Computational Brain & Behavior (2018) 1:69–89
https://doi.org/10.1007/s42113-018-0005-5

Do People Ask Good Questions?

Anselm Rothe[1] · Brenden M. Lake[1,2] · Todd M. Gureckis[1]

Published online: 9 July 2018

Abstract
People ask questions in order to efficiently learn about the world. But do people ask good questions? In this work, we designed an intuitive, game-based task that allowed people to ask natural language questions to resolve their uncertainty. Question quality was measured through Bayesian ideal observer models that considered large spaces of possible game states. During free-form question generation, participants asked a creative variety of useful and goal-directed questions, yet they rarely asked the best questions as identified by the Bayesian ideal observers (Experiment 1). In subsequent experiments, participants strongly preferred the best questions when evaluating questions that they did not generate themselves (Experiments 2 and 3). On one hand, our results show that people can accurately evaluate question quality, even when the set of questions is diverse and an ideal observer analysis has large computational requirements. On the other hand, people have a limited ability to synthesize maximally informative questions from scratch, suggesting a bottleneck in the question asking process.

Keywords Question asking · Question generation · Information search · Active learning · Bayesian modeling

Introduction

Asking questions is a hallmark of human intelligence which enables us to flexibly learn about, navigate in, and adapt to our environment. For example, a simple question to a fellow traveler (“Where is the uptown train platform?”) can save us from wandering around until we find our way. Similarly, a doctor asking a patient “Have you traveled internationally in the past 21 days?” might be able to rule out a large number of exotic diseases if the answer is “no.” Questions also are important in course of cognitive development. For example, Nelson (1973) found that most children acquire an utterance for asking questions within their first few words (e.g., “Eh?” or “Doh?” to mean “What is that?”). Such words may actually help bootstrap the process of learning language by coordinating information requests between the child and caregiver (Nelson 1973). There is little doubt then that asking questions is a powerful cognitive and linguistic tool, but are people actually effective at asking questions? Specifically, given all the possible questions someone could ask in a given situation, do people generally ask the best or most informative questions?

In recent decades, there has been a growing scientific interest in how children and adults ask questions. Much of the past work on this topic is largely qualitative, essentially cataloging the different types of questions people ask in different situations. The quantitative work on this topic, which has the potential to objectively assess the *quality* of human questions, has tended to focus on relatively simple scenarios where the range of allowed questions is limited to the features and class membership of each object (e.g., in the “Guess Who?” game, a question might be “Is your person wearing a hat?”). As a result, it is unclear whether people ask objectively good (or maximally informative) questions given their knowledge and goals in more unconstrained scenarios such as those encountered in everyday life.

In this paper, we attempt to explore this issue in a novel way using a relatively unconstrained question asking task which is nonetheless amenable to computational analysis. After reviewing the past literature on question asking and active learning, we describe the task we used to study question asking. Next, we describe alternative models of question evaluation that allow us to objectively measure the quality of the questions people asked in the experiment.

✉ Anselm Rothe
anselm@nyu.edu

[1] Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA

[2] Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA

We then report empirical results from three experiments in which people either asked questions in order to resolve some ambiguous situation, or evaluated the questions generated by other people. To foreshadow, our results highlight interesting limitations in the intuitive question asking abilities of humans which we argue results, in part, from the immense computational demands of optimal question asking.

Past Work on Question Asking

The way people ask information-seeking questions has attracted considerable attention in cognitive science in both laboratory studies and observational designs.

Qualitative Studies of Question Asking

Studies in psychology and education investigating human question asking in classroom settings have focused on rather general distinctions between good and bad questions (see Graesser et al. 1993; Graesser and Person 1994; Davey and McBride 1986; Chin and Brown 2002; Dillon 1988). For instance, Chin and Brown (2002) distinguished basic information questions from "wonderment" questions, characterized as deeper questions used for planning or making predictions. In a reading comprehension study, Davey and McBride (1986) judged participants' questions as better if the questions targeted central ideas, used a "wh" stem, and required more than a yes/no response. Building on a classification scheme of questions that children ask in the classroom (Graesser et al. 1992), Graesser and colleagues defined good questions as having a question type that fits to the type of knowledge structure that the questioner wants to learn about. If the questioner was learning about a taxonomic structure of musical instruments, then "What are the types of X?" (e.g., "To which groups does a soprano clarinet belong?") constituted a good question while "How can you create X?" (e.g., "How do you build a soprano clarinet?") was a bad one because it targeted an answer unlikely to be informative about the taxonomic structure (Graesser et al. 1993). In subsequent work, Graesser and colleagues defined some of their categories as "deep" (e.g., why, why not, how, what-if, what-if-not) in contrast with "shallow" questions (e.g., who, what, when, where) and found that the proportion of deep questions asked in a class correlated with students' exam scores (see Graesser and Person 1994).

Although these studies offer interesting insight into the causes and consequences of question asking, they leave unresolved if any *particular* question is informative from the perspective of an individual learner. For example, even "deep" questions can be uninformative if the learner already knows the answer. This is largely because in observational classroom studies, it is difficult to measure and control the amount of knowledge that different learners have as well as to account for differing individual goals during learning. In this paper, we focus on the quality of questions from an individual's perspective by controlling the background and prior knowledge that participants had about our experimental task as well as their goals.

Quantitative Studies of Question Asking

Although the above studies often focus on question asking in natural settings such as a classroom, quantitative studies often rely more heavily on laboratory tasks. Such tasks typically create a scenario (such as a game) where a learner must ask questions or perform certain actions in order to acquire information (e.g., Coenen et al. 2015; Markant and Gureckis 2014a, b; Meder and Nelson 2012; Nelson et al. 2014; Ruggeri and Feufel 2015; Ruggeri et al. 2016).[1] The key concern in this work is if children and adults ask the "best" or most informative question as measured by computational models.

In order to apply computational models to this data, often these experiments are simplified relative to real-life inquiry. One view of the purpose of question asking is to resolve between alternative world states. For instance, when you ask a question like "Is this the uptown train platform?" you are attempting to resolve between two hypothetical worlds, one where you are standing on the uptown train and one where you are not. The answer to the question helps to resolve that ambiguity enabling you to act accordingly. Good questions are those that, from the perspective of the individual, rule out alternative world states.

In most laboratory tasks that try to mimic these real-life situations, the space of possible hypotheses (or alternative world states) that the learner is trying to discriminate is relatively curtailed, ranging from around 20 hypotheses (Ruggeri and Feufel 2015; Ruggeri et al. 2016; Nelson et al. 2014) to as few as two (Meder and Nelson 2012; Coenen et al. 2015; Markant and Gureckis 2014a). Similarly, many tasks allow only yes/no questions, which constrains the size of the set of answers (e.g., Ruggeri et al. 2016 only allowed yes/no questions but provided "some"

[1] In some studies, participants performed information-seeking *actions*, such as clicking on a certain part of an object, to obtain information about the object, which is, for our purposes, equivalent to asking information-seeking *questions*. For instance, participants could click on either the eye or claw of a plankton creature presented on a computer screen, to reveal the eye/claw color and then categorize the plankton based on that information (Meder and Nelson 2012), which is equivalent to asking "What is the color of the eye/claw?" Similarly, in Coenen et al. (2015), participants could click on one of three nodes in a causal network and subsequently observe which of the other nodes would turn on, which is equivalent to asking "Which nodes will turn on when I activate this node?"

as an additional, third answer; Coenen et al. 2015 had as many as four possible nodes or components that could be intervened on one at a time). Finally, a common strategy in the laboratory literature has been to effectively provide people with a predetermined list of questions allowing them to select the best (e.g., Nelson et al. 2014; Coenen et al. 2015; Meder and Nelson 2012). Although this approach simplifies data analysis considerably, it relieves learners from the burden of generating interesting and informative questions from scratch. As we highlight in the experiments below, the distinction between question *generation* and *evaluation* is a psychologically significant part of the task of asking question in more complex tasks and everyday life.

One notable exception to this trend is the work by Ruggeri and colleagues, who formally analyzed the question quality of relatively unconstrained yes/no questions that adults and children generated in order to identify a target category (Ruggeri et al. 2016). They reported relatively high performance by adults, who on average asked a first question that was very close to optimal. Our experiments similarly examine open-ended question asking performance, but with a much broader range of questions and a more complex task.

In summary, past work has tended to organize into observational studies of real-world question asking, where the issue of question quality is addressed relatively qualitatively, or careful laboratory settings, which are more simplified but allow precise measurement of the information value of different queries. The goal of the present study is to combine elements of both traditions by studying question asking in a rich and unconstrained setting that is nonetheless amenable for formal mathematical modeling.

Studying Question Asking in the Battleship Game

In light of the issues laid out above, we identified a few key features for studying question asking in an experimental setting. First, we wanted to provide participants with ambiguous situations, in which they can ask a variety of questions with the goal of resolving the ambiguity. Second, we wanted participants to share the same understanding of what that ambiguity is. Thus, the situations should be defined by instructions that are easy for people to understand (e.g., as part of an intuitive task). Third, we wanted situations that are amenable to formal modeling, that is, constrained enough such that all possible ways to resolve the ambiguous situation can be represented in a mathematical model.

These features are ideally captured by an active learning task that is called the Battleship game due to its similarity to a single-player version of the popular children's game (Gureckis and Markant 2009; Markant and Gureckis 2012, 2014b). The goal of the game is to determine the location and size of three non-overlapping rectangular "ships" on a 6×6 grid (see Fig. 1). The ships have a width of 1 tile, are 2 to 4 tiles long, and are horizontally or vertically

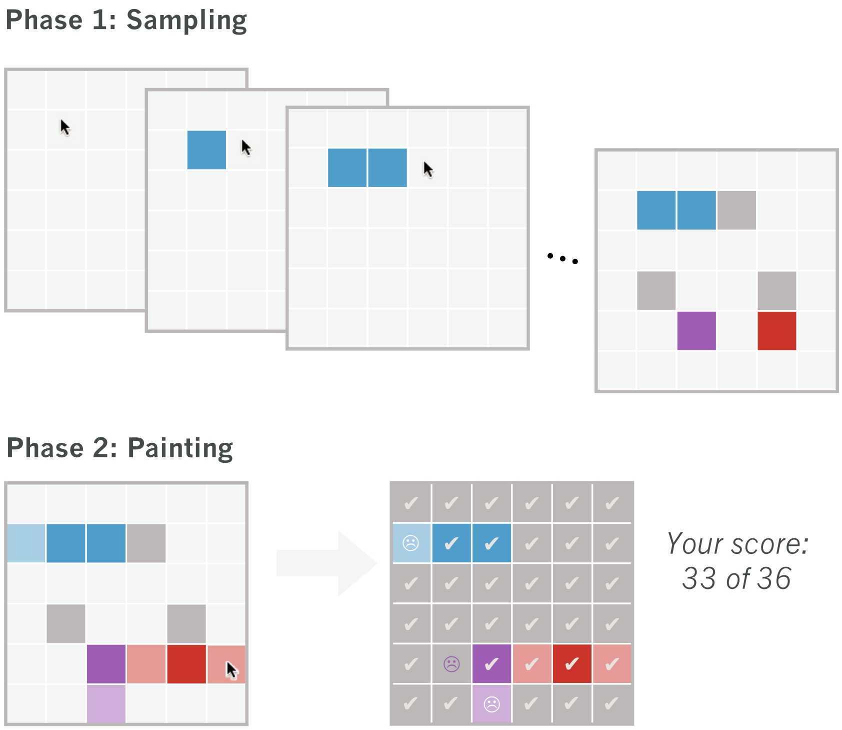


Fig. 1 Battleship game boards as viewed by participants. *Sampling phase*: A participant sequentially clicks on tiles to turn them over. The revealed color indicates a ship (blue, red, or purple) or water (dark gray). *Painting phase*: At a certain point, the sampling phase is stopped and the participant guesses the color of the remaining tiles. For each correctly painted tile, one point is awarded

oriented. In past work that has used this information search task, a participant sequentially clicked on tiles to uncover either the color of the underlying ship part or an empty water part (*sampling phase*, Fig. 1). An efficient active learner seeks out tiles that are expected to reduce uncertainty about the ship locations and avoids tiles that would provide redundant information (e.g., when the hidden color can be inferred from the already revealed tiles). At a certain point, the sampling was stopped and participants were asked to fill in the remaining tiles with the appropriate color, based on their best guess (*painting phase*, Fig. 1). The score they receive was a decreasing function of the number of observations made in the sampling phase and an increasing function of the number of correctly painted tiles.

The task is well suited for the present study because the underlying hypothesis space (i.e., possible ship configurations that define a possible gameboard) is relatively large (1.6 million possible game boards) but is easy to explain to participants prior to the start of the task. In addition, the game is interesting and fun for participants while being amenable to an ideal observer analysis (see below). In previous work using this task, the only means a participant has for acquiring information is by sampling or turning over individual tiles (Settles 2009; Markant and Gureckis 2012, 2014b). In this paper, we allow participants to ask any question they want in natural language (e.g., “Are the ships touching?” or “What is the total area of the ships?”). This modification allows participants to use much more powerful tools to gain information (e.g., general purpose question asking) and allows us to study rich, natural language question asking in the context of a well-understood active learning task. Importantly, the design implied that people were conversing with an English-speaking oracle who knew the true hidden gameboard and would always answer honestly, similar to the assumption that would apply to clicking a tile to uncover it.

The Battleship domain has also proved useful to study how a machine can generate informative, human-like questions from scratch (Rothe et al. 2017).

Defining the Computational Problem of Asking Good Questions

A computational-level analysis describes, in a formal way, the goals and constraints of computations relevant for the cognitive system to solve a task (Marr 1982; Anderson 1990). Let us begin with the very general notion that any question we might ask has a certain utility with respect to our goals. This idea can be formalized as the *expected utility* of question x,

$$\mathrm{EU}(x) = \mathbb{E}_{d \in A_x}[U(d; x)] \tag{1}$$

where d is an answer from the set of possible answers A_x to question x, and $U(d; x)$ is the utility of that answer for that question. Using an *expectation* is required here given that, at the time of asking, we do not know yet what the answer to the question is going to be. Under this framework, the task of asking the best question x^* can then be cast as a search over the set Q of all possible questions one could ask,

$$x^* = \arg\max_{x \in Q} \mathrm{EU}(x). \tag{2}$$

The utility of a question $U(d; x)$ can include a range of factors depending on the agent’s goals and situation. For example, question asking is often a social activity implying a listener and speaker, and assumptions about each along with the limited bandwidth of spoken communication may enter into the calculation of utility. For instance, research has shown that a learner may consider how difficult a question is to answer or how much information is conveyed to the answerer about one’s intentions (e.g., Clark 1979; Hawkins et al. 2015). Any of such social and pragmatic considerations can be included as factors in the computation of a question’s utility $U(d; x)$.

Another issue concerns how much the utility of a question is influenced by the costs and rewards of the current task. For example, Chater, Crocker, and Pickering distinguish between cost-insensitive (or “disinterested”) utilities and cost-senstive (or “interested”) utilities (Chater et al. 1998; Markant and Gureckis 2012). A cost-insensitive, information-maximizing utility values knowledge by itself without reference to task-specific rewards, as for example reflected in the spirit of basic research or genuine curiosity. For instance, attempting to reduce one’s uncertainty about the position of the ships in the Battleship game as much as possible follows this strategy and is captured by the *expected information gain* (EIG) model (see below).

In contrast, a cost-sensitive, utility-maximizing strategy values information only to the degree that it will lead to later rewards or avoid costs, making it a more economically oriented strategy. For example, students who want to minimize their study time and decide to ignore information that is unlikely to be tested in an exam engage in such a strategy, where the cost structure is given by the time spent studying as well as points lost in the exam. In the Battleship game, one utility-maximizing strategy is captured by the *expected savings* (ES) model (see below), which evaluates information with respect to the errors that can be avoided in the painting task.

In the present paper, we compare these two ways of assigning utility to a question (ES versus EIG). The two models provide alternative “yardsticks” for objectively evaluating the quality of people’s questions with respect to the goals of the task. To give an intuitive example, EIG assigns a high value to a question such as “How many tiles are occupied by ships?” because every answer allows the learner to rule out many hypothesized ship configurations that are inconsistent with the obtained answer. On the other hand, ES assigns a low value because such abstract information about the number of ship tiles does often not help much with the painting task.

The EIG versus ES contrast is also interesting because past work found that people’s strategies were more in line with the cost-insensitive EIG than the cost-sensitive ES model (Markant and Gureckis 2012). Before defining these models formally, we introduce a Bayesian ideal observer of our task, which forms the foundation of both measures of utility for question asking.

Bayesian Ideal Observer Analysis of the Battleship Game

In a given Battleship game context, a player aims to identify a hidden configuration corresponding to a single hypothesis h in the space of possible configurations H. We model her prior belief distribution over the hypothesis space, $p(h)$, as uniform over ship sizes. The prior is specified by first sampling the size of each ship from a uniform distribution and second sampling uniformly a configuration from the space of possible configurations given those sizes. The player can make a query x (uncovering a tile or asking a natural language question) and receives the response d (the answer). The player can then update her posterior probability distribution over the hypothesis space by applying Bayes’ rule,

$$p(h|d; x) = \frac{p(d|h; x)p(h)}{\sum_{h' \in H} p(d|h'; x)p(h')}. \tag{3}$$

The semi-colon notation indicates that x is a parameter rather than a random variable. The posterior $p(h|d; x)$ becomes the next step’s prior $p(h|D; X)$, with X representing all past queries and D representing all past responses,

$$p(h|d, D; x, X) = \frac{p(d|h; x)p(h|D; X)}{\sum_{h' \in H} p(d|h'; x)p(h'|D; X)}. \tag{4}$$

The likelihood function $p(d|h; x)$ models the oracle that provides answer d. The likelihood is zero if d is not a valid response to the question x, and $\frac{1}{n}$ otherwise, where n is the number of correct answers that the oracle chooses from uniformly. For most questions that we collected, there was a single correct answer, $n = 1$. But for example when asking for the coordinates of any one of the tiles that contain a blue ship, n is defined by the number of blue ship tiles in the true configuration. The posterior predictive value of a new query x resulting in the answer d can be computed as

$$p(d|D; x, X) = \sum_{h \in H} p(d|h; x)p(h|D; X). \tag{5}$$

Cost-Insensitive Utilities: Expected Information Gain

A primary goal of most question asking is to gain *information* about the world relevant for the learner’s goal. In this case, the utility of a question and its answer is entirely or at least heavily determined by their expected informativeness. More precisely, we can define the utility of a question and its answer as the amount of gained information,

$$U(d; x) = I[p(h|D; X)] - I[p(h|d, D; x, X)], \tag{6}$$

where $I[\cdot]$ is the Shannon entropy (uncertainty, Shannon 1948) of the previous belief distribution $p(h|D; X)$ (i.e., before receiving an answer d to the new query x, but with prior knowledge D and X; Eq. 4) and of the posterior belief distribution $p(h|d, D; x, X)$ (i.e., after receiving an answer d to question x) over the hypothesis space.

For illustration, consider the example of asking a friend about the location of your car keys. The hypothesis space, H, spans the space of possible locations you can consider (e.g., locations in your apartment). Assuming there have been no previous questions X and answers D, $p(h)$ represents your subjective belief in these possible locations (e.g., twice as likely to be in the office than the kitchen), and the entropy, $I[p(h)]$, indicates how uncertain you are about the key location (this scalar measure will be zero when you know its location and will be high if you assign equal belief to each possibility). Likewise, $I[p(h|d; x)]$ is the uncertainty about the key location after the answer is revealed. Some answers d can be far more informative than others. For example, imagine asking your friend, “Where are my car keys?”. The answer “Somewhere in your apartment” is intuitively less informative than the more precise answer “On your desk, to the left of your laptop.”

Of course, when asking a question such as “Where are my car keys?”, we do not yet know which answer we will receive, but under a computational analysis, we can simulate how our uncertainty would hypothetically be reduced for each possible answer. Combining Eqs. 1 and 6, we define the expected utility of question x as the *expected information gain* (EIG),

$$\begin{aligned} \mathrm{EU}(x) := \mathrm{EIG}(x) &= \mathbb{E}_{d \in A_x}\left[I[p(h|D; X)] - I[p(h|d, D; x, X)]\right] \\ &= \sum_{d \in A_x} p(d|D; x, X)\left[I[p(h|D; X)] \quad - I[p(h|d, D; x, X)]\right] \end{aligned} \tag{7}$$

or the average amount of information we can expect from each of the possible answers to the question (e.g., Oaksford and Chater 1994; Coenen et al. in press). EIG is a commonly used metric in machine learning approaches to active learning (Settles 2012) and has a long history of study as a model of human information gathering (Oaksford and

Chater 1994). The value of a query is measured by EIG in the unit of bits.

Assuming the learner is motivated to quickly gain information, it would make sense for people to optimize information with their questions. That is, out of the possibly infinite set of questions, Q, we want to find the optimal question, $x \in Q$, that maximizes the expected information (see Eq. 2).

Cost-Sensitive Utilities: Expected Savings

According to ES, a query x is valued according to the expected rewards or avoided costs that the learner is expected to accrue after learning the answer to the question. For example, some questions might have high utility not just because they convey information but because they allow the agent to act effectively in the world (e.g., "Is this mushroom poisonous to eat?"). Cost-sensitive utilities are highly task-dependent in the sense that they depend on what the question asker plans to do with the acquired information (asking about the safety of a poisonous mushroom is not useful if you never planned to eat it, in which case the same question is essentially trivia). As a result, these utilities are defined differently for almost all tasks and goals the learner might have (unlike cost-insensitive utilities, which only depend on what the learner knows).

In the case of the Battleship, the primary goal might be reducing errors in the painting task (Fig. 1), leading to the following utility function,

$$U(d; x) = \mathrm{EC}[p(h|D; X)] - \mathrm{EC}[p(h|d, D; x, X)]. \tag{8}$$

The function $\mathrm{EC}[p(h|v)]$ is used to denote the expected cost when coloring tiles in the painting task according to a particular belief distribution $p(h|v)$, using v as shorthand for past questions and responses. Expected cost is defined as

$$\mathrm{EC}[p(h|v)] = \sum_i \sum_l p(l|v; i) \times [C_{\text{hit}}\, p(l|v; i) + C_{\text{miss}}(1 - p(l|v; i))], \tag{9}$$

where the belief that tile i has color l is given by $p(l|v; i) = \sum_{h \in H} p(l|h; i) p(h|v)$. The choice to actually paint the tile in that color is here given by $p(l|v; i)$ again because we assume during the painting phase participants will use a probability matching decision strategy to choose the color of each tile. $C_{\text{hit}} = 0$ and $C_{\text{miss}} = 1$ indicate the costs associated with painting a tile correctly or incorrectly, respectively.

As with EIG, the question asking agent does not know the answer d to question x in advance. We define the expected savings (ES) as the expected reduction of errors in the painting task averaged across all possible answers A_x of the query

$$\mathrm{EU}(x) := \mathrm{ES}(x) = \sum_{d \in A_x} p(d|D; x, X) \Big[\mathrm{EC}[p(h|D; X)] - \mathrm{EC}[p(h|d, D; x, X)]\Big]. \tag{10}$$

Thus, in the Battleship game, ES measures a query's value in units of expected (average) number of correctly painted tiles. As above, we want to find the optimal question, $x \in Q$, that maximizes the expected savings. Note that because of the conjunctive consideration of hypotheses and actions (painting tiles, Eq. 9), ES cannot be recast as a mere weighted variant of EIG, which only concerns hypotheses.

Example Calculation and the Computational Challenge of Question Asking

To make the equations just described more concrete, consider the computations involved in finding your lost keys by asking questions according to EIG. If, based on your prior knowledge $p(h)$, the kitchen is an unlikely place for the keys, then "Are my car keys in the kitchen?" is an intuitively uninformative question. According to the analysis above and Eq. 7, we would evaluate for each of the two possible answers to this question, $A_x = \{$"yes,""no"$\}$, how much our uncertainty would be reduced. If the answer is $d =$ "no" (which is very likely the answer), then our uncertainty about the location did not change much relative to what we believed already. If the answer is $d =$ "yes," then we would rule out most of the other locations in our apartment, accordingly update our belief $p(h|d; x)$, and we could see that our uncertainty $I[p(h|d; x)]$ is much smaller than before. Overall, since $d =$ "yes" is unlikely given that the kitchen is an unlikely place for the keys, the expected information gain of the question is low, as captured by the weighted expectation $\mathbb{E}_{d \in A_x}$.

Although it might seem straightforward to simply ask "Where are my car keys?" given the goal of finding them, there are a myriad of questions one could ask instead (e.g., "What room did you go to after you parked my car?," "Where do you usually put the keys?," "Are the keys in the living room?," etc.). Each of these has a slightly different semantic meaning and provides additional information or leaves additional ambiguity depending on our goal.

This computational analysis of question asking provides one way to formalize the notion of a "good question." However, it raises interesting computational challenges. For example, the computations under this information-theoretic model grows with the number of hypotheses (computing $I[p(h)]$ and $I[p(h|d; x)]$), the number of possible answers (computing $\mathbb{E}_{d \in A_x}$), and the size of the question space

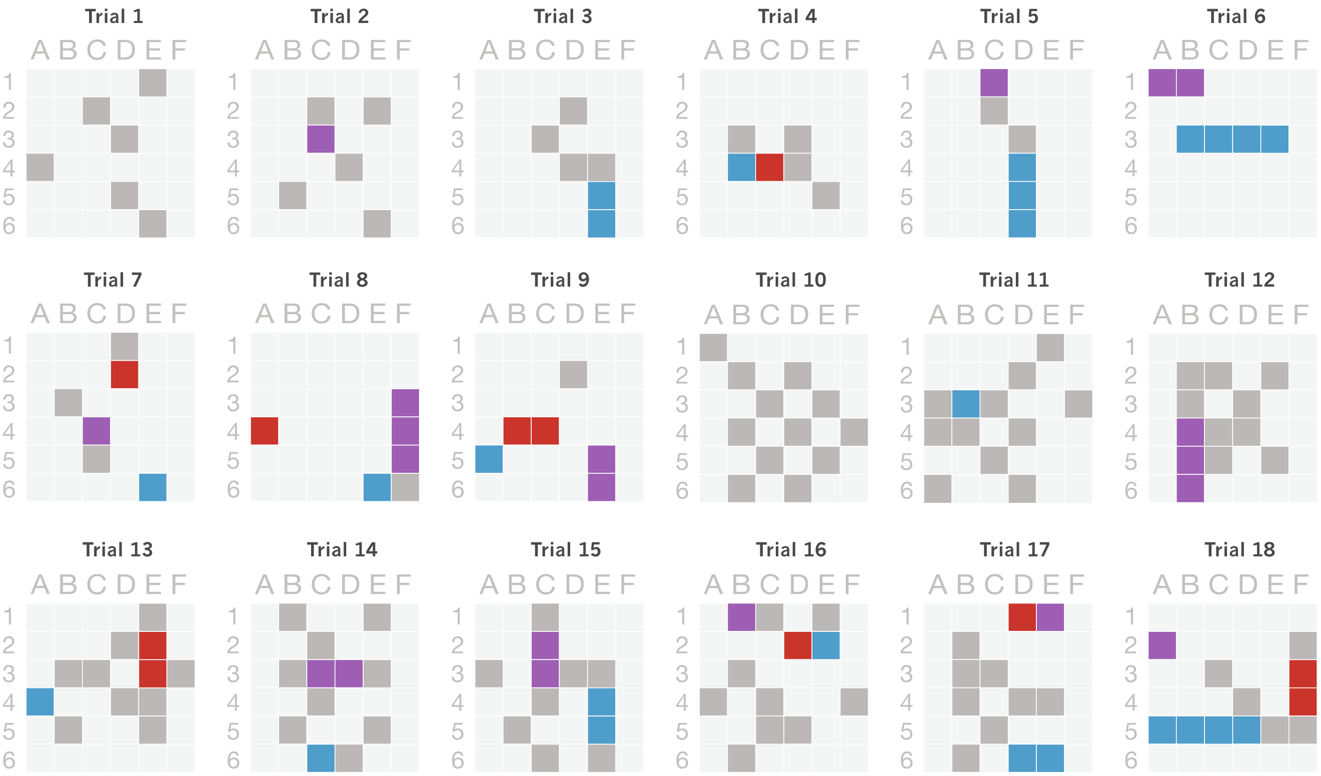


Fig. 2 Game boards shown to participants when eliciting questions. Each board is a partially revealed state of the game

(scoring each EIG(x) for $x \in Q$). Moreover, if Q is defined in terms of natural language, as is often the case in naturalistic inquiry, it is hardly a small enumerable set of possibilities. Instead, it is a very large (possibly infinite) set of options, each with a complex internal structure that determines how the question x when applied to state h determines a distribution on answers d. (Note, the same argument applies to ES and any other model that seeks to maximize an expected utility via Eqs. 1 and 2.) In the methods and analysis of Experiment 1, we highlight some of the complexities and challenges in characterizing both the question space Q and particular questions x. In our case, we formalize questions x as functions that can be applied to a state h to return d. All things considered, it is surprising that people can ask questions at all given the immense computational demands this behavior implies for most realistic situations (see Coenen et al. in press, for a full discussion). The apparent difficulty of creating a tractable computational account of optimal question asking raises the core focus of the current study: Are people actually are good or effective question askers?

Experiment 1—Question Generation

There is an infinite number of questions that can be asked in any situation. However, most of them would have little or no information value while others would be highly informative. Our first experiment explored how people generate free-form, natural language questions in a modified version of the Battleship game. After familiarizing participants with the mechanics and parameters of the game, we collected the questions they asked to resolve ambiguity about the positions of ships on a game board. Our ultimate goal is to relate open-ended natural language questions to the models of information utility just described.

Participants

Forty participants recruited on Amazon Mechanical Turk using *psiTurk* (Gureckis et al. 2016), with restriction to the US pool, were paid a base of $2 with a performance-based bonus of up to $3.60. Participants were awarded a bonus of $0.20 for each generated question that was in line with the task rules, encouraging a minimum level of question quality without providing monetary incentives for especially rich and creative questions.[2]

[2] We decided against paying people based on question quality. Participants would have to reason about what we, the experimenters, expect to be good questions.

Method

Before eliciting the natural language questions, we took a number of steps to help the participants understand the task. These included detailed tutorial-like instructions that explained the task and comprehension quizzes to verify understanding. The comprehension quizzes had questions such as "How many ships can appear on the grid?" and "You will never see ships on the grid that..." and participants had to select the correct multiple-choice answer (here, "Exactly 3" and "...have the same color," respectively). In addition, key task information about the number and the possible colors as well as the possible sizes and orientations of the ships remained visible on a side panel throughout the whole experiment.

In a warm-up phase, participants played five rounds of the standard Battleship game to ensure understanding of the basic game play. Each warm-up round included sampling (i.e., clicking on tiles to find the ships) followed by painting (i.e., guessing the color of the remaining tiles; see Fig. 1). Then, in the main phase, participants were given the opportunity to ask free-form questions in the context of 18 partly revealed game boards, presented in the order shown in Fig. 2. To produce a variety of different types of partial knowledge states, we varied the number of uncovered tiles (6 or 12), the number of partly revealed ships (0 to 3), and the number of fully revealed ships (0 to 2). These factors were varied independently while excluding impossible combinations leading to a total of 18 contexts.

Figure 3 shows the procedure of a trial in the main phase. At the beginning of a trial, we introduced participants to a context by requiring them to click on a pre-determined sequence of tiles (which are the past queries X and answers D in Eq. 4). We chose this format of tile-uncovering moves, resembling the warm-up phase, to give the impression that a human was playing a game that was paused in an unfinished state. Subsequently, as a comprehension check, participants were asked to indicate the possible colors of each covered tile subject to the rules of the game (e.g., whether the tile could plausibly still be hiding a piece of the red ship, see Fig. 3, *Ship indication*). The task would only continue after all tiles were indicated correctly (or a maximum of six guesses were made).

Next, participants were given the following prompt: "If you had a special opportunity to ask any question about the grid, ships, or tiles, what would you ask?" (represented as x in Eq. 4). A text box recorded participants' responses. The only two restrictions were that combinations of questions were not allowed (i.e., putting two questions together with "and" or "or") and questions had to be answerable with a single piece of information (e.g., a word, a number, true/false, or a single coordinate). Thus, participants could not ask for the entire latent configuration at once, although their creativity was otherwise uninhibited. Due to practical

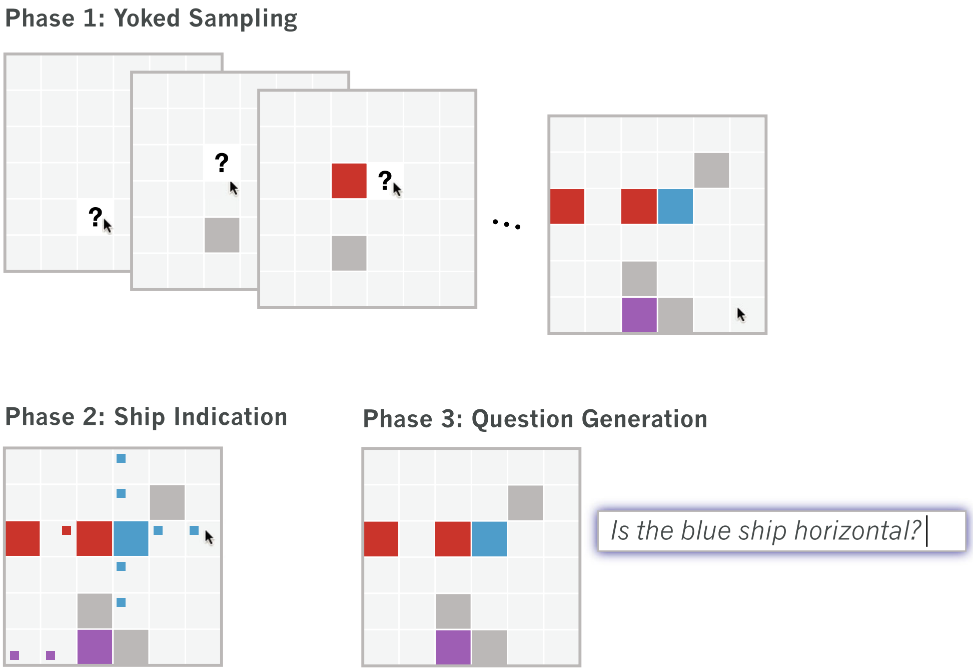


Fig. 3 One trial of the Battleship game in Experiment 1. *Yoked sampling* Participants click on the question marks to reveal a pre-determined sequence of tiles, leading them to a partly revealed game board context. *Ship indication* To ensure participants' attention, participants have to indicate all possible ship tiles. *Question generation* Participants type a question they would like to ask about the hidden configuration in a text box

limitations participants asked only one question per trial, no feedback was provided and there was no painting phase. We emphasized to participants that they should ask questions as though they were playing the game they already had experience with in the earlier part of the experiment.

Results

We recorded 720 questions (18 contexts × 40 participants). Questions that did not conform with the rules or that were ambiguous were discarded (13%). An example of an invalid question is "Where is the red ship?," which cannot be answered with a single coordinate (in contrast to the similar, legal question "Where is one tile of the red ship?"). A few individual questions (3%) were not included because they had features that made them computationally challenging to model.[3] The remaining 605 questions (84%) were categorized by type (see Table 1), and the full data set is available online.[4]

Question Content

We first manually coded commonalities in the meaning of questions independent of the specific wording used. For example, the questions "How many squares long is the blue ship?" and "How many tiles is the blue ship?" have the same meaning for our purposes and were formalized as shipsize(blue), where shipsize is a function with parameter value blue. Since the function shipsize also works with red and purple as parameter values, it represents a cluster of analogous questions. Within these coder-identified clusters, we then calculated the frequency with which such questions were generated across the 18 contexts to get a sense of how the participants approach our task (first column in Table 1).

At a broad level, there are a few natural groups of question types (Table 1). While this partitioning is far from the only possible scheme, it helps to reveal qualitative differences between questions. An important distinction contrasts *location/standard queries* with *rich queries*. Location queries ask for the color of a single tile and are the only question type afforded by the "standard"

Table 1 A comprehensive catalog of the natural language questions obtained in Experiment 1 (regularized across slightly different wordings of the same question)

N	Question
	Location/standard queries
24	What color is at [row][column]?
24	Is there a ship at [row][column]?
31	Is there a [color_incl_water] tile at [row][column]?
	Region queries
4	Is there any ship in row [row]?
9	Is there any part of the [color] ship in row [row]?
5	How many tiles in row [row] are occupied by ships?
1	Are there any ships in the bottom half of the grid?
10	Is there any ship in column [column]?
10	Is there any part of the [color] ship in column [column]?
3	Are all parts of the [color] ship in column [column]?
2	How many tiles in column [column] are occupied by ships?
1	Is any part of the [color] ship in the left half of the grid?
	Ship size queries
185	How many tiles is the [color] ship?
71	Is the [color] ship [size] tiles long?
8	Is the [color] ship [size] or more tiles long?
5	How many ships are [size] tiles long?
8	Are any ships [size] tiles long?
2	Are all ships [size] tiles long?
2	Are all ships the same size?
2	Do the [color1] ship and the [color2] ship have the same size?
3	Is the [color1] ship longer than the [color2] ship?
3	How many tiles are occupied by ships?
	Ship orientation queries
94	Is the [color] ship horizontal?
7	How many ships are horizontal?
3	Are there more horizontal ships than vertical ships?
1	Are all ships horizontal?
4	Are all ships vertical?
7	Are the [color1] ship and the [color2] ship parallel?
	Adjacency queries
12	Do the [color1] ship and the [color2] ship touch?
6	Are any of the ships touching?
9	Does the [color] ship touch any other ship?
2	Does the [color] ship touch both other ships?
	Demonstration queries
14	What is the location of one [color] tile?
28	At what location is the top left part of the [color] ship?
5	At what location is the bottom right part of the [color] ship?

Column *N* reports the number of questions people generated of that type, and brackets denote an argument ("[size]" can be replaced by 2, 3, or 4). Questions are organized into broad classes (headers) that reference different aspects of the game

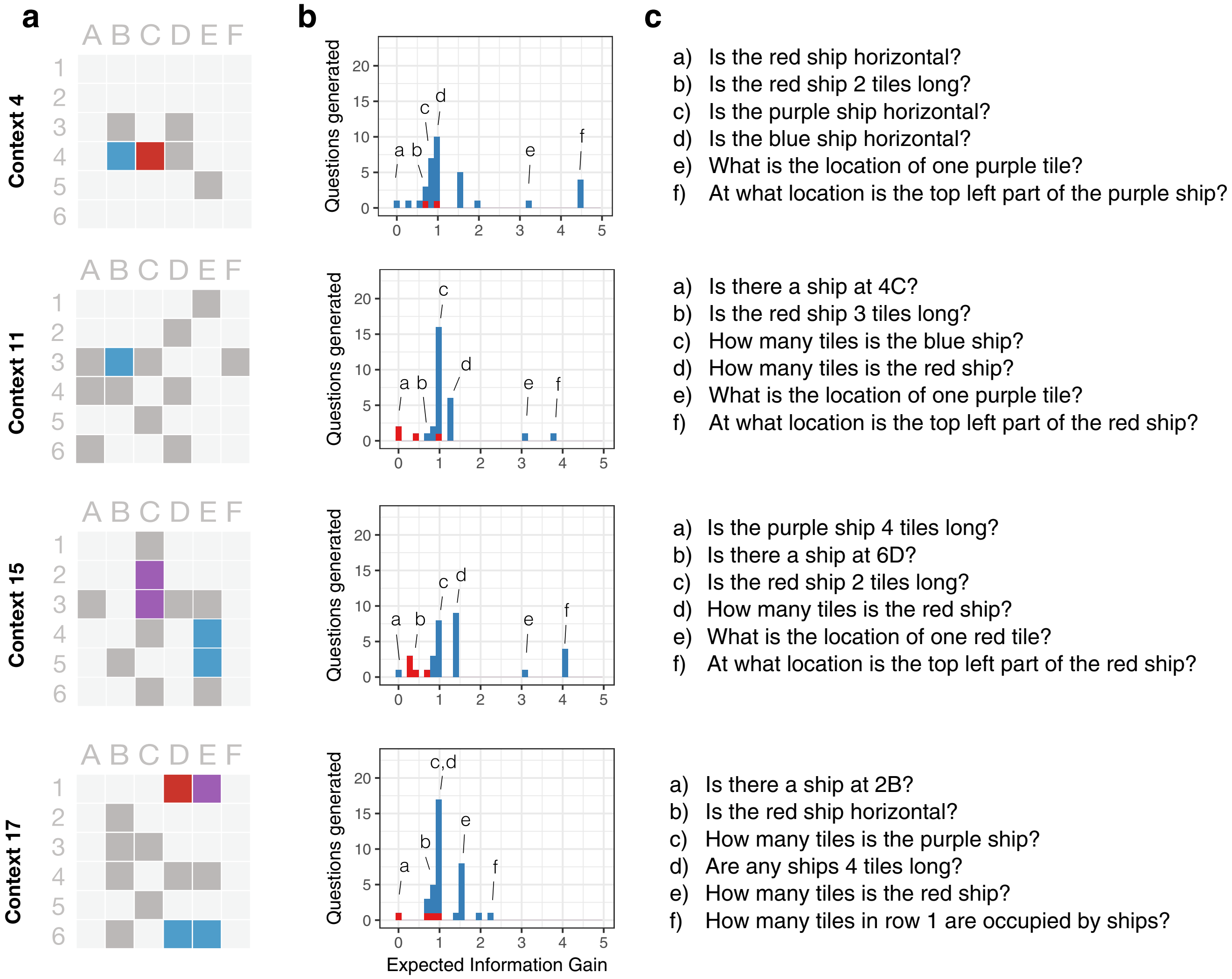
[3] An example from the small set of questions that were not formalized asked whether the purple ship was larger than the part of it that was so far revealed on the board. This question is equivalent with asking "Is the purple ship larger than n?," where n is the number of purple tiles already revealed in the particular context. Interestingly, only a small fraction ($\sim$ 1%) of questions necessitated such dynamic reference to the partly revealed board, while all others could be answered by only accessing the information of the true underlying board. The dropped questions did not seem to be especially informative; thus, it is unlikely that leaving them out changed the results significantly. At least in principle, all of these questions can be formalized in our model given sufficient computational power.

[4] https://github.com/anselmrothe/question_dataset

Battleship task (Markant and Gureckis 2012, 2014b). Rich queries incorporate all other queries in Table 1 and reference more abstract properties of the game. Of the rich queries, *demonstration queries* ask for an example given a reference label (the term demonstration comes from the fact that the learner is essentially asking the oracle to demonstrate a positive example). In the case of Battleship, demonstration queries ask for an example tile of a ship, whereas most other rich queries ask about a part or feature of the game board configuration. Demonstration queries can be especially helpful in active learning settings where the set of positive examples is relatively small (Cakmak and Thomaz 2012; Hendrickson et al. 2016), as is the case in Battleship. Examples of high value demonstration queries are questions e) and f) in Context 4 in Fig. 4c.

Question Frequencies

Among all 605 questions, only 13% were of the location/standard query type. In other words, being freed from the constraints of typical active learning studies, our participants creatively invented a vast array of questions. Only 47 questions (8%) were demonstration queries despite the fact that these can be especially useful (see below). In sum, there were 139 unique questions that were repeated with different frequencies. The most popular questions was "How many tiles is the [blue/red/purple] ship?" (n = 185). When grouping the questions, almost half of all generated questions (n = 289) addressed the size of one or several ships (Table 1). Another large group of questions targeted the orientation of the ships (n = 116).



Context 4:
- a) Is the red ship horizontal?
- b) Is the red ship 2 tiles long?
- c) Is the purple ship horizontal?
- d) Is the blue ship horizontal?
- e) What is the location of one purple tile?
- f) At what location is the top left part of the purple ship?

Context 11:
- a) Is there a ship at 4C?
- b) Is the red ship 3 tiles long?
- c) How many tiles is the blue ship?
- d) How many tiles is the red ship?
- e) What is the location of one purple tile?
- f) At what location is the top left part of the red ship?

Context 15:
- a) Is the purple ship 4 tiles long?
- b) Is there a ship at 6D?
- c) Is the red ship 2 tiles long?
- d) How many tiles is the red ship?
- e) What is the location of one red tile?
- f) At what location is the top left part of the red ship?

Context 17:
- a) Is there a ship at 2B?
- b) Is the red ship horizontal?
- c) How many tiles is the purple ship?
- d) Are any ships 4 tiles long?
- e) How many tiles is the red ship?
- f) How many tiles in row 1 are occupied by ships?

Fig. 4 Four game contexts and example questions produced by participants. The partly revealed configuration (**a**) is shown next to the corresponding histogram of question qualities (EIG; (**b**)), for all of the questions produced by participants in Experiment 1. Red bars signify simple queries of the type used in earlier Battleship experiments (Markant and Gureckis 2012), and blue bars signify rich queries. For reference, a yes/no question in a context where both answers are equally likely has EIG = 1. The six questions in (**c**) were sampled from those obtained in Experiment 1 and used in Experiments 2 and 3

Question Quality

The computational analysis laid out in the introduction assumes that people ask the best possible question (or at least the best one found through a mental search process). One way to assess the quality of a participant's question is to compute its utility according the two "yardstick" models described above, EIG and ES. Even assuming that people have somewhat noisy estimates of either of these two measures (e.g., following a softmax function), one reasonable prediction is that there should be a positive relationship between the frequency a question is asked and its objective quality. That is, the better a question, the more often it will be asked by people.

To compute EIG and ES scores, all questions were represented by functions or programs that would return the answer computed against a true or hypothesized game board configuration. Formally, if a question x was represented by a function f(), then f(h) = d, where d is the answer to question x given the true or hypothetical configuration h. The functions were written by the experimenter by hand to capture the semantic meaning of the natural language questions asked by participants. To give an example, with h being the particular configuration shown in Fig. 1, the question "How many tiles is the blue ship?" would return shipsize(blue, h) $=$ 3. This process could be repeated for all hypothesized configurations. The function itself would access a representation of the true gameboard configuration and run a short algorithm to determine the answer. The same function could be run on representations of a hypothetical gameboard (i.e., those that are plausible solutions to the game). In conjunction with Eqs. 4 to 10, this provided the machinery to evaluate all 605 questions by the EIG and the ES model. Specifically, for each question-context pair, where the context was the partly revealed game board in which the question had been generated, one EIG and one ES score were computed, providing objective quality scores for people's questions.

A few uninformative questions (i.e., EIG $=$ ES $=$ 0) where asked (n $=$ 19, 3%, by 11 participants across 13 trials). These questions asked for information that was already given by the context (see question (a)

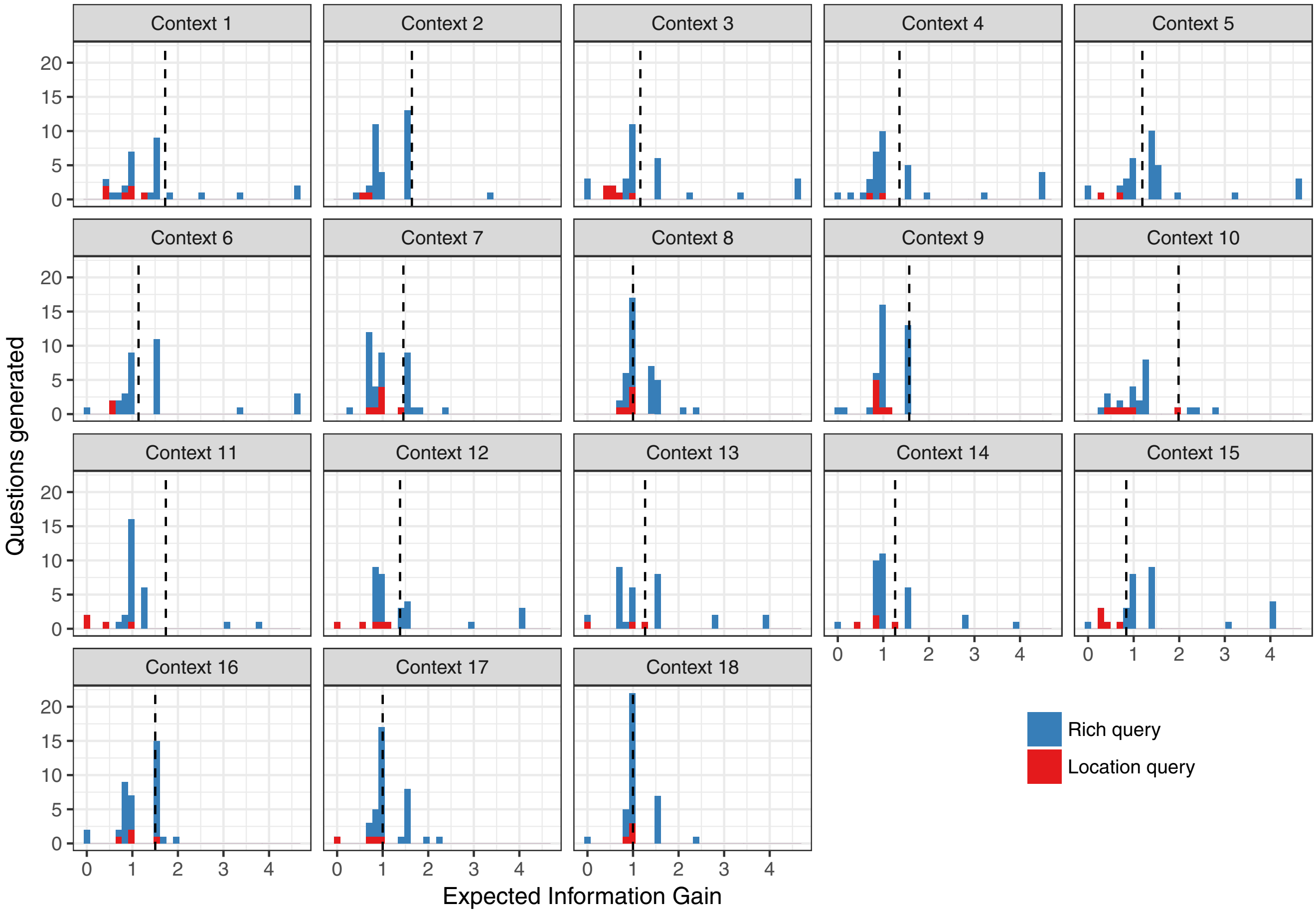


Fig. 5 Histograms of question qualities in Experiment 1. Dashed lines mark the upper limit on the quality of the location queries (i.e., the best possible location query). Note that the mode of each distribution is in the intermediate range of the distribution and not at its right end

in Fig. 4c). Surprisingly, most questions having some potential information concentrated at an intermediate level of informativeness while the objectively best questions were generated rarely. The histograms in Fig. 4b demonstrate this pattern for a few example contexts. For reference, labels were added that mark the quality of the example questions listed in Fig. 4c. For the histograms, the bin width was chosen such that a single count of a question resulted in a square. Notably, each histogram showed a pronounced peak, indicating that people preferably asked questions in a particular quality range. However, this peak was nowhere near the range of high-quality questions. Indeed, it turned out that in all 18 contexts, the histogram had a mode which was substantially below the maximum range of values (see Fig. 5).

A second analysis corroborated the finding that a better question was not necessarily asked more often. There was only a very weak relationship between the frequency of a question and its quality scores (i.e., EIG or ES). The correlation between frequency and EIG was computed for each context and then averaged across contexts, leading to a mean correlation of 0.16 *HDI* = [0.08, 0.23].[5] For ES and question generation frequency, the mean correlation was 0.04, *HDI* = [−0.06, 0.15].

Another interesting comparison considers the quality of rich queries and location/standard queries. In every context, there were a number of rich queries that people asked which outperformed the best standard queries (based on EIG). In all but one context where there was a draw, rich queries even beat standard queries when using the theoretical upper limit for standard queries as benchmark (indicated by a dashed line in Fig. 5). When using ES to measure question quality, rich queries were similarly dominant, defeating standard queries in 16 out of 18 contexts. The mean EIG for standard queries was 0.75 *HDI* = [0.64, 0.85] compared to 1.26 *HDI* = [1.18, 1.35] for rich queries, suggesting increased informativeness of the more sophisticated natural language questions (see red vs. blue in Fig. 4b). The difference of the two means had *HDI* = [0.38, 0.64], and 100% of the posterior probability mass were above 0, indicating high confidence that rich queries performed better. The mean ES for standard queries was 0.62 *HDI* = [0.47, 0.77] compared to 0.72 *HDI* = [0.62, 0.83] for rich queries. The difference of the two means had *HDI* = [−0.08, 0.29], and 87% of the posterior probability mass were above 0. Therefore, the advantage of rich queries over standard queries was less pronounced when measuring question quality with ES instead of EIG.

We tested whether removing questions that participants asked after failing the ship indication all six times resulted in a cleaner data set. Since participants failed often, almost a fifth of the questions would need to be removed under this criterion. However, when doing so, the average EIG score of questions goes up by only 0.06 *HDI* = [0, 0.11] (by 0.03 for ES, *HDI* = [−0.03, 0.09]). In retrospect, the ship indication task might have been overly difficult at times for some participants. We therefore decided to not exclude questions based on ship indication performance.

Question Length

The produced questions varied in their number of words from 1 (a participant chose to make a simple location query by typing "1C") to 18, with *M* = 7.07, SD = 2.15. An analysis of question length found that shorter questions were not asked more often. Equivalent to the analysis of EIG and question frequency described above, we correlated frequency with the average number of words (while the EIG values were constant across questions with identical meaning, the number of words vary and hence we needed to take the average). As above, the correlation was computed separately for each context, resulting in a mean correlation of 0.04 *HDI* = [0.02, 0.06].

Context Specificity

A good question in a certain context is not necessarily a good question in a different context. To estimate the context sensitivity of the generated questions, we permuted the contexts each question was associated with across all 605 questions and evaluated the EIG for each new context-question pair (see Fig. 6). Note that this was only possible because the formalized question functions were entities that represented the meaning of the question independently

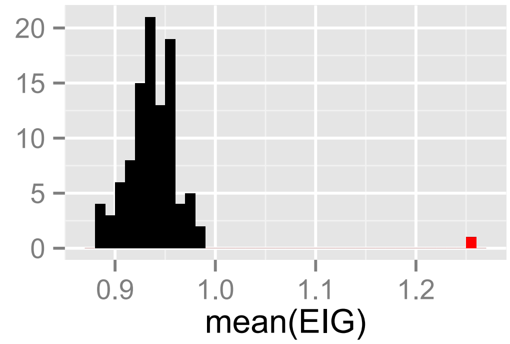


Fig. 6 Context specificity in question asking. If the questions are evaluated in a different context (black, 100 bootstrap samples), they clearly lose quality compared to the original context they were asked in (red)

[5] For each estimation, we report the 95% highest density interval (*HDI*), based on Bayesian data analysis. In all of these estimations, broad priors as described by Kruschke (2013) were used.

from the context they were asked in. Thus, every question function could be evaluated in a variety of new contexts. The average EIG across questions in the original data set was larger than in all 100 permutation sets ($p < 0.01$). An analysis for ES on a subset of the contexts obtained the same result (due to the higher computational demands of ES the analysis could only be run on six contexts). In summary, people produced a range of questions that were both rich and context-sensitive.

Discussion

In Experiment 1, we observed that people generated a variety of natural language questions targeting various aspects of the Battleship game boards in a context-sensitive fashion. By formalizing these questions as functions, we could compute their exact information-theoretic value with respect to the Battleship game's massive hypothesis space. We found that people asked more informative questions than when limited to simple location queries, which only return the color of a particular board tile. The context sensitivity analysis shows that people flexibly altered their questions to adapt to the idiosyncratic circumstances of particular game contexts. This finding appears to rule out heuristic strategies that only target questions generally useful for the overall game (e.g., always asking "What is the size of the red ship?").

On the other hand, the very best questions in our dataset were asked by only a small fraction of participants. This indicates that the questions that came to people's minds where not perfectly tuned to the informational structure of the game (otherwise all or most participants would ask the objectively "best" questions). This also means that neither EIG nor ES is a complete account of human question generation.

There are a variety of possible reasons that our subjects failed to universally select high information (or expected savings) questions. One hypothesis is that people had difficulty thinking of good questions to ask, effectively limiting the search for the best possible expression (i.e., Eq. 2). Alternatively, people may have difficulty evaluating the questions (thinking that an intermediate quality question was actually better than a more informative one). To decide between these alternatives, in Experiment 2, we removed the demand of generating question by providing people with a list of possible questions, thus separating question *generation* from question *evaluation*. If people have difficulty identifying the best questions to ask, we expect to see a similar pattern of results in Experiment 2. Alternatively, if the process of creatively generating a highly informative question is the major bottleneck, we expect people's behavior will be more aligned with our models in Experiment 2.

Experiment 2—Question Evaluation

The results of Experiment 1 showed that people failed to agree with the models about the best questions to ask. In Experiment 2, we instead provided people with a list of question selected from those asked by participants in Experiment 1.

Participants

Forty-five participants on Amazon Mechanical Turk were paid \$2 with a potential performance-based bonus of up to \$3.78.

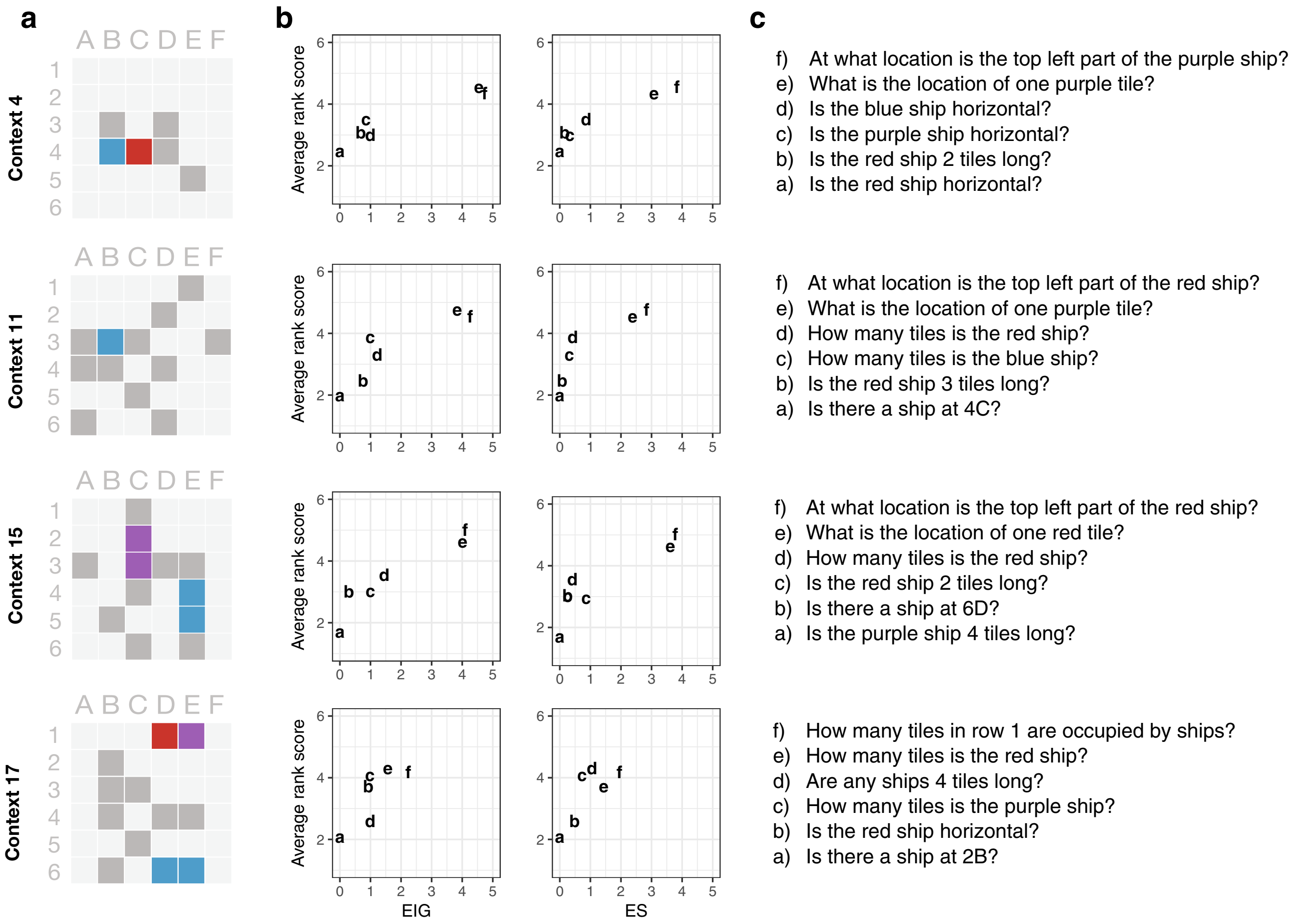
Method

The materials and procedure were nearly identical to Experiment 1, except that instead of entering a question into the text box at the end of each trial, participants ranked a provided set of natural language questions according to perceived informativeness for the given setting. Participants viewed the same 18 board configurations (contexts) along with a selection of six natural language questions. They were asked to rank the questions for quality by positioning them from best to worst in a sortable list. After sorting, they were asked to select their favorite question, presumably the question at the top of the ranked list.

For each context, the six question options were sampled from the full list of human-generated questions from the corresponding trial in Experiment 1 (see Fig. 7c). This reduction was necessary, as the intention of this experiment was to study question evaluation without the burden of having to consider a large number of possible questions. We used a simple algorithmic sampling procedure designed to include the most frequently generated questions, the highest quality questions (according to EIG), and some questions that were neither frequent nor high quality.[6] The questions were regularized into a simple grammatical form, removing typos and odd constructions from the original human generated versions while preserving the semantic meaning.

To ensure that people read each question they ranked, they were asked to classify each question by the form of its possible answers (either a color, a coordinate on the grid, a number, or yes/no, which span all possible answers to the questions in Table 1). Participants were awarded a bonus of \$0.015 for each correct answer, and they did not receive feedback. As in Experiment 1, this

[6] The free-from questions for each context in Experiment 1 were placed in a 2D space with EIG and generation frequency as dimensions. We then sampled 1000 six-question subsets and took the sample with the largest average pairwise distance between questions in the subset.



Context 4:

f) At what location is the top left part of the purple ship?
e) What is the location of one purple tile?
d) Is the blue ship horizontal?
c) Is the purple ship horizontal?
b) Is the red ship 2 tiles long?
a) Is the red ship horizontal?

Context 11:

f) At what location is the top left part of the red ship?
e) What is the location of one purple tile?
d) How many tiles is the red ship?
c) How many tiles is the blue ship?
b) Is the red ship 3 tiles long?
a) Is there a ship at 4C?

Context 15:

f) At what location is the top left part of the red ship?
e) What is the location of one red tile?
d) How many tiles is the red ship?
c) Is the red ship 2 tiles long?
b) Is there a ship at 6D?
a) Is the purple ship 4 tiles long?

Context 17:

f) How many tiles in row 1 are occupied by ships?
e) How many tiles is the red ship?
d) Are any ships 4 tiles long?
c) How many tiles is the purple ship?
b) Is the red ship horizontal?
a) Is there a ship at 2B?

Fig. 7 **a** Four selected contexts exemplifying the partly revealed configuration. **b** Participants' rank orderings of the questions (higher rank score means better question) strongly correlated with the Bayesian expected information gain (EIG) and Bayesian expected saving (ES) model scores. **c** Participants ranked six questions that were sampled from those obtained in Experiment 1 (for comparison, identical questions were shown in Fig. 4C but in reversed order for readability)

bonus encouraged attention but did not provide a monetary incentive tied to the quality of the ranked question list.

Results

In the following, we will present three comparisons. First, we will compare the two dependent variables in Experiment 2 (choosing and ranking), then, we will compare the results from Experiment 2 with those from Experiment 1, and finally, we will compare the two models EIG and ES.

Question Evaluation

Participants ranked the six questions and subsequently selected their favorite from the list. In our analysis, a higher rank score represents a better question (i.e., 6 for the highest position and 1 for the lowest). In 74% of the cases, participants selected the question they also had ranked highest. Vice versa, for all questions that were selected, across participants and contexts, the mean and median rank scores were 5.4 and 6, respectively. This suggests that the two measures were capturing similar (but not completely identical) variables. For brevity, the following analyses only used the rank data as it provides a more fine-grained measure of people's judgments. A detailed presentation of the ranking of the questions in four of the contexts, plotted against the scores that the EIG model and the ES model assigned to those questions, is shown in Fig. 7.

Question Generation Versus Evaluation

Figure 8 shows data for all 18 contexts, with the data for Experiment 1 constrained to the subset of questions that were also used in Experiment 2. Restating the finding for the full data in Experiment 1, the *frequency* with which people generated questions was not significantly correlated to the questions' quality as measured by the models, both in the full set of generated questions and the subset used in Experiment 2. The mean correlation across contexts was $r = -0.04$, $HDI = [-0.14, 0.06]$, for EIG (panel a)

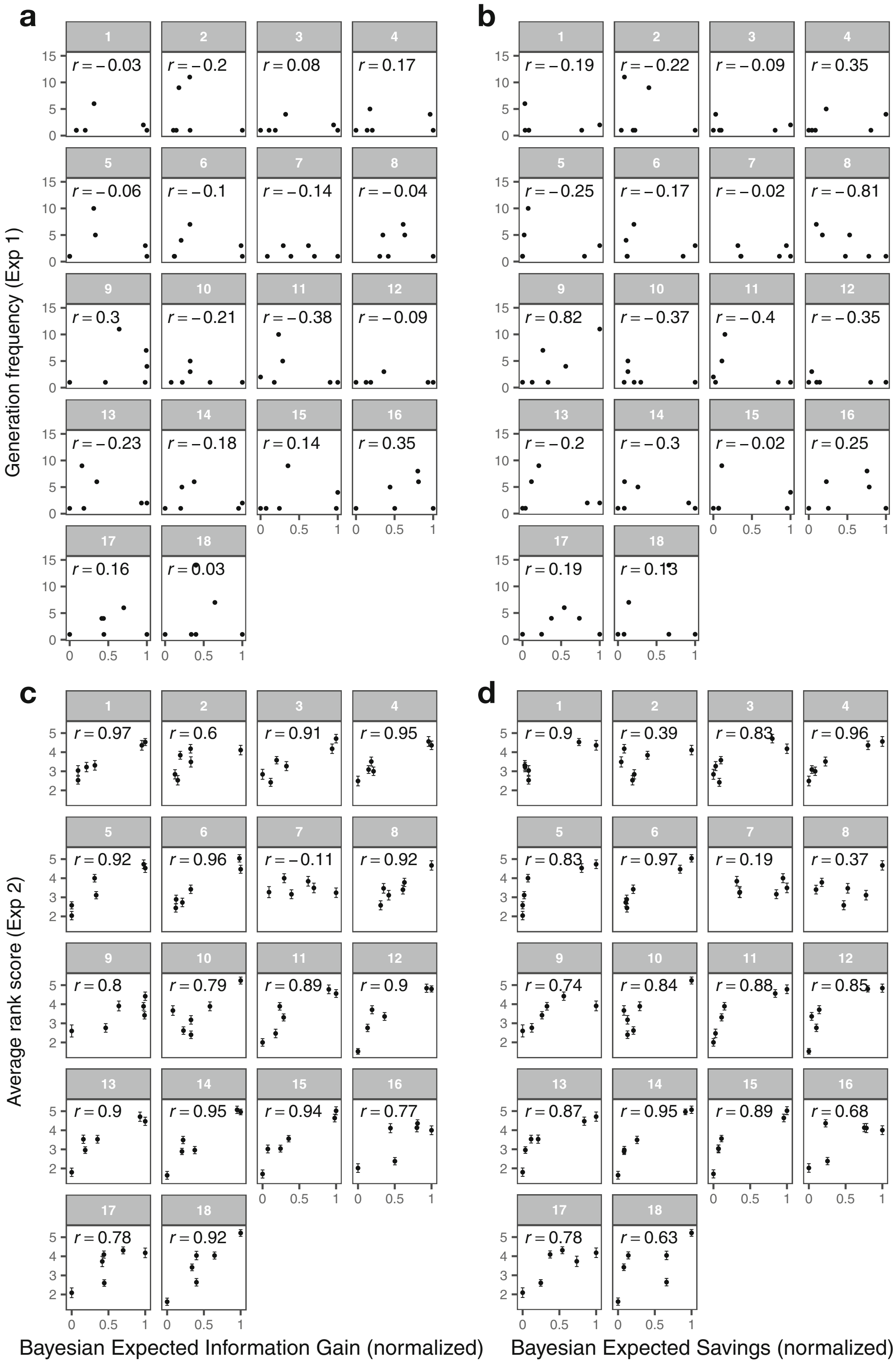


Fig. 8 Comparing judgments of question value for people versus the Bayesian ideal learners. In Experiment 1 where participants generated questions, questions that score highly according to the ideal learners (EIG and ES) were not necessarily asked more frequently (**a**+**b**). In Experiment 2 where participants evaluated questions, questions that score highly according to the ideal learners tended to be ranked more highly by participants (**c**+**d**). The *x*-axis showing the ideal learner scores was normalized so that the maximum value is 1. The same six questions per context are shown in (**a**+**b**) versus (**c**+**d**) for comparison, meaning these six were only a subset of the questions in Experiment 1 while constituting the full set for Experiment 2

and −0.11, *HDI* = [−0.28, 0.06], for ES (panel b). In contrast, people's *ranking* of the questions in Experiment 2 was highly correlated with the model scores, with mean correlation $r = 0.89$, *HDI* = [0.84, 0.94], for EIG (panel c) and 0.78, *HDI* = [0.66, 0.89], for ES (panel d) (see Fig. 7b for detailed examples). Only context 7 did not follow this pattern and basically showed no correlation, but this might be a floor effect as the list of questions shown to participants in Experiment 2 was less diverse in its EIG/ES scores than for other contexts (participants did not produce any high-quality questions for context 7 in Experiment 1). A Bayesian analysis for EIG confirmed the strong difference between the correlations in Experiment 2 (ranking vs. EIG) and Experiment 1 (generation frequency vs. EIG) with *HDI* = [0.81, 1.05] and 100% of the posterior probability mass above 0. Equivalently, for ES there was a strong difference between Experiment 2 (ranking vs. ES) and Experiment 1 (generation frequency vs. ES), with *HDI* = [0.66, 0.89] and 100% of the posterior above 0.

Which "Yardsticks" of Question Quality?

We analyzed which of the two models of question quality, EIG and ES, provides the best fit to participants rankings of natural language questions. The key distinction between the models is that the EIG model purely focuses on the uncertainty reduction about the correct configuration, while the ES model takes the tile-painting task into account and focuses on minimizing the number of painting mistakes, and therefore considers the overall task costs.

The different preferences of the models become clear with the example of the question "How tiles is the red ship?" in Context 15 (question (d) in Fig. 7). Under EIG, this question is useful because every answer will allow the learner to rule out all hypothesized configurations that are inconsistent with that answer. Under ES, this question is hardly useful because in Context 15, the learner does not know yet where on the board the rep ship is located at all, so the abstract information about the ship size would barely increase the chances of painting the red ship correctly (compare Fig. 7b, c).

On the other hand, model simulations showed that there are many questions for which both models make very similar predictions. With respect to the human data, Fig. 8 shows how EIG and ES make predictions that are not identical but overall are quite similar. Instead of comparing correlation coefficients, we conducted a more sensitive model comparison that takes guessing behavior into account. Model scores were transformed into choice probabilities via the softmax function,

$$p(x) = \frac{e^{-\beta M(x)}}{\sum_x e^{-\beta M(x)}} \tag{11}$$

where $M(x)$ is the model score (i.e., EIG(x) or ES(x)) and β is the free temperature parameter, capturing more guessing behavior as $\beta \rightarrow 0$. For each model, β was fit per participant to the choice data from Experiment 2, and the resulting log-likelihood computed, $\log p(\text{data}) = \sum_{x \in \text{data}} \log p(x)$. We found that ES had a higher log-likelihood for 36 out of 45 participants (80%).

Discussion

Both the EIG and ES models were highly predictive for the people's rankings of provided questions. This is impressive given that the models incorporate a Bayesian ideal observer analysis without any free parameters. The finding that people were able to objectively evaluate question qualities in line with the models is striking.

In our direct comparison, we found that the expected savings (ES) model accounted better for human judgments than the expected information gain (EIG) model. This suggests that people were somewhat congnizent of the cost structure of the task (i.e., guessing the colors of the remaining tiles in the painting task immediately after learning the answer to the question), which is better captured by ES than EIG. Interestingly, Markant and Gureckis (2012) reported that EIG provided a better fit to participant's query behavior than does ES in a similar task. A possible explanation for the reversal of findings is that the natural language question asking exposes a broader range of possible queries, some of which might more clearly distinguish the different sampling models.

One remaining concern is that in Experiment 2 the cost structure of the task may not have been sufficiently obvious. For example, people were not actually provided answers to their questions and they did not actually have to perform the painting task. In Experiment 3, we conducted a conceptual replication of Experiment 2 but made the cost structure of the task much more apparent. Specifically, participants received an answer to their question and a bonus payment based on their performance in the painting task. Our expectation was that this would increase the use of the ES strategy and remove some of the variance from the prior experiment.

Experiment 3—Question Evaluation with Explicit Cost Structure

We replicated Experiment 2 while providing people with a clear objective when they evaluated the natural language questions. People received the answer to their chosen question and subsequently guessed the underlying game board configuration, through the process of explicitly performing the painting phase task (Fig. 1). A monetary bonus payment was based on their accuracy in this final phase.

Participants

Forty-one participants on Amazon Mechanical Turk were paid \$6 with a potential performance-based bonus of up to \$3.60. We paid more than in Experiments 1 and 2, as this experiment was longer.

Method

The materials and procedure were nearly identical to Experiment 2, except that participants received the answer to their selected question and could use this information in a subsequent painting phase. Also, we streamlined the question selection procedure in that the top-ranked question was automatically coded as the participant's "chosen" question, and the trial order was randomized. The same six-question subsets from in Experiment 2 were used, albeit 9% of the questions were different due to resampling.

For each question, we precomputed the answers assuming a randomly chosen underlying "true" configuration that was consistent with the partially revealed gameboard (this target configuration was constant across participants). After a participant had ranked the questions, the top-ranked question was highlighted and the corresponding answer displayed. This information remained visible painting phase, where the participant painted in the empty tiles in the partly revealed game board. For each correctly painted tile, we awarded a potential bonus of \$0.10. For fairness, already revealed tiles were counted as correct, thus resulting in a possible bonus of \$3.60 for all 36 tiles. The bonus was only paid for a single trial, selected by a lottery at the end of the experiment. This allowed us to award a higher bonus per tile and also kept people motivated for the entire task.

Results

Overall, we obtained results very similar to those in Experiment 2. Figure 9 shows again high correlations between human rank scores and model scores (for EIG, see panel a, mean $r = 0.87$, $HDI = [0.73, 0.98]$; for ES, see panel b, mean $r = 0.86$, $HDI = [0.80, 0.93]$). Equivalently to the analysis in Experiment 2, we compared the correlations between Experiments 3 and 1. Again, a Bayesian estimation of $\mu_{\text{diff}} = \mu_{\text{Exp3}} - \mu_{\text{Exp1}}$ found a strong difference between the correlations in Experiment 3 (ranking vs. EIG) and Experiment 1 (generation frequency

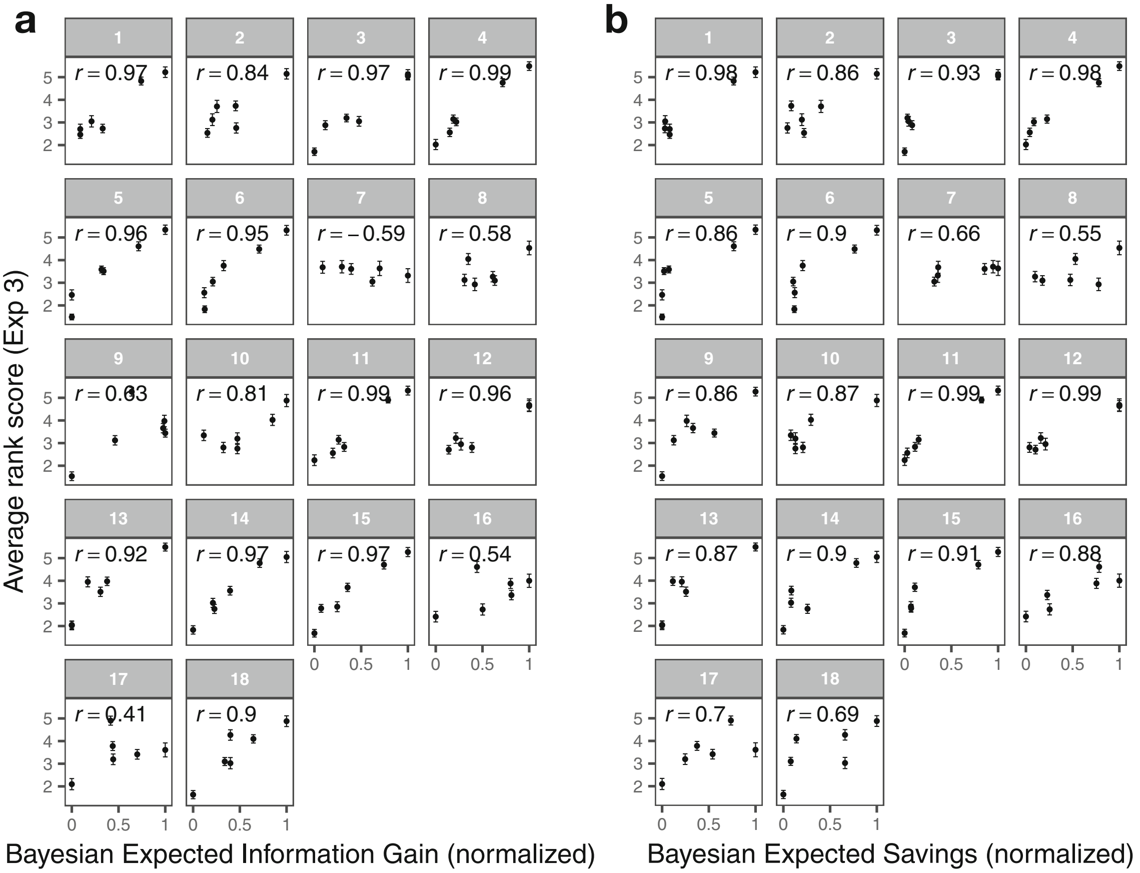


Fig. 9 Comparing judgments of question value for people versus Bayesian ideal learners. There were strong correlations between human rank scores and EIG scores (**a**) as well as ES scores (**b**) in Experiment 3

vs. EIG), *HDI* = [0.73, 1.08], 100% posterior mass above 0, and a similarly strong difference when replacing EIG with ES, *HDI* = [0.78, 1.15], 100% posterior mass above 0.

To test the similarity between the correlations in Experiments 3 and 2, $\mu_{\text{diff}} = \mu_{\text{Exp3}} - \mu_{\text{Exp2}}$, we defined −0.1 to 0.1 to be a *region of practical equivalence* (ROPE) around the null value of μ_{diff} (Kruschke 2013). If nearly all of the posterior (e.g., 95%) falls into this region, we can accept the null hypothesis that there is no practically relevant difference between the correlations. For EIG, the *HDI* of μ_{diff} ranged from −0.06 to 0.09 and thus fell completely into the ROPE. Therefore, the result with respect to EIG was practically the same in both experiments. For ES, the *HDI* ranged from −0.03 to 0.21, and while only 60% of the posterior fell into the ROPE, 92% of the posterior mass were above 0. This result indicated slightly stronger correlations with ES in Experiment 3 than in Experiment 2.

For a direct model comparison, as in the Experiment 2 analysis, a softmax temperature parameter was fit to the choice data (i.e., highest-ranked question) for each participant for each model. We found that ES had a higher log-likelihood than EIG for 30 out of 41 participants (73%), which was slightly below the 80% in Experiment 2.

Discussion

Experiment 3 used the same stimuli as Experiment 2, but in contrast to Experiment 2, participants received answers to their chosen questions and then had to complete the painting task. As a result, Experiment 3 had a more explicit incentive structure that encouraged participants to ask questions in order to perform well in the painting task. Given these differences, we expected people's choices to be more in line with the ES model than before because ES exactly captures the incentive structure of the experiment. Overall, the results were very similar to the results of Experiment 2. The ES model provided a good account of people's question ranking in both experiments, and the correlations were slightly stronger in Experiment 3. In the direct model comparison, ES outperformed EIG in both experiments for predicting the judgments of individual participants, although by a slightly smaller margin in Experiment 3 than Experiment 2. This is somewhat surprising given that participants could have used their experience with receiving answers and painting to inform their question selection strategy. Overall, the ES model is still a strong predictor of how the participants evaluated questions.

General Discussion

We began this paper by asking, "Do people ask good questions?" Although we are still far from a complete account, our experiments and analyses represent a step forward in understanding the effectiveness of human question asking in computational terms. Bridging previous qualitative and quantitative approaches to the study of question asking, we studied free-form natural language question asking in a rich and intuitive setting that was nonetheless still amenable to Bayesian ideal observer analysis. Through these ideal observer measures of question quality, we were able to analyze the wide variety of natural language questions that people asked, placing them together on a common scale for comparison. The design additionally allowed the comparison of question generation and question evaluation within the same framework, providing insight into when and why people may fail to ask the best possible question in a particular scenario.

In Experiment 1, we found that people asked a variety of rich and interesting questions that grouped into various identifiable subtypes and were highly tuned to the particular contexts in which they were asked. Some of the questions people came up with were rather direct attempts to ascertain hidden properties of the game board such as "Is the red ship 3 tiles long?" or "Is the red ship horizontal?". Yet other were striking in their creativity. For example, in context 4, four participants asked, "What is the top left part of the purple ship?" a clever approach (obtaining information about the location and, potentially, orientation and size of the purple ship in a single query) that even the authors of this paper did not consider. In addition, we found that questions were highly context sensitive, responsive to the particular situation on the partially revealed gameboard rather than being driven by blind heuristics. However, interestingly, people rarely asked the objectively best questions relative to more mundane, intermediate quality questions.

In contrast, in Experiment 2, we found that people strongly preferred the objectively best questions, once they were provided with the questions and did not have to generate them from scratch. We found evidence that people's rankings of question were better described by the expected savings (ES) model that values questions that improve people's performance on the inference task directly (i.e., the painting phase), compared to the expected information gain (EIG) model that values information irrespective of the costs and benefits for the immediate task, something we discuss further below. Moreover, the close correspondence with the models also rules out that participants in the first experiment were not capable of evaluating or agreeing upon the quality of the questions they asked.

In Experiment 3, we replicated the finding from the Experiment 2 in a setting with a much more explicit cost structure. In this experiment, people received the answer to their selected question and were paid for their performance in a subsequent test, for which the answers provided potentially useful information. Once again, people's preferences among questions were better described by ES than EIG, and the results sharply diverged from the results of Experiment 1.

The free-form question asking task in Experiment 1 contrasts with past work on active inquiry, which has often provided participants with lists of question or queries. We argue that such a procedure dramatically simplifies the computational problem, resulting in a more favorable view of human performance in inquiry tasks than might be true in everyday life. Our results also highlight the need for further computational work aimed at understanding question and query synthesis, given its key role in human inquiry in naturalistic settings.

Question Evaluation

The finding in Experiments 2 and 3 that people's evaluation of questions aligned more closely with the cost-sensitive ES model is surprising in light of previous work, which found the opposite pattern in a similar task (Markant and Gureckis 2012). Markant and Gureckis (2012) explain that a cost-sensitive strategy can be computationally more demanding as every piece of information needs to be evaluated in light of the decision policy the learner will adopt in the test phase. However, there are reasons to suspect that people could intuitively align more with ES even without direct calculation. For example, as noted above some high-EIG questions are seemingly abstract (e.g., "What is the total number of water tiles?") in the sense that they provide a lot of information but do not help the learner with the immediate task of painting the tiles to complete the game board. In contrast, if one tile of the red ship is revealed and the learner asked "What is the orientation of the red ship?," they can use this to make more informed guesses about the location of the red ship in the painting phase. The possible discrepancy between our results and those of Markant and Gureckis (2012) may stem from the fact that these types of differences are much more apparent in the case of natural language question asking due to the rich space of possible queries.

The finding that people's question *evaluations* were so well captured by the Bayesian ideal observer model is surprising in light of the immense computation required in our models to determine the quality of each question, and the fact that the models have no free parameters. In part the large computational burden stemmed from of the massive hypothesis space, H, and the large space of possible answers to certain question types, A_x (c.f. Eq. 7 in the introduction). While the sets of answers in our task domain were straightforward, everyday life questions often have an potentially infinite amount of possible answers, escalating the computational problem even further.

Although our ideal learners are defined at the computational level, additional work is needed to understand how these computations can be approximated or mimicked at the algorithmic level. A recent review article, Coenen et al. (in press), provides a fairly extensive consideration of these issues. A key point though is that whatever mechanism people use to evaluate questions it seems to correlate strongly with the information-theoretic, probabilistic analysis of the problem that we advanced.

Question Generation

In interpreting the results of Experiment 1, we consider three alternative explanations for why people rarely generated the best questions. First, it is possible that participants were just not motivated to ask good questions in the game. This seems unlikely, however, given that the incentive structure was basically identical between the generation and evaluation experiments (Experiments 1 and 2), and unmotivated participants would have also failed to recognize the best questions in the evaluation experiment. Second, it is possible that participants were not sure what questions were allowed and did not want to risk asking a good but illegal question. Points that speak against this objection are that the instructions were fairly simple to follow and that we had extensive and explicit checks of participants' understanding. Participants were also repeatedly reminded of the allowed answer types (a word, a number, true/false, or a single coordinate), making clear what counted as a question that could be answered by a single piece of information. Third, it is possible that it is difficult for people to come up with good questions from scratch when they are unfamiliar with the settings of the task, but one good example question is all they would have needed to prevail. This claim is challenged by anecdotal evidence showing that even after having seen the many examples in this paper, it is hard to generate novel highly informative questions. For instance, we invite the reader to consider what question would be better than any of those in Fig. 7. A participant that just repeats or slightly tweaks a previously seen question is not engaging in the creative generation process we intend to study here. In fact, we were careful to not provide any examples of rich questions in the instructions of the generation experiment to avoid biasing participants by influencing their creative process.

What then makes the generation of questions so difficult? Although the exact mechanism by which people synthesize novel questions remains elusive, one idea is that question asking is akin to a search for the maximally useful question (i.e., Eq. 2) within some defined syntactic and semantic space, Q. As a consequence, as this space of candidate questions, Q, grows, the search task becomes harder for an agent with limited computational resources (e.g., limited memory or time). In fact, in many situations, including our experiment, the question space is approximately infinite, rendering an exhaustive search a futile effort. Instead, people might settle on local minima in the space (i.e., “good enough” questions), perhaps by terminating the search after a certain time and then asking the best question discovered so far.

A key takeaway from our study is that any computational account of question asking not only needs to model the evaluation of questions (such as Eq. 1) but also has to account for where questions come from (i.e., the question space Q in Eq. 2). Many of the generated questions shared concepts that resembled game features that the instructions of the experiment had “built-in” (e.g., the size or the orientation of a ship). However, participants also referred to features that are inductive in nature (e.g., about ships touching each other, about one ship being larger than another, or about ships having parallel orientation). A computational model of question generation must be able to synthesize such new features or transfer them from previously seen tasks. We have begun laying out the groundwork for such a model, building on the data set of rich questions obtained in the present study (Rothe et al. 2017). In brief, a probabilistic generative model can be defined over a space of questions Q to provide an estimate of how likely a person is to ask any particular question in the space, given the current context and knowledge state of the question asker. As in the current paper, questions are formalized as functions/programs, and the space of questions Q can be defined compositionally with a context-free grammar that incorporated building blocks such as the size of a ship and the color of a tile, as well as generic sub-functions for performing logical, arithmetic, and set operations. Although the approach still requires domain-specific engineering, it demonstrates how novel and informative questions can be generated in new contexts that the model has never been trained on. We see this, and related efforts to formally characterize Q and develop generative models, as an important tool for further exploring the computational mechanisms for question asking and more precisely characterizing its successes and limitations.

Conclusion

Do people ask good questions? In this paper, we addressed this question through a series of behavioral and computational experiments, using rich game-based scenarios that were tractable for ideal observer measures of question quality. Even in these games, the computational burden of an ideal question asker is immense, scaling with factors that can be intractably large: the number of possible questions, the number of possible answers to each question, and the size of the hypothesis space. Nevertheless, with only minimal familiarity with the task at hand, people generated a catalog of interesting and creative questions that were goal-directed and context sensitive. By the standards of contemporary AI systems and active learning algorithms, no algorithm comes close to matching the flexibility and sophistication of human question asking. On the other hand, people rarely asked the best question available in any given scenario (Experiment 1), even though they could accurately estimate the value of the best questions when presented as a set of fixed alternatives (Experiments 2 and 3). However impressive, the human ability to ask questions is not without limitations, and we see additional computational work addressing query synthesis as key to further unraveling the mystery of human question asking.

Funding Information This research was supported by a National Science Foundation grant BCS-1255538, the John Templeton Foundation “Varieties of Understanding” project, a John S. McDonnell Foundation Scholar Award to TMG, and the Moore-Sloan Data Science Environment at NYU.

References

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale: Lawrence Erlbaum Associates.

Cakmak, M., & Thomaz, A.L. (2012). Designing robot learners that ask good questions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 17–24). https://doi.org/10.1145/2157689.2157693.

Chater, N., Crocker, M., Pickering, M. (1998). The rational analysis of inquiry: The case of parsing. In Oaksford, M., & Chater, N. (Eds.) *Rational models of cognition* (pp. 441–468). Oxford University Press.

Chin, C., & Brown, D.E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*, *24*(5), 521–549.

Clark, H. (1979). Responding to indirect speech acts. *Cognitive Psychology*, *11*(4), 430–477.

Coenen, A., Nelson, J.D., Gureckis, T.M. (in press). Asking the right questions about human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*.

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

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, *78*(4), 256–262.

Dillon, J.T. (1988). The remedial status of student questioning. *Journal of Curriculum Studies*, *20*(3), 197–210.

Graesser, A.C., Langston, M.C., Bagget, W.B. (1993). Exploring information about concepts by asking questions. *The Psychology of Learning and Motivation*, *29*, 411–436.

Graesser, A.C., Person, N., Huber, J. (1992). Mechanisms that generate questions. In Lauer, T.W., Peacock, E., Graesser, A.C. (Eds.) *Questions and information systems* (pp. 167–187). Hillsdale, Erlbaum.

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 *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Chan, P. (2016). psiTurk: an open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Hawkins, R.X.D., Stuhlmuller, A., Degen, J., Goodman, N.D. (2015). Why do you ask? Good questions provoke informative answers. In Noelle, C.D., et al. (Eds.) *Austin, TX: Cognitive Science Society*.

Hendrickson, A.T., Navarro, D.J., Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62.

Kruschke, J.K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Markant, D.B., & Gureckis, T.M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Markant, D.B., & Gureckis, T.M. (2014a). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Markant, D.B., & Gureckis, T.M. (2014b). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Marr, D.C. (1982). *Vision*. San Francisco: W.H. Freeman and Company.

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

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*(1), 74–80.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, *38*(1-2, Serial No. 149), 1–135.

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

Rothe, A., Lake, B.M., Gureckis, T.M. (2017). Question asking as program generation. *Advances Neural Information Processing Systems*, *30*, 1046–1055.

Ruggeri, A., & Feufel, M.A. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, *6*.

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.

Settles, B. (2009). Active learning literature survey (Tech. Rep.). University of Wisconsin-Madison.

Settles, B. (2012). *Active learning*. San Rafael: Morgan & Claypool Publishers.

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