[0.02, 0.56]$, though the former effect was not statistically significant (i.e., the 95% CI included 1). Mirroring these results, use of Question Set 2 primitives

was more likely in the QS 2 condition compared to both the Baseline condition, $OR = 0.17$, 95% CI $[0.01, 1.97]$, and the QS 1 condition, $OR = 0.03$, 95% CI $[0.002, 0.39]$, though the former effect was not statistically significant. Note that each target question template is comprised of multiple primitives, and we examined only a subset. However, this analysis provides preliminary evidence that people do not solely repeat entire question templates—instead, even when generating new, never-before-seen questions, people recombine elements of previously encountered questions.

How do people reuse questions? Next, we investigate *how* people reuse question templates. Critically, the boards in the generation task were selected so that the informativeness of Question Set 1 was higher for Board Group A than B, and vice versa for Question Set 2. Therefore, we can test whether question reuse is adaptive: are people more likely to ask repeat questions in situations where those questions are informative? Indeed, there was evidence for a significant interaction between condition (QS 1, QS 2) and board group (A, B), $\chi^2 = 14.67, p < .001$ (see Fig. 3). In the QS 1 condition, participants asked more repeat questions for Board Group A (62% of questions) compared to Board Group B (32% of questions), $OR = 5.96$, 95% CI $[2.57, 13.82], \chi^2 = 12.58, p < .001$. In the QS 2 condition, participants asked more repeat questions for Board Group B (40% of questions) compared to Board Group A (35% of questions), but this difference was not significant, $OR = 1.42$, 95% CI $[0.59, 3.45], \chi^2 = 0.58, p = .44$. This provides preliminary evidence that questions are reused adaptively, though it is an open question why there was no evidence for adaptive reuse of Question Set 2.

Second, we ask how people select among candidate repeat questions. Participants in the QS 1 and QS 2 conditions were exposed to three question templates, most of which had several possible variants. Therefore, even if participants *only* generated repeat questions, they still faced the difficult task of choosing between them. We compare the informativeness of participants' repeat questions to a simple question-selection model that chooses the highest-EIG question from

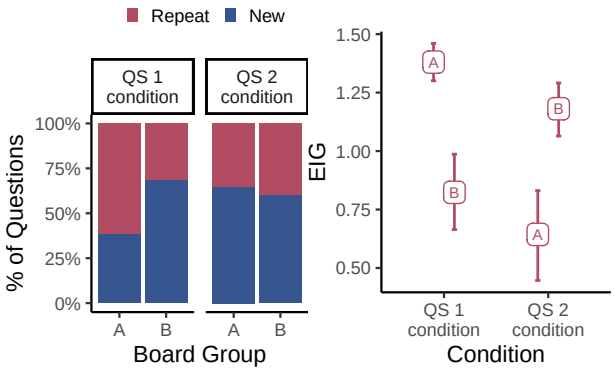

Figure 3: Left: Percentage of repeat vs. new questions for each condition and board group. Right: Average EIG (with bootstrap 95% CIs) of repeat questions asked for each condition and board group.

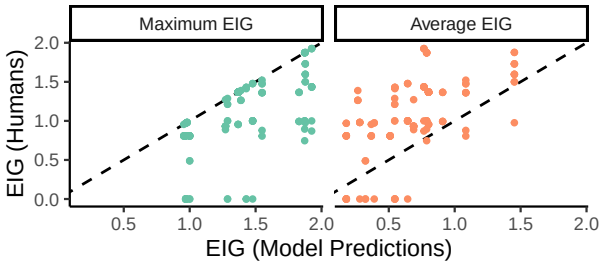

Figure 4: Comparison between repeat question EIG and predicted EIG according to two models. Most questions are less informative than the maximum EIG of all possible repeat questions (below dashed line on left), but more informative than the average EIG (above dashed line on right).

all possible repeat questions. Participants' repeat question EIG was significantly lower than model-predicted EIG, $b = -0.34$, 95% CI $[-0.42, -0.25], \chi^2 = 55.92, p < .001$. However, the model successfully explained 38% of the variance in EIG, and 19% of repeat questions exactly matched the model's predictions. Moreover, repeat question EIG was significantly higher than the average EIG of all possible repeat questions—in other words, the expected EIG if a repeat question were selected at random, $b = 0.40$, 95% CI $[0.32, 0.48], \chi^2 = 80.52, p < .001$. (see Fig. 4). Therefore, when asking a repeat question, participants appear to find a question that is reasonably (though not maximally) informative in the present situation.

What are the informational consequences of reuse? Finally, we further investigate the informational consequences of question reuse. First, we tested whether experience with the sorting task generally enhanced the informativeness of participants' questions in the later generation task. However, there was no evidence for an overall effect of condition (QS 1, QS 2, Baseline) on EIG, $\chi^2 = 2.54, p = .28$. Therefore, it appears the effect of question exposure is selective, increasing the likelihood that people ask the questions to which they were exposed—but not the likelihood that people ask good questions in general.[2]

If this is the case, we would also expect participants' questions to be more informative in some situations than others. In particular, participants in the QS 1 condition should ask more informative questions for Board Group A than Board Group B, and vice versa for participants in the QS 2 condition. Consistent with this, in a regression model predicting question EIG, there was a significant interaction between condition (QS 1, QS 2) and board group (A, B), $\chi^2 = 10.11, p = .001$. Participants in the QS 1 condition asked higher-EIG

[2]To contextualize this result, we also estimated the informativeness of the six question templates most frequently used by participants in Rothe et al. (2018). If the most informative of these questions was selected for each board in our task, the average EIG across the 10 boards is 1.43. In comparison, the average EIG for Question Set 1 is 1.34, and the average EIG for Question Set 2 is 1.36. Therefore, there is little reason to suspect that question reuse would increase the informativeness of participants' questions when averaged across the 10 boards.

questions for Board Group A ($M = 1.18$) than Board Group B (1.05), while participants in the QS 2 condition asked higher-EIG questions for Board Group B (1.05) than Board Group A (0.92). However, the within-condition differences were not significant, QS 1 condition: $b = -0.15$, 95% CI $[-0.42, 0.12], \chi^2 = 1.39, p = .24$; QS 2 condition: $b = 0.13$, 95% CI $[-0.11, 0.36], \chi^2 = 1.30, p = .25$.

Importantly, this analysis takes into account both repeat and new questions, but only the informativeness of repeat questions should vary across board groups. Indeed, for repeat questions, the interaction between condition and board group was significant, $\chi^2 = 58.32, p < .001$, and the effect of board group within each condition was also significant, QS 1 condition: $b = -0.52$, 95% CI $[-0.70, -0.35], \chi^2 = 17.47, p < .001$; QS 2 condition: $b = 0.45$, 95% CI $[0.04, 0.85], \chi^2 = 4.55, p = .03$. That is, when participants asked repeat questions, those questions were more informative in some situations than in others (see Fig. 3). These results provide evidence that question exposure has implications for later question informativeness: reusing questions is only beneficial in certain situations.

General Discussion

Questions enable us to efficiently learn. However, because the space of possible questions is too large to search exhaustively or at random, generating and selecting an informative question is a difficult challenge. By investigating how prior exposure to questions influences the questions people ask in new situations, the present research tested unexplored mechanisms behind question asking.

We compared the questions asked by participants who were exposed to two distinct sets of questions, as well as participants who had no exposure. This led to three main findings. First, participants frequently repeated the questions to which they were exposed, and even novel questions tended to reuse components of the exposure questions. Second, when participants asked repeat questions, they successfully found questions that were reasonably (but not maximally) informative in the present context. However, third, question reuse had informational consequences: the informativeness of participants' questions depended on the situation. In particular, participants asked more informative questions in situations where previously encountered questions were still useful.

These results provide new insight into how humans succeed in asking informative questions (Callanan & Oakes, 1992; Chouinard, 2007; Rothe et al., 2018). Rather than searching through the entire space of possible questions in each situation, we can constrain our search using the questions we've encountered in the past, as we do for other decision problems (Bear et al., 2020; Morris et al., 2021; Phillips et al., 2019). This has important implications for computational models of question asking, which have instead modeled question generation as random search through the entire space of questions (Rothe et al., 2017). Future research might modify these models to incorporate a search mechanism that prioritizes previously asked questions.

This work also raises a number of new questions. For example, does exposure to questions influence question generation or question evaluation? Prior research on other decision problems (Morris et al., 2021) has shown that the prior quality of a solution influences the later generation of that solution, but not its evaluation. Consistent with this, we found that participants asked repeat questions that were reasonably informative, indicating the questions were evaluated and selected at least partly by their current informativeness rather than their previous use. However, further research is needed to cleanly separate question generation and evaluation.

Second, how pervasive are the effects of question exposure? In the present research, we demonstrated that people reuse previously encountered questions in new situations, but within the same general task (i.e., the Battleship game) and in a single experimental session. Would question reuse extend across tasks and across time? Relatedly, we found only modest evidence for "adaptive" reuse: selectively reusing questions in situations where those questions are informative. Further research is needed to determine what features of the current situation determine whether past questions are asked.

Several limitations of this work must be noted. First, an unexpectedly large proportion of our participants asked invalid questions, and therefore, it is likely that many of our participants were not paying close attention to the task. Notably, our results remain consistent even when limiting analyses to the 30 participants who asked a valid question for all 10 generation task rounds. Nonetheless, it is possible that more attentive participants would be more or less likely to use repeat questions, and therefore further research is needed to test the extent to which our results generalize to different samples.

In addition, we investigated how exposure to a small set of questions affects question asking in a small number of situations. This allowed us to test precise predictions about the informativeness of repeat questions in different situations. However, questions asked in these situations might not be representative of questions asked across all situations—and our task and model certainly do not capture all contexts in which questions are asked in everyday life. For example, we did not take into account non-informational motives for asking questions (Hawkins et al., 2015; Markant & Gureckis, 2012; Meder & Nelson, 2012; Rothe et al., 2018). Therefore, it remains to be tested how exposure to questions impacts question generation in a range of settings, for a range of informational and non-informational goals.

Despite these limitations, the present research provides new insight into the cognitive mechanisms that enable question asking. Rather than generating questions from scratch in each new situation, as existing models of question asking predict, people (sometimes) draw upon previously used questions. It is likely that question reuse provides a computationally efficient means of generating reasonable questions in novel situations—though this has consequences for how informative our questions are likely to be.

Acknowledgements

We thank Angela Radulescu and Pat Little for their helpful comments on an earlier draft, and we thank the anonymous reviewers for their valuable feedback. This work was supported by the James S. McDonnell Foundation (Scholar Award to T.M.G.) and National Science Foundation grants #2121102 and #2021060.

References

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, 104057.

Birbili, M., & Karagiorgou, I. (2009). Helping children and their parents ask better questions: An intervention study. *Journal of Research in Childhood Education*, *24*(1), 18–31.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, *72*(1), 1–129.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, *42*(5), 1410–1456.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Du, X., Shao, J., & Cardie, C. (2017). *Learning to Ask: Neural Question Generation for Reading Comprehension.* (https://arxiv.org/abs/1705.00106v1)

Good, T. L., Slavings, R. L., Harel, K. H., & Emerson, H. (1987). Student passivity: A study of question asking in K-12 classrooms. *Sociology of Education*, *60*(3), 181–199.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104–137.

Gureckis, T. M., & Markant, D. B. (2009). Active learning strategies in a spatial concept learning game. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 3145–3150). Cognitive Science Society.

Hawkins, R. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. In D. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 878–883).

Jain, U., Zhang, Z., & Schwing, A. G. (2017). Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6485–6494).

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629.

King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal*, *27*(4), 664–687.

King, A. (1991). Effects of training in strategic questioning on children's problem-solving performance. *Journal of Educational Psychology*, *83*(3), 307–317.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Markant, D. B., & 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* (pp. 719–724). Cognitive Science Society.

Markant, D. B., & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 958–963). Cognitive Science Society.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, *32*(11), 1731–1746.

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

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience Matters: Information Acquisition Optimizes Probability Gain. *Psychological Science*, *21*(7), 960–969.

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

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556.

Ronfard, S., Zambrana, I. M., Hermansen, T. K., & Kelemen, D. (2018). Question-asking in childhood: A review of the

literature and a framework for understanding its development. *Developmental Review*, *49*, 101–120.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*(2), 181–221.

Rothe, A., Lake, B. M., & Gureckis, T. (2017). Question asking as program generation. In *Advances in neural information processing systems* (pp. 1046–1055).

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2019). Asking goal-oriented questions and learning from answers. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 981–986). Cognitive Science Society.

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Ten, A., Kaushik, P., Oudeyer, P.-Y., & Gottlieb, J. (2021). Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, *12*(1), 5972.

Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 2686–2697). Curran Associates, Inc.

Wang, Z., & Lake, B. (2021). Modeling Question Asking Using Neural Program Generation. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 2595–2601). Cognitive Science Society.