Journal of Experimental Psychology:
Learning, Memory, and Cognition
2017, Vol. 43, No. 8, 1274–1297

0278-7393/17/$12.00 http://dx.doi.org/10.1037/xlm0000374

Asking Better Questions: How Presentation Formats Influence Information Search

Charley M. Wu and Björn Meder
Max Planck Institute for Human Development, Berlin, Germany

Flavia Filimon
Max Planck Institute for Human Development, Berlin, Germany, and Humboldt University

Jonathan D. Nelson
Max Planck Institute for Human Development, Berlin, Germany

While the influence of presentation formats have been widely studied in Bayesian reasoning tasks, we present the first systematic investigation of how presentation formats influence information search decisions. Four experiments were conducted across different probabilistic environments, where subjects ($N = 2{,}858$) chose between 2 possible search queries, each with binary probabilistic outcomes, with the goal of maximizing classification accuracy. We studied 14 different numerical and visual formats for presenting information about the search environment, constructed across 6 design features that have been prominently related to improvements in Bayesian reasoning accuracy (natural frequencies, posteriors, complement, spatial extent, countability, and part-to-whole information). The posterior variants of the icon array and bar graph formats led to the highest proportion of correct responses, and were substantially better than the standard probability format. Results suggest that presenting information in terms of posterior probabilities and visualizing natural frequencies using spatial extent (a perceptual feature) were especially helpful in guiding search decisions, although environments with a mixture of probabilistic and certain outcomes were challenging across all formats. Subjects who made more accurate probability judgments did not perform better on the search task, suggesting that simple decision heuristics may be used to make search decisions without explicitly applying Bayesian inference to compute probabilities. We propose a new take-the-difference (TTD) heuristic that identifies the accuracy-maximizing query without explicit computation of posterior probabilities.

Keywords: information search, presentation formats, Bayesian reasoning, probability gain, optimal experimental design

Supplemental materials: http://dx.doi.org/10.1037/xlm0000374.supp

Before it is possible to arrive at the correct answer, one must first find the right question. The ability to ask good questions is essential for cognition and decision-making, because the choice of query determines what information is acquired, influencing all subsequent inferences and decisions. Consider a doctor choosing a test for diagnosing a patient, where a medical test is an example of a search query with binary outcomes (e.g., positive or negative test results). Not all tests are equally useful, and a doctor must consider the probabilities and diagnostic implications of each test outcome to identify the best test for diagnosing a patient. The term "information search" applies to any decision-making task where the goal is to actively acquire information, including directing eye movements toward informative parts of a scene (Legge, Klitz, & Tjan, 1997; Najemnik & Geisler, 2005, 2008; Nelson & Cottrell, 2007; Renninger, Verghese, & Coughlan, 2007) or conducting experiments to differentiate between competing hypotheses (Lindley, 1956; Slowiaczek, Klayman, Sherman, & Skov, 1992).

Choosing the right search query requires evaluating the usefulness, or, more precisely, *expected* usefulness of each potential query, because the outcome of a query is not known before it is

This article was published Online First March 20, 2017.

Charley M. Wu and Björn Meder, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany; Flavia Filimon, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development and Berlin School of Mind and Brain, Humboldt University; Jonathan D. Nelson, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development.

This research was supported by Grant ME 3717/2-2 to BM and Grant NE 1713/1-2 to JDN from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). This article is based in part on CMW's master's thesis at the University of Vienna, and early results were presented at the SPUDM 2015 conference in Budapest, Hungary. We thank Matthias Gloel for designing and piloting a preliminary version of the Turtle Island experimental scenario, and Gary Brase, Michelle McDowell, Patrice Rusconi, Laura Martignon, Katya Tentori, and Vincenzo Crupi for useful feedback.

Correspondence concerning this article should be addressed to Charley M. Wu, Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: cwu@mpib-berlin.mpg.de

posed. Forming an expectation about the usefulness of a query depends on the probabilities of each outcome and their implications for the hypotheses under consideration (a *preposterior analysis*; Raiffa & Schlaifer, 1961). Computationally, calculating the expected usefulness of a query can be demanding, with part of the complexity arising from the derivation of posterior probabilities (e.g., probability that a patient has the disease given a positive test result). There is a large literature on elementary Bayesian reasoning (for reviews see Barbey & Sloman, 2007; Brase & Hill, 2015; Koehler, 1996), studying the influence of presentation formats on the ability to derive a posterior probability; however, little is known about how presentation formats affect human behavior in other kinds of probabilistic reasoning tasks, such as information search. Currently, the only literature on this topic has shown that 1–2 hr of experience-based learning (i.e., sequential presentation of naturally sampled stimuli with immediate feedback) greatly increased the proportion of accuracy-maximizing search queries when compared with using numerical conditional probabilities (Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Our key goal was to fill this gap in the literature by systematically investigating how various design features of presentation formats influence information search behavior. We specifically focused on descriptive presentation formats, including both numerical and visual formats, because they have the advantage of not requiring the extensive training time of experience-based learning.

Presentation Formats and Bayesian Reasoning

Human performance in Bayesian reasoning tasks can be substantially improved by presenting information in terms of *natural frequencies* (Gigerenzer & Hoffrage, 1995; Meder & Gigerenzer, 2014; for a meta-analysis see McDowell & Jacobs, 2016). Natural frequencies represent the Bayesian reasoning problem in terms of joint frequencies (e.g., number of patients who test positive and have the disease), which resembles how an individual experiences events in daily life (Cosmides & Tooby, 1996). Natural frequencies also simplify the calculations necessary to apply Bayes's rule for deriving a posterior probability, because base rate information is preserved (Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; Kleiter, 1994). This is in contrast to the *standard probability format*, which provides information in terms of conditional probabilities (e.g., probability of a positive test result given that the individual has the disease), which requires introducing base rate information via Bayes's rule (Gigerenzer & Hoffrage, 2007). To illustrate, consider a Bayesian reasoning task with binary hypotheses (disease or no disease) and a single binary outcome (positive or negative test result). The posterior probability that a patient has the disease given a positive test result can be computed using Bayes's rule:

$$P(\text{disease} \mid \text{pos}) = \frac{P(\text{disease})P(\text{pos} \mid \text{disease})}{P(\text{disease})P(\text{pos} \mid \text{disease}) + P(\text{no disease})P(\text{pos} \mid \text{no disease})} \quad (1)$$

In Equation 1, the use of conditional probabilities requires explicitly considering base rate information, $P(\text{disease})$ and $P(\text{no disease})$. Substituting natural frequencies for conditional probabilities, the same posterior probability can be computed without reintroducing the base rate:

$$P(\text{disease} \mid \text{pos}) = \frac{N(\text{pos} \wedge \text{disease})}{N(\text{pos} \wedge \text{disease}) + N(\text{pos} \wedge \text{no disease})} \quad (2)$$

where N denotes the number of cases for each combination of disease and positive test result.

The set of presentation formats that have been shown to improve Bayesian reasoning also includes formats that visualize natural frequencies (e.g., bar graphs and icon arrays), which have systematically yielded superior estimation accuracy over numerical representations (using only words and numbers) in Bayesian inference tasks (Ancker, Senathirajah, Kukafka, & Starren, 2006; Brase, 2009; Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero & Hoffrage, 2013; Sedlmeier & Gigerenzer, 2001; but see Martire, Kemp, Sayle, & Newell, 2014). Additionally, studies where subjects sequentially experienced single events naturally sampled from the environment also yielded improvements over the standard probability format (Lindeman, van den Brink, & Hoogstraten, 1988; Medin & Edelson, 1988).

From Bayesian Reasoning to Information Search

While many Bayesian reasoning tasks in cognitive psychology deal with probability judgments about a single binary outcome, here we are concerned with information search in classification problems (Skov & Sherman, 1986), which are more complex probabilistic reasoning tasks. In the studies presented here, subjects were asked to choose between two information search queries, each with binary outcomes. The goal was to choose the query that would be most useful for improving classification accuracy. To determine which query is most useful, one must consider the probability of each outcome (e.g., probability of a positive or negative test result), as well as the usefulness of each outcome (e.g., the informational value of a test result for diagnosing a patient) for the purpose of making a classification decision. Thus, this type of information search task is considerably more complex than elementary Bayesian reasoning tasks, because it requires reasoning about two binary outcomes instead of one, interpreting the usefulness of each outcome, and forming an expectation to derive the overall usefulness of the query.

There are many ways to define the usefulness of a query outcome (e.g., the reduction of uncertainty or improvement of classification accuracy), corresponding to different information search goals (for reviews, see Nelson, 2005, 2008). Because it is outside the scope of this study to determine which goal is normatively appropriate or optimal, or which goal is the most accurate description of human search behavior in general, we used an information search task where the goal was explicitly defined as the maximization of classification accuracy. This goal corresponds to selecting queries according to their expected increase in classification accuracy (i.e., *probability gain* or any equivalent metric; Baron, 1985). Nelson and colleagues (2010) found probability gain to be the best account for how humans select queries in probabilistic categorization tasks, when information about the search environment was acquired through experience-based learning.

Goals and Scope

How do presentation formats influence information search decisions? We systematically examine both numerical and vi-

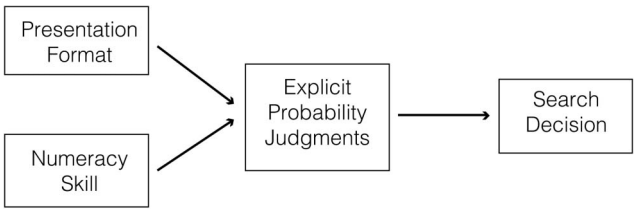


Figure 1. The hypothesized relationships between presentation format, numeracy, probability judgment accuracy, and information search.

sual methods of presenting probabilistic information to assess the important design features that could help facilitate better information search decisions. We address two main questions. First, how are search decisions influenced by presentation formats and design features? Second, is search behavior mediated by probabilistic reasoning, numeracy skill, or both? The first question serves the practical purpose of identifying how information should be presented to improve search behavior. Because natural frequencies and visualizations thereof have been shown to improve people's ability to calculate posterior probabilities (Gaissmaier, Wegwarth, Skopec, Müller, & Broschinski, 2012; Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2002; Micallef, Dragicevic, & Fekete, 2012), these formats may also positively influence information search behavior. Our second question examines the basis of how information search is related to probabilistic reasoning and numeracy skill. Because the usefulness of a query depends on the statistical structure of the environment, we elicited judgments about the relevant probabilistic variables that made up the search environment, with the hypothesis that more accurate probability judgments would be a predictor for better search decisions. We also hypothesized that probability judgment accuracy would be a function of both presentation format and individual numeracy skill, as numeracy has been found to be positively related with Bayesian reasoning accuracy (Chapman & Liu, 2009; Hill & Brase, 2012; McNair & Feeney, 2015) and other decision making tasks (Bodemer, Meder, & Gigerenzer, 2014; Peters et al., 2006; Reyna, Nelson, Han, & Dieckmann, 2009). Figure 1 summarizes our initial hypotheses about the relationships between key variables.

In the following section, we present prominent models of information search and describe how they can make different predictions about which query is most useful. Subsequently, we describe the theoretical motivations behind the presentation formats we investigated and present them in detail. Finally, we describe four experiments investigating the influence of these presentation formats on human search behavior, each using the same stimulus and procedure, but with different probabilistic environments.

Models of Information Search

We considered both Bayesian statistical models and simple heuristic strategies for modeling search behavior. Bayesian *Optimal Experimental Design* (OED) models provide a means to quantify the usefulness of a query, based on the probabilistic structure of the environment. The OED framework has been widely used to construct normative and descriptive models of human information acquisition (Baron, 1985; Bramley, Lagnado, & Speekenbrink, 2015; Klayman & Ha, 1987; Nelson, 2005; Savage, 1954; Skov & Sherman, 1986; Slowiaczek et al., 1992). OED models currently provide the best available computational-level description (Marr, 1982) of many probabilistic information search tasks (Gureckis & Markant, 2012; Nelson et al., 2010; Ruggeri, Lombrozo, Griffiths, & Xu, 2016; Ruggeri & Lombrozo, 2015; for reviews see Markant & Gureckis, 2012; Nelson, 2005). Yet, it has also been shown in some cases that simple heuristic strategies can approximate or even exactly implement particular OED models, thereby establishing a link to psychologically plausible mechanisms (Navarro & Perfors, 2011; Nelson, 2005, 2008; Nelson, Meder, & Jones, 2016). In the following sections, we introduce prominent OED and heuristic models of information search.

OED Models

OED models quantify the expected usefulness of possible queries within a Bayesian decision-theoretic framework (Savage, 1954). For the purposes of our study, each OED model is a candidate descriptive model of how people might intuitively evaluate the expected usefulness of a query. We describe several OED models, each using a distinct *informational utility function*[1] (also known as sampling norm; Nelson, 2005). We use capital Q to denote a query, where lowercase $q_1, \ldots, q_m$ are the m possible outcomes of the query (e.g., medical test results, or forms of a features). Because the search tasks presented in our experiments are classification tasks, we use $C = \{c_1, \ldots, c_n\}$ to denote the different hypotheses (i.e., categories). Equation 3 shows the general framework used by all OED models to quantify the expected usefulness of a query, $eu(Q)$:

$$eu(Q) = \sum_{j=1}^{m} P(q_j)u(q_j) \tag{3}$$

The various OED models differ in how they calculate the usefulness of each individual query outcome, $u(q_j)$, which may lead to disagreements about which query is most useful, corresponding to different information acquisition goals (Nelson, 2005; see Figure 2). We describe several prominent OED models below.

Probability gain. Probability gain (PG; Baron, 1985; Nelson, 2005) values a query in terms of its expected improvement in classification accuracy, assuming that the most probable category will always be chosen. The model's informational utility function is shown in Equation 4, where the max operators choose the leading (i.e., most likely) hypothesis given the outcome of a query and the initially leading hypothesis before any query. The difference between the two terms is the probability gain of a query outcome:

$$u_{PG}(q_j) = \max_i P(c_i|q_j) - \max_i P(c_i) \tag{4}$$

Probability gain corresponds to what Martignon and Krauss (2003) have called average validity. The highest probability gain

[1] Note that we only consider disinterested utility functions, which do not factor in situation-specific payoffs or search costs (see Meder & Nelson, 2012). Our experiments use a cover story where payoffs correspond to classification accuracy and there are no explicit search costs.

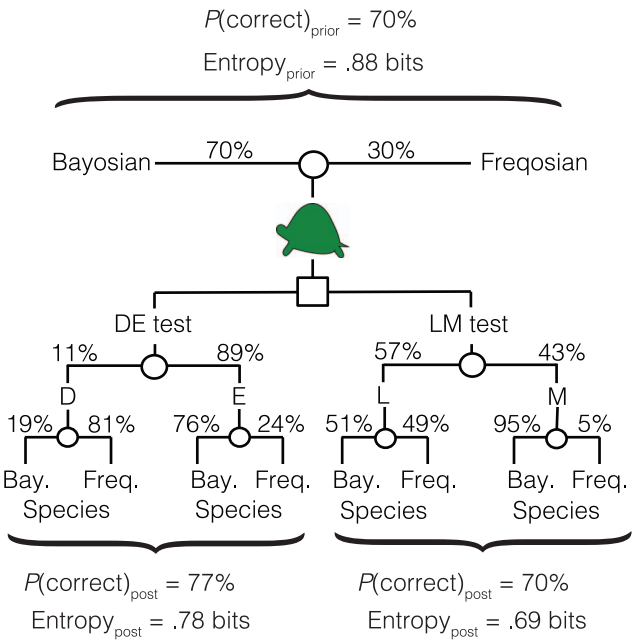


Figure 2. An illustration of how models can make divergent predictions about which query is more useful. In Experiment 1, probability gain predicts selection of the DE test, while information gain predicts selection of the LM test. Probability gain measures the usefulness of a query based on how much the outcome is expected to increase the probability of making a correct classification, $P(\text{correct})_{post} - P(\text{correct})_{prior}$, while information gain measures usefulness in terms of the reduction in Shannon entropy (in bits), $\text{Entropy}_{prior} - \text{Entropy}_{post}$. Although the LM test has a larger expected reduction in entropy, there is no change in expected classification accuracy, whereas both outcomes of the DE test result in an increase in classification accuracy. Thus, probability gain predicts selection of the DE test, while information gain predicts the LM test. See the online article for the color version of this figure.

query is the query that results in the highest expected improvement in classification accuracy over the initial best guess (i.e., relying solely on base rate information). Probability gain has been used as model for a variety of information acquisition tasks, such as experience-based categorization (Nelson et al., 2010), the prediction of eye movements where the goal is to find a target in a cluttered environment (Najemnik & Geisler, 2008), and in medical test selection (Baron, Beattie, & Hershey, 1988). In all of our experiments in this article, we explicitly identify the goal of the information search task as the maximization of classification accuracy, which implies that probability gain should be used to identify the most useful query.

Information gain. Information gain (IG) quantifies how much a query outcome reduces the uncertainty about the hypotheses, where uncertainty is measured using Shannon (1948) entropy[2] (Lindley, 1956):

$$u_{IG}(q_j) = \sum_{i=1}^{n} P(c_i) \log_2 \frac{1}{P(c_i)} - \sum_{i=1}^{n} P(c_i|q_j) \log_2 \frac{1}{P(c_i \mid q_j)} \quad (5)$$

While the reduction of Shannon entropy can correspond to improved classification accuracy, sometimes information gain and probability gain strongly disagree about which query is more useful (Figure 2; see Nelson, 2005; Nelson et al., 2010). In Experiments 1 and 2, we specifically studied search behavior in environments where the query with the highest expected reduction of Shannon entropy does not increase expected classification accuracy at all. It should also be noted that the expected Kullback-Leibler (KL) divergence of a query (Kullback & Leibler, 1951) is exactly equivalent to its expected information gain (Oaksford & Chater, 1996), although the usefulness of individual outcomes may differ. Therefore, when we describe information gain making a prediction about query selection, it should be understood that KL divergence always makes the same prediction.

Impact. Impact quantifies the usefulness of a query as the absolute change in beliefs (Nelson, 2005, 2008; Wells & Lindsay, 1980), from the prior probability to the posterior probability of the hypotheses conditional on a query outcome (Equation 6):

$$u_{Impact}(q_j) = \sum_{i=1}^{n} |P(c_i|q_j) - P(c_i)| \quad (6)$$

In the case of a binary category classification task where the base rates are equiprobable (Experiment 4), the highest impact query also has the highest probability gain (Nelson, 2005). If the base rates are not equiprobable, probability gain and impact can disagree about which query is more useful, as illustrated by the search environments used in Experiments 1 and 2.

Heuristic Models

In addition to several OED models, we considered the possibility that human search behavior might be best described with simple heuristic strategies. In principle, all of the strategies we consider, OED models and heuristics alike, can be applied to all presentation formats and probabilistic environments in our experiments. However, the heuristics operate directly on specific subsets of the environmental probabilities, for example test likelihoods, and may be easier to use given particular presentation formats. More important, under certain conditions heuristic strategies implement the same behavior as OED models (Klayman & Ha, 1987; Navarro & Perfors, 2011; Nelson, 2005; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Nelson et al., 2016).

Likelihood difference heuristic. The likelihood difference heuristic (likDiff; Nelson, 2005; Slowiaczek et al., 1992) chooses the query with the largest absolute difference in feature likelihoods (i.e., test outcomes), for either query outcome. We can describe this using an expected usefulness function, analogous to the OED models, even though the likelihood difference heuristic does not explicitly compute expectations:

$$eu_{likDiff}(Q) = |P(q_1|c_1) - P(q_1|c_2)| = |P(q_2|c_1) - P(q_2|c_2)| \quad (7)$$

The likelihood difference heuristic has been proven to invariably select the query with the highest impact[3] (Nelson, 2005). In a medical diagnosis scenario, one could apply the likelihood difference heuristic by selecting the test with the largest absolute difference between the

[2] We use Shannon entropy with $\log_2$ to measure information gain in bits, although the choice of logarithm base is arbitrary.

[3] One caveat is that the likelihood difference heuristic only applies for binary classification tasks with binary features, whereas impact can apply in situations with multivalued features and any number of categories (Nelson, 2005).

true positive, $P(\text{positive}|\text{disease})$, and the false positive rate, $P(\text{positive}|\text{no disease})$.

Probability of certainty heuristic. The probability of certainty heuristic (ProbCertainty) selects the query (if any) with the highest probability of an outcome granting certainty about the true hypothesis (Nelson et al., 2010). A query outcome has certainty when the posterior probability of one of the hypotheses given an outcome q_j is 1. Analogous to the likelihood difference heuristic, we can describe the probability of certainty heuristic using the OED framework by assigning the usefulness of a query outcome to 1 if it implies certainty about a hypothesis, and zero otherwise:

$$u_{ProbCertainty}(q_j) = \begin{cases} 1 & \text{if } \max_{i=1}^{n} P(c_i|q_j) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Note that an informational utility function numerically very similar to probability of certainty can be derived by substituting a high order Tsallis (1988) entropy in place of Shannon entropy within the information gain model (Crupi, Nelson, Meder, Cevolani, & Tentori, 2016). Thus, probability of certainty could also be viewed as a type of generalized information gain model. The probability of certainty heuristic was tested in Experiments 2 and 3, where we introduced certain query outcomes, and explored how the possibility of obtaining a certain outcome influenced search behavior.

Information Search Scenario

We devised an information search task involving the classification of artificial turtle stimuli to examine the influence of presentation formats on search behavior.[4] In all experiments, participants were told that 100 turtles live on a remote island, with each turtle belonging to either the "Freqosian" or "Bayosian" species. The two species of turtles look identical, but differ in their genetic makeup. Thus, identification of a turtle can be aided with the help of one of two genetic tests, which constitute the available information search queries. The DE test yields truthful information about whether the D form or the E form of the DE gene is present in the turtle, while the LM test yields truthful information about the L form or the M form of the LM gene. While each turtle possesses one form of the DE gene and one form of the LM gene, the extent to which these gene forms are present varies between the two species of turtles. Thus, each test outcome provides probabilistic information about the true species of the turtle. This is a binary information search task, where there are two possible classes, $C = \{Bayosian, Freqosian\}$, and two possible search queries, Q_{DE} and Q_{LM}, with each query having two possible outcomes, $Q_{DE} = \{D, E\}$ and $Q_{LM} = \{L, M\}$. Participants were given the explicit goal of choosing the test that maximizes the probability of correctly classifying a randomly encountered turtle from the island.

Presentation Formats

We presented information about the search scenario using 14 different numeric and visual formats. While many presentations formats have been studied in the context of Bayesian reasoning, including tree diagrams (Binder, Krauss, & Bruckmaier, 2015; Sedlmeier & Gigerenzer, 2001), signal detection curves (Cole, 1989; Cole & Davidson, 1989), and Bayesian boxes (Burns, 2004), our formats were chosen for the purpose of systematically studying the influence of six design features (see Table 1) that have been prominently related to improvements in elementary Bayesian reasoning. All design features are binary and are linked to theories about how people perform probabilistic reasoning, such as presenting natural frequencies instead of conditional probabilities, or highlighting part-to-whole information. All formats provide complete information about the probabilistic search environment, and can be used to implement any of the OED or heuristic strategies. However, the formats are not equivalent in the ease of extracting or computing specific pieces of information about the search environment, such as test likelihoods or posterior probabilities. In Herbert Simon's (1978) terms, the formats are informationally equivalent, but not necessarily computationally equivalent. Thus, people's tendency to use particular heuristic or OED strategies could depend on which format is used to present the probabilistic task information.

Numerical formats. Four conditional probability formats and four natural frequency formats collectively comprise the numerical formats, which use words and numbers to express probabilistic information (see Table 2). All conditional probabilities were rounded to the nearest whole percentage point, while natural frequencies were derived from a population of 100 turtles. Each set of four formats was constructed using a 2×2 factorial design, (a) varying the type of information presented (either likelihoods or posteriors) and (b) varying the presence or absence of complement information about each binary variable. These two factors are described in detail in the subsequent design feature section.

Visual formats. We tested six different visual formats, comprised of two types of icon arrays, two types of bar graphs, and two types of dot diagrams (see Table 3). Each visual format has a likelihood variant and a posterior variant, providing a visual representation of the corresponding natural frequency format (FreqLik+ or FreqPost+). Stimuli were grouped by species (likelihoods) or by test outcome (posteriors) in the same way as the numerical natural frequency formats. The colors used in all visual formats were defined using a tool called Paletton (http://paletton.com), which allowed us to construct a symmetrically divergent and colorblind-friendly palette.

Icon arrays and *bar graphs* are two of the most common visualizations explored in the Bayesian reasoning literature (Ancker et al., 2006; Brase, 2009; Gaissmaier et al., 2012; Galesic et al., 2009), using the number of icons or the length of a bar to represent the number of data points with each joint outcome (i.e., each relationship between species and test outcome). Icon arrays and bar graphs are unique out of the 14 formats, because they utilize spatial extent (i.e., the length of a bar or icon array) to visualize natural frequency information. We used a variant of icon arrays that are more comparable with bar graphs (i.e., a separate array for each joint outcome), rather than stacked into a single unit (in contrast to Brase, 2009; Galesic et al., 2009; Garcia-Retamero & Hoffrage, 2013), allowing us to examine differences in countability (see Table 1).

We introduce the *dot diagram* as a novel format inspired by Euler diagrams (alternatively called "Venn diagrams"), which have been found to be successful at improving Bayesian reasoning

[4] For screenshots of the experiment see supplemental material Figures 1–3.

Table 1
Design Features of Presentation Formats

	Numerical formats								Visual formats					
	Conditional probabilities				Natural frequencies				Icon arrays		Bar graphs		Dot diagrams	
Design feature	ProbLik	ProbLik+	ProbPost	ProbPost+	FreqLik	FreqLik+	FreqPost	FreqPost+	IconLik	IconPost	BarLik	BarPost	DotLik	DotPost
Natural frequencies	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Posteriors	0	0	1	1	0	0	1	1	0	1	0	1	0	1
Complement	0	1	0	1	0	1	0	1	1	1	1	1	1	1
Spatial extent	0	0	0	0	0	0	0	0	1	1	1	1	0	0
Countability	0	0	0	0	1	1	1	1	1	1	0	0	1	1
Part-to-whole information	0	0	0	0	1	1	1	1	0	0	0	0	1	1

Note. 1 denotes the presence of a design feature and 0 denotes the absence.

(Brase, 2009; Micallef et al., 2012; Sloman, Over, Slovak, & Stibel, 2003). One main difference is that Euler diagrams present only one particular result of a binary query, whereas the dot diagram is designed to show both of the possible test outcomes equivalently, to fairly present an information search task, as opposed to a Bayesian reasoning task. The dot diagram uses Uniform Poisson Disk Sampling (Lagae & Dutré, 2008) to place the individual dots in an approximation of a uniform random distribution. Dots (each representing a single item) are distributed within containers that highlight the part-to-whole relationships between species and test outcomes. To avoid idiosyncrasies of a particular random distribution of dots affecting search behavior, we generated 20 dot diagrams for each set of probabilities and selected one at random for each participant assigned to the dot diagram condition.

Design Features

Natural frequencies. Conditional probabilities present different quantitative information than natural frequencies (both numeric and visual representations, where all visual formats in this investigation are representations of natural frequencies). *Conditional probability* formats normalize likelihoods and posterior probabilities (i.e., to the interval [0, 1]) irrespective of the prior or marginal probabilities (Gigerenzer & Hoffrage, 2007), whereas *natural frequencies* express information about likelihoods and posteriors without normalization, incorporating the base rate and marginal probabilities. Natural frequencies are constructed as outcomes of natural sampling (Kleiter, 1994) and are frequently associated with higher Bayesian reasoning accuracy than conditional probabilities (Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2002; Zhu & Gigerenzer, 2006; for reviews see Brase & Hill, 2015; McDowell & Jacobs, 2016). We hypothesized that the advantage of natural frequencies would also carry over to our information search task.

Posterior information. For all formats, we tested a variant presenting information in terms of likelihoods and a variant presenting posterior probabilities. The *likelihood variants* provide information about the base rates (i.e., the distribution of the turtle species before any tests being performed), along with the likelihoods of each test outcome relative to a given species (e.g., the likelihood of a D outcome of the DE test if the turtle is Bayosian). The *posterior variants* provide information about the marginals (i.e., the marginal probability of a test outcome independent of species), as well as the posterior probability of a turtle belonging to a species given a specific outcome (e.g., the probability that a turtle is Bayosian given the D outcome of the DE test). It should be understood that because natural frequencies do not renormalize information, the same set of numerical quantities are presented in both likelihood and posterior variants, but are grouped differently (grouped by species or by outcome, respectively), whereas the conditional probability formats present distinctly different numerical quantities (because of normalization). We hypothesized that posterior variations may be more helpful for information search tasks, since all of the OED models make use of the posterior probabilities in the calculation of usefulness. If the posteriors do not need to be inferred, this simplifies the required computations. However, the likelihood difference heuristic does not require posterior probabilities, and it is possible that other heuristic strategies

Table 2
Numerical Formats

Format	Quantity	Example
Standard probability (ProbLik)	P(species)	Consider a turtle picked at random from the 100 turtles on the island: The probability that it is a Bayosian turtle is 70%.
	P(outcome|species)	If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is 3%.
Standard probability with complement (ProbLik+)	P(species)	Consider a turtle picked at random from the 100 turtles on the island: The probability that it is a Bayosian turtle is 70%, and the probability that it is a Freqosian turtle is 30%.
	P(outcome|species)	If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is 3%, and the probability that it has the E form is 97%.
Posterior probability (ProbPost)	P(outcome)	Consider a turtle picked at random from the 100 turtles on the island: The probability that it has the D form of the DE gene is 11%.
	P(species|outcome)	If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is 19%.
Posterior probability with complement (ProbPost+)	P(outcome)	Consider a turtle picked at random from the 100 turtles on the island: The probability that it has the D form of the DE gene is 11%, and the probability that it has the E form is 89%.
	P(species|outcome)	If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is 19%, and the probability that it is a Freqosian turtle is 81%.
Natural frequency (FreqLik)	N(species)	Out of the 100 turtles on the island, 70 are Bayosian turtles.
	N(outcome $\wedge$ species)	Out of the 70 Bayosian turtles, 2 turtles have the D form of the DE gene.
Natural frequency with complement (FreqLik+)	N(species)	Out of the 100 turtles on the island, 70 are Bayosian turtles and 30 are Freqosian turtles.
	N(outcome $\wedge$ species)	Out of the 70 Bayosian turtles, 2 turtles have the D form of the DE gene, and 68 turtles have the E form of the gene.
Posterior frequency (FreqPost)	N(outcome)	Out of the 100 turtles on the island, 11 have the D form of the DE gene.
	N(species $\wedge$ outcome)	Out of the 11 turtles with the D form of the DE gene, 2 are Bayosian turtles.
Posterior frequency with complement (FreqPost+)	N(outcome)	Out of the 100 turtles on the island, 11 have the D form of the DE gene, and 89 turtles have the E form of the gene.
	N(species $\wedge$ outcome)	Out of the 11 turtles with the D form of the DE gene, 2 are Bayosian turtles and 9 are Freqosian turtles.

Note. Examples of conditional probability and numerical natural frequencies values from Experiment 1. The corresponding values for all experiments are available in supplemental material Tables 1–4. The short names for each format are provided in brackets below the full name. $P(\bullet)$ refers to the probability of an item and $N(\bullet)$ refers to the natural frequency of an item.

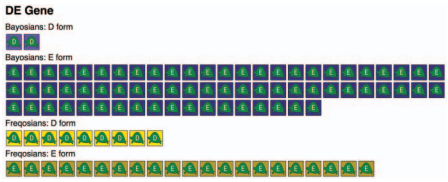
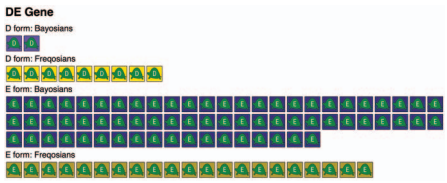
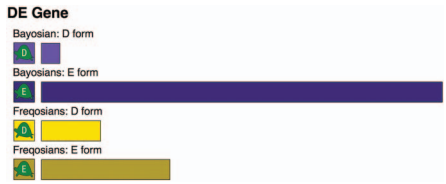
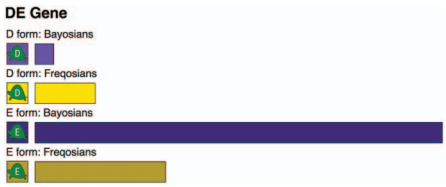
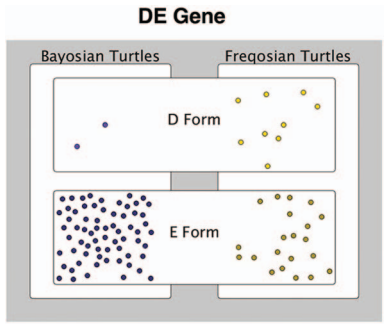
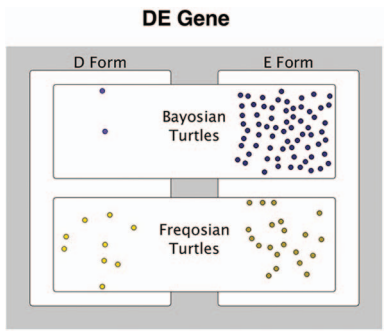
making the same predictions as OED models would not need to compute posteriors either.

Complement. Each variable in the probabilistic search environment is binary, that is, a turtle is either Bayosian or Freqosian. We tested the differences between presenting information *with complement* (e.g., 70% of turtles are Bayosian and 30% of turtles are Freqosian) and presenting information *without complement* (e.g., 70% of turtles are Bayosian). Rusconi and McKenzie (2013) have shown that the inclusion of complement information with the standard probability format improved sensitivity to the informativeness of an answer, which we hypothesized would also be helpful for making search decisions. This design feature was not manipulated in the visual formats, all of which intrinsically include complement information.

Spatial extent. The icon arrays and bar graphs presented in this investigation convey quantitative information using the spatial extent of an array of icons or the length of a bar, where there is a fixed ratio of area to length. In contrast, the numerical formats use words and numbers to convey information, and the dots in the dot diagrams are restricted to a container of fixed size, conveying quantitative information using density rather than spatial extent. Empirical work by Cleveland and McGill (1985; Heer & Bostock, 2010) has shown that graphical representations exploiting basic perceptual abilities, such as length comparisons, can engage automatic visual perceptual mechanisms, thus, reducing the required mental computation and leading to higher quantitative reasoning capabilities. Perceptual accuracy is highest when judging lengths against a common scale (e.g., when comparing the lengths of bars in a bar graph with a common axis), and progressively worse for comparing area (e.g., circles) and volumes or densities (Ancker et al., 2006). This would suggest that visualizations of natural frequencies using spatial extent (icon arrays and bar graphs) would reduce the computational complexity of the task and lead to better search decisions than formats using numbers (conditional probabilities and natural frequencies) or density (dot diagrams).

Countability. By countability we refer to the presentation of quantitative information in discrete units. Conditional probabilities are not countable because they represent statistical information on a continuous scale (i.e., range from 0 to 1), whereas the numerical natural frequencies use words and numbers to present discrete frequencies. However, when natural frequencies are translated into visualizations, countability is not necessarily preserved. With respect to the six design features (see Table 1), the only difference between the bar graphs and icon arrays in our experiments is that the discrete number of icons in an array can easily be counted, whereas the length of a bar represents information on a continuous scale rather than in discrete units. If the countability of formats has a positive effect on Bayesian reasoning or information search, we would expect to see higher performance for icon arrays compared with bar graphs. The dot diagrams are countable in principle, although the random distribution makes it considerably more difficult to arrive at the exact number. Brase (2009) found that dotted Venn

Table 3
Visual Formats in Experiment 1

Format	Example
Icon array (IconLik)	
Posterior icon array (IconPost)	
Bar graph (BarLik)	
Posterior bar graph (BarPost)	
Dot diagram (DotLik)	
Posterior dot diagram (DotPost)	

Note. Full format examples for all experiments available in supplemental material Tables 1–4. Short names for each format are provided in brackets below the full name. See the online article for the color version of this figure.

diagrams sometimes resulted in better Bayesian reasoning performance than Venn diagrams solely using the area of a container to communicate quantitative information, while Stone et al. (2003) independently proposed that the ability to derive exact quantities from a visual format accounted for increased risk perception (i.e., increased willingness to pay for a product with a lower posterior probability of a negative outcome). These hypotheses were further tested by Micallef et al. (2012), who found that Euler diagrams with randomly distributed dots (similar to dotted Venn diagrams) led to the highest Bayesian reasoning performance, compared with other variations that incorporated ordered icon arrays and better facilitated counting, in contrast to Brase (2009). Thus, it is not yet clear whether countability is important or helpful for Bayesian reasoning, although it may be beneficial in a search task.

Part-to-whole information. By part-to-whole information we refer to the ability to relate the proportion of a set of objects (e.g., patients who test positive and have the disease) to any larger set of objects (e.g., patients who test positive). A review by Ancker and colleagues (2006) on the effectiveness of visual formats in conveying health-related risk proposed that part-to-whole information substantially improved risk perception. Similar notions of "nested-set relations" (Neace, Michaud, & Bolling, 2008; Sloman et al., 2003; Stone et al., 2003) or the "subset principle" (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999) have also been proposed as explanation for the effectiveness of natural frequencies in Bayesian reasoning tasks, which by definition contain part-to-whole information, whereas conditional probabilities do not (Gigerenzer & Hoffrage, 2007). Because numeric formats are also differentiated by other design features, we used differences between visual formats to examine the influence of part-to-whole information on search decisions. We designed the dot diagrams to provide accessible information about the part-to-whole relationships between all features of the environment, for example, that the number of Bayosian turtles with the D form of the DE gene represents a portion of the total number of Bayosian turtles, as well as a portion of the total number of turtles with the D form.

In contrast, the design of our icon arrays and bar graphs do not present part-to-whole information with the same accessibility as the dot diagrams, because each combination of gene form and species is expressed as a separate entity. Numeric natural frequencies are always expressed in relation to the larger set of objects (e.g., out of the 70 Bayosian turtles, 2 turtles have the D form of the DE gene), and provide more accessible part-to-whole information than icon arrays or bar graphs. If part-to-whole information has a positive influence on search behavior, we would expect to see more correct search choices for the dot diagrams and numeric natural frequencies than for icon arrays or bar graphs.

Experiments

We conducted four different experiments using the same procedure, but each with a unique probabilistic environment. We used optimal experimental design principles to generate statistical environments in which various subsets of OED and heuristic models made divergent predictions about which query is more useful. Accordingly, these environments are not necessarily representative of the distribution of search environments in the real world, but are meant to "stress test" the presentation formats in scientifically interesting cases where competing models disagree. Table 4 provides the parameters for each experiment, while Table 5 provides an overview of model predictions across the search environments.

Experiment 1

In Experiment 1 we used a search environment where probability gain maximally disagreed with information gain, under the constraint that no queries can lead to certain outcomes.[5] The result is a search environment where probability gain predicts selection of the DE test, while information gain, impact, and the likelihood difference heuristic make the opposite prediction (LM test); probability of certainty makes no prediction in this experiment because no query outcome provides certainty about the true species. Here we present the general methods used in all experiments.

Method

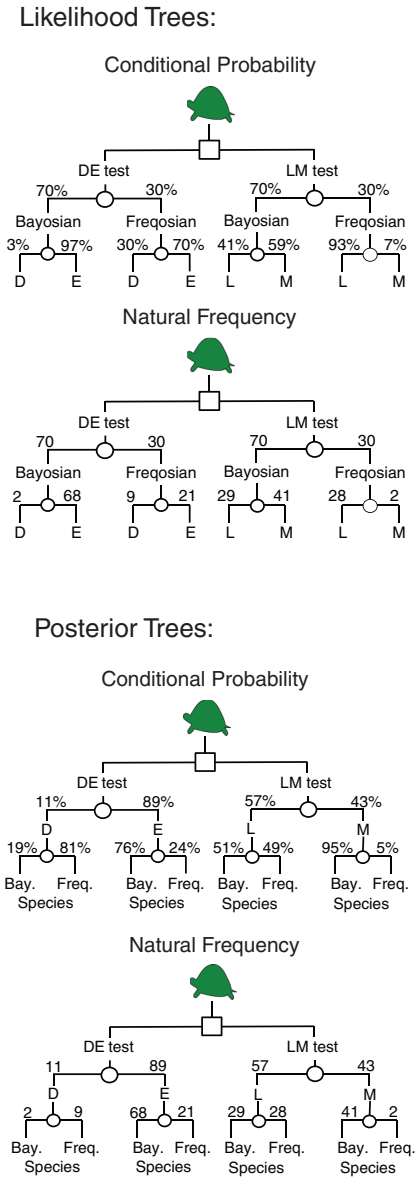
Participants and design. All experiments were conducted using the Amazon Mechanical Turk (AMT) Platform. The Ethics Committee of the Max Planck Institute for Human Development approved the methodology and all participants consented to participation through an online consent form at the beginning of the survey. Human Intelligence Tasks (HITs) were published exclusively to experienced AMT workers who had completed at least 1,000 HITs and had at least a 95% acceptance rate on previous tasks. In total, 821 participants completed Experiment 1, with four excluded because of missing data or because of a self-reported elementary or limited understanding of English. The final sample included 817 participants (46% female, median age of 32 years, range 18–76). To make sure there was no overlap of participants between experiments, once a HIT was completed, the subject's AMT worker ID was added to the exclusion criteria for subsequent studies. Each participant who completed the HIT was paid a fixed amount of $1.50 USD, meeting the minimum hourly wage recommended by Paolacci, Chandler, and Ipeirotis (2010) for experiments on the AMT platform. The same exclusion criteria were applied to all other experiments. Table 6 provides a full description of demographic information for all experiments.

In Experiment 1, participants were randomly assigned to 1 of 14 different presentation formats upon accepting the HIT. In subsequent experiments, we adopted a prerandomized list, to better equalize the number of participants within each condition. Within each presentation format, participants were also randomly assigned to 1 of 16 different randomizations of the probability values, to

[5] A two feature, binary category search environment can be fully described using a single prior and four likelihood probabilities, assuming class-conditional independence (Jarecki, Meder, & Nelson, 2016). We randomly generated 1 million random variations of these five probabilities under the constraint that all posterior probabilities belonged in the range of .05 and .95. The environment used in Experiment 1 had the largest pairwise disagreement strength between probability gain and information gain, where we measured the preference strength of a model as the normalized difference between the expected usefulness of two queries, $PStr_m = 100(eu_m(Q_1) - eu_m(Q_2))/maxPstr_m$, and disagreement strength as the geometric mean between the preference strength of two opposing models m1 and m2, $DStrm1m2 = (|PStrm1| \times |PStrm2|)0.5$, if $PStrm1 \times PStrm2 \leq 0$. See Nelson et al. (2010) for a full description of this process.

Table 4
Search Environments

Parameters	Probabilities				Tree diagrams for Exp. 1
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	
Base rate					
P(Bayosian)	.7	.7	.72	.5	
P(Freqosian)	.3	.3	.28	.5	
Likelihoods					
P(D|Bayosian)	.03	.04	.03	.1	
P(E|Bayosian)	.97	.96	.97	.9	
P(D|Freqosian)	.3	.37	.83	.8	
P(E|Freqosian)	.7	.63	.17	.2	
P(L|Bayosian)	.41	.43	.39	.1	
P(M|Bayosian)	.59	.57	.61	.9	
P(L|Freqosian)	.93	1	1	.3	
P(M|Freqosian)	.07	0	0	.7	
Marginals					
P(D)	.11	.14	.25	.45	
P(E)	.89	.86	.75	.55	
P(L)	.57	.6	.56	.2	
P(M)	.43	.4	.44	.8	
Posteriors					
P(Bayosian|D)	.19	.2	.09	.11	
P(Freqosian|D)	.81	.8	.91	.89	
P(Bayosian|E)	.76	.78	.94	.82	
P(Freqosian|E)	.24	.22	.06	.18	
P(Bayosian|L)	.51	.5	.5	.25	
P(Freqosian|L)	.49	.5	.5	.75	
P(Bayosian|M)	.95	1	1	.56	
P(Freqosian|M)	.05	0	0	.44	



Note. Squares in the tree diagrams denote decision nodes and circles denote chance nodes. See the online article for the color version of this figure.

avoid possible confounds with aspects of the stimulus (e.g., the naming of the turtle species, the naming of the genes, and choice of colors in the visual formats).[6] For the analyses, the 16 randomizations were recoded to match a single canonical variation, described in Table 4. For simplicity, we present each of the experiments using a randomization where selecting the DE test is the correct choice given the goal of maximizing classification accuracy.

Procedure and materials. Once participants gave their consent to participate in the study, they were presented with the Turtle

[6] For example, swapping the query outcomes for the DE and LM test changes which query is most useful respective to each of the considered models. We then recoded the test selection and the probability judgments relative to the assigned randomization.

Table 5
Expected Utilities and Query Predictions for Each Search Environment

Experiment	Model	Prediction (test)	DE test					LM test				
			P(D)	*u*(D)	*P*(E)	*u*(E)	***eu*(DE)**	*P*(L)	*u*(L)	*P*(M)	*u*(M)	***eu*(LM)**
1	Probability gain	DE	11%	.11	89%	.06	**.07**	57%	−.19	43%	.25	0
	Information gain	LM		.18		.09	.1		−.12		.6	**.19**
	Impact	LM		1.02		.12	.22		.38		.5	**.44**
	Prob. certainty	—		0		0	0		0		0	0
2	Probability gain	DE	14%	.1	86%	.08	**.08**	60%	−.2	40%	.2	0
	Information gain	LM		.16		.12	.13		−.12		.88	**.28**
	Impact	LM		1		.16	.28		.4		.6	**.48**
	Prob. certainty	LM		0		0	0		0		1	**.4**
3	Probability gain	DE	25%	.19	75%	.22	**.21**	56%	−.22	44%	.28	0
	Information gain	DE		.44		.51	**.49**		−.14		.86	.29
	Impact	DE		1.26		.44	**.65**		.44		.56	.5
	Prob. certainty	LM		0		0	0		0		1	**.44**
4	Probability gain	DE	45%	.39	55%	.32	**.35**	20%	.25	80%	.06	.1
	Information gain	DE		.5		.32	**.4**		.19		.01	.05
	Impact	DE		.78		.64	**.7**		.5		.12	.2
	Prob. certainty	—		0		0	0		0		0	0

Note. Predictions for each model correspond to the search environments described in Table 4. For each test, *P*(•) denotes the probability of the outcome, *u*(•) denotes the utility of the outcome, and *eu*(•) denotes the expected utility of the test. All of the OED models compute *eu*(•) through a normalized mean, weighting each individual outcome utility, *u*(•), by the probability of the outcome, *P*(•). The expected utility for the most useful query is shown in bold. The likelihood difference heuristic invariably makes the same prediction as impact (proof in Nelson, 2005), so it is not listed separately. The probability of certainty (prob. certainty) heuristic makes no predictions in Experiments 1 and 4, because neither of the queries contains a certain outcome.

Island story. Before proceeding, participants had to correctly answer a comprehension question to ensure that they had understood the nature of the task. Specifically, participants were asked one of four randomly selected questions about the binary nature of the DNA test, for example, "If a turtle does not have the D form of the DE gene, then which form of the gene does it have?" Responses were selected from one of the following options: "D form," "E form," "L form," or "M form." Participants were required to correctly answer this question before continuing the study. If an incorrect answer was given, participants were asked to reread the instructions and attempt a new randomly selected question.

Search task and probability judgments. After completing the comprehension question, participants were given information on the environmental probabilities in one of the 14 randomly assigned formats, and asked to choose the test that would yield the highest chance of correctly classifying the species of a randomly selected turtle. Specifically, they were asked: "Which test is better for having the highest chance of correctly classifying the animal as either a Freqosian or a Bayosian turtle?"

Once participants made a search decision, they were asked to give their estimates for 11 different probabilistic variables in the search environment. To avoid potential memory confounds, the assigned presentation format was shown again for the probability judgment task. The questions were arranged in blocks, with three questions referring to the prior probability of the species and the

Table 6
Participant Demographics

Variable	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Final *N*	817	683	681	677
Completed HITs	821	690	684	686
Excluded	4	7	3	9
Average completion time in minutes (±*SD*)	13.8 (±7.7)	14 (±19.9)	13.1 (±8.8)	14.5 (±8.5)
Gender	46% female	45% female	48% female	50% female
Age				
Median	32	31	31	32
Mean	34	34	34	35
Range	18–76	18–83	18–73	18–71
Education				
High school	14%	15%	14%	12%
Some university	34%	33%	37%	39%
Bachelor's degree or higher	51%	52%	48%	49%
Other	1%	0%	1%	1%

Note. Participants were excluded because of missing data or a self-reported elementary or limited understanding of English. Percentages may not add up to 100% because of rounding.

marginal probability of the outcomes of each test, four questions concerning the likelihood of a test outcome given the species, and four questions regarding the posterior probabilities of the species given the test outcomes (supplemental material Figures 4–6). The order of the three blocks was randomized across participants. Responses were given using a visual analog scale slider. As the slider was adjusted, text displayed the two complementary values of the slider (e.g., "Bayosian turtles: 70%" and "Freqosian turtles: 30%") corresponding to its position to the nearest percentage point.

Numeracy test. Subsequently, participants completed a numeracy test to assess their ability to comprehend and reason with numerical information. Numeracy has been shown to correlate positively with accuracy in Bayesian reasoning (Brown et al., 2011; Hill & Brase, 2012) and other tasks (Peters et al., 2006). We used a hybrid test consisting of the Schwartz, Woloshin, Black, and Welch (1997) numeracy test and the adaptive Berlin Numeracy Test (BNT; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). The choice to combine the two tests follows the recommendation of Cokely and colleagues (2012), who found that this method offers the best discriminability specific to the broad range of numeracy levels within an AMT sample population. The BNT is well suited for discriminating among highly numerate populations, whereas the Schwartz et al. test is better for discriminating among relatively less numerate populations. The scores for the two tests were added together, yielding a combined numeracy score in the range of 1 to 7. Upon completion of the numeracy test, participants provided demographic information and were debriefed.

Results

Information search decisions. Figure 3A shows the results of the information search task, with the height of the bars representing the proportion of correct (i.e., accuracy-maximizing) queries for each format. Overall, 61% of participants chose the correct query, with the

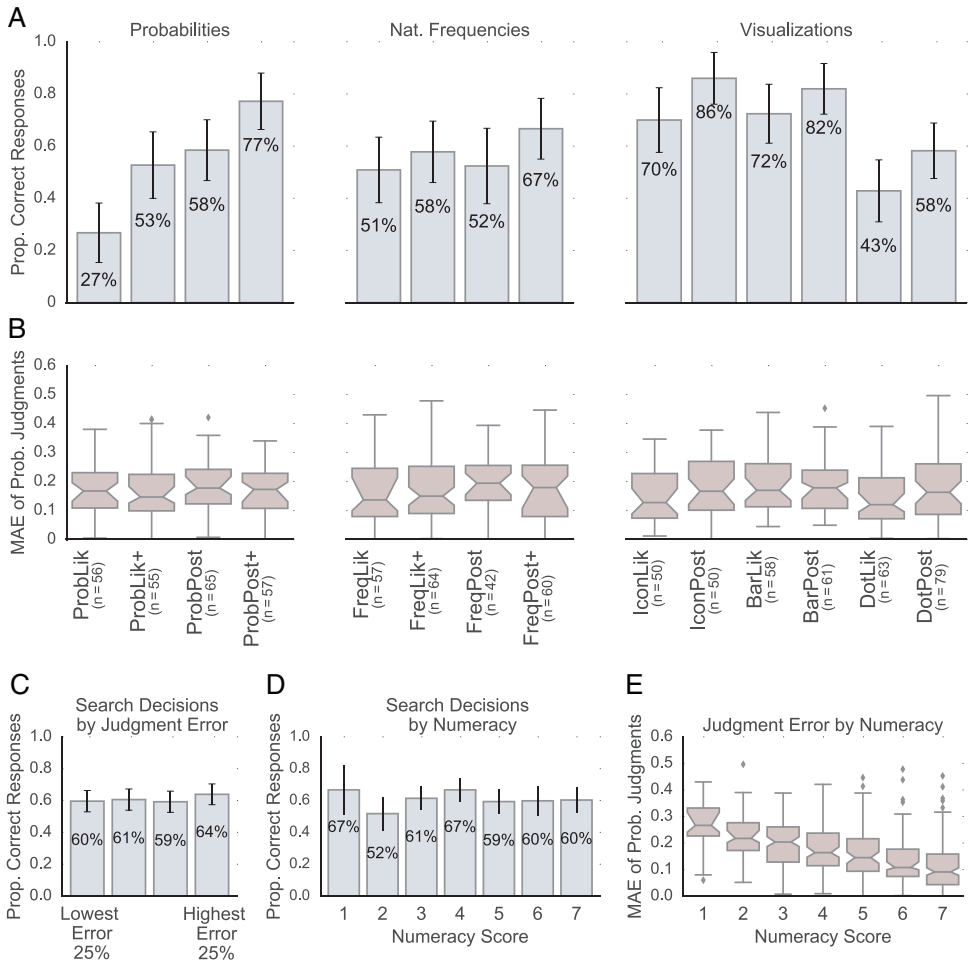


Figure 3. Experiment 1 results. (A) Proportion of correct (accuracy-maximizing) search queries by format, (B) probability judgment error by format, (C) correct search responses by judgment error split into quartiles, (D) correct search responses by numeracy score, and (E) probability judgment error by numeracy. Bar graphs displays Agresti-Coull 95% confidence interval (CI). Box plot whiskers indicate 1.5 IQR, with the notch illustrating bootstrapped 95% CI of the median (10k replications). See the online article for the color version of this figure.

proportion varying between 27% and 86% across formats. A χ^2 test found that search decisions were strongly affected by format, $\chi^2(13, N = 817) = 78.57$, $p < .001$. The standard probability format (ProbLik) was the least helpful format for identifying the correct query, with only 27% of participants making this choice. This result was virtually identical to Nelson et al. (2010) who used the standard probability format in a similar search task. Participants given the posterior icon array (82% correct choices) or posterior bar graph (86% correct choices), both visual formats, achieved the highest proportion of correct choices, and were able to meet the same levels of performance as those who had undergone 1–2 hr of experience-based learning in Nelson et al. (2010).

Consistent with the findings reported in the Bayesian reasoning literature (e.g., Gigerenzer & Hoffrage, 1995), the natural frequency format (FreqLik) yielded more correct responses (51%) than the standard probability format (ProbLik; 27%). However, different variants of numerical natural frequency formats did not consistently outperform the different variations of conditional probability formats. The posterior probability with complement (ProbPost+) was most effective out of all numerical formats (77% correct choices), suggesting that conditional probabilities can be helpful in information search tasks if the complement is added and if posterior probabilities are given, rather than in the form of priors and likelihoods associated with the standard probability format.

We conducted a logistic regression to model search decisions, using the design features of the assigned presentation format, individual numeracy level, and probability judgment error as predictors (Table 7). Design features were coded as a binary vector relative to each presentation format (see Table 1), while numeracy levels ranged from 1 to 7, and mean absolute error was used as the measure of probability judgment error. The regression analysis provides a comparison of search behavior at the level of design features, with the result that spatial extent, posterior information, and including complements were the strongest predictors for correct search decisions. For visual formats, using spatial extent and posterior groupings (IconPost and BarPost) were the most effective at soliciting correct search decisions, while for numeric formats, presenting posterior and complement information (ProbPost+ and FreqPost+) were also relatively effective. Our regression model also indicates that the use of natural frequencies was not a reliable predictor for search behavior, in contrast to our initial hypothesis. Countability and part-to-whole information also failed to consistently predict correct search decisions.

Formats and probability judgments. In contrast to search behavior (i.e., identification of the most useful test), probability judgment accuracy did not vary substantially across formats (Figure 3B; distribution of probability judgment error per format). Accordingly, a one-way between-participants analysis of variance (ANOVA) found that the mean absolute error[7] of probability judgments was not significantly influenced by format, $F(13, 804) = 1.15$, $p = .31$. As a manipulation check, we tested differences between formats on specific subsets of probability questions. Likelihood formats (numerical and visual) had less error than their posterior counterparts on the base-rate questions, $t(816) = -.73$, $p < .001$ and on the likelihood question, $t(816) = -.33$, $p < .001$, whereas the posterior variants had lower error on the posterior questions than the likelihood variants, $t(816) = -.66$, $p < .001$. Likelihood and posterior formats did not systematically differ on the marginal probability questions,[8] $t(816) = -.8$, $p = .42$. Consistent with findings from Bayesian reasoning studies (Brase & Hill, 2015; Gigerenzer & Hoffrage, 1995), natural frequency formats (FreqLik, FreqLik+) had less error on posterior probability judgments than their conditional probability counterparts (ProbLik, ProbLik+), $t(231) = -3.45$, $p < .001$. However, the natural frequency formats were not better when aggregated over all probability judgments questions.

Probability judgments and search decisions. In contrast to our initial hypothesis there was virtually no correlation ($r_{pb} = .04$) between probability judgment error and the proportion of correct search decisions (Figure 3C; participants split into quartiles based on judgment error). This is supported by the regression analysis, which controls for design features and individual numeracy level (see Table 7). Similar results are obtained when judgments are broken down by question type (i.e., base rate, marginals, likelihoods, and posteriors). This suggests that contrary to our hypothesis lower probability judgment error does not necessarily lead to better search decisions.

Numeracy. There was no correlation between numeracy and search decisions ($r_{pb} = .003$; Figure 3D), with the regression model also finding no significant relationship (see Table 7). However, higher numeracy was correlated with lower error on the probability judgment task (Pearson $r = -.4$; Figure 3E). Thus, participants with higher numeracy performed better on the probability judgment task, but had no advantage on the information search task.

Discussion

Experiment 1 tested a search environment where the correct response (DE test) as predicted by probability gain was in contradiction to the predictions made by information gain, impact, and the likelihood difference heuristic (LM test). Search decisions were strongly influenced by presentation format, with the posterior icon array and posterior bar graphs providing the most effective means to help participants identify the correct test. The key advantage of these graphical formats is that they were able to elicit the same (correct) intuitions about the relative usefulness of a query as experience-based learning (Nelson et al., 2010), but without the same time requirements (several hundred trials of training, over 1–2 hr). Spatial extent, posteriors, and complements were the strongest predictors for correct search decisions. Contrary to our initial hypotheses, neither numeracy nor probability judgment accuracy were reliable predictors of search behavior. Most participants presented with the posterior icon array or the posterior bar graph were able to identify the correct query, even though they were not

[7] Using mean squared error (*MSE*) to measure performance on the probability judgment tasks yields equivalent results in all experiments. We use MAE because it provides a more intuitive representation of the magnitude of error.

[8] The manipulation check results were replicated in all subsequent experiments, with the exception of Experiment 3, where the posterior formats had significantly lower error on the marginal questions than the likelihood formats, $t(680) = 2.40$, $p = .017$. We do not separately report the corresponding t tests for the subsequent experiments.

Table 7
Logistic Regression Results

	Dependent variable: Correct search decision					
	Model					
Variable	(1) Experiment 1	(2) Experiment 2	(3) Experiment 3	(4) Experiment 4	(5) All experiments	(6)
Design features						
Natural frequencies	−.114 (.375)	.028 (.354)	−.127 (.362)	.333 (.454)	.034 (.177)	.042 (.187)
Posteriors	.662*** (.150)	.232 (.162)	.223 (.163)	−.076 (.202)	.287*** (.077)	.306*** (.082)
Complement	.483** (.179)	−.040 (.201)	−.114 (.198)	.196 (.247)	.120 (.094)	.133 (.099)
Spatial extent	.961*** (.271)	.533* (.250)	.197 (.263)	−.265 (.296)	.343** (.126)	.378** (.133)
Countability	.056 (.330)	.127 (.291)	.497 (.305)	−.744* (.376)	.007 (.150)	.009 (.158)
Ind. differences						
Numeracy	.043 (.048)	−.035 (.053)	.140** (.053)	.361*** (.066)	.072** (.025)	.101*** (.026)
Probability judgments (MAE)	.953 (.837)	.994 (.880)	−.409 (.802)	−1.910 (1.183)	−.049 (.420)	.209 (.441)
Environment						
Certainty						−.973*** (.083)
OED model disagreement						−1.081*** (.083)
Constant	−.766* (.349)	−.930* (.397)	−.253 (.384)	.437 (.446)	−.250 (.180)	.621** (.198)
Observations	817	683	681	677	2,858	2,858
Classification accuracy	.62	.64	.63	.80	.60	.67
Akaike Information Criterion	1,042.758	891.693	890.972	636.189	3,815.619	3,520.556

Note. Log odds are shown with *SE* in brackets. The part-to-whole design feature is not presented as an independent predictor, because it is entirely accounted for by taking natural frequencies without spatial extent. MAE = mean absolute error; OED = Optimal Experimental Design. Classification accuracy is 10-fold cross validation prediction accuracy.
* $p < .05$. ** $p < .01$. *** $p < .001$.

any better at the probability estimation task than participants assigned to other presentation formats.

Experiment 2

In Experiment 2 we introduced the possibility of certain outcomes with respect to the hypotheses under consideration, that is, one of the two search queries had a probabilistic outcome that provided certainty about the true species of a turtle. We used a search environment that was similar to Experiment 1, but adjusted so that the M result of the LM test gives 100% certainty that the turtle belongs to the Bayosian species. However, the M result is only present in 40% of the turtles, and the alternative L result gives maximal uncertainty about the species (50% probability for each species). Even though the LM test has a possibility of certainty, the DE test leads to higher expected classification accuracy. In Experiment 2, the probability of certainty heuristic along with information gain, impact, and the likelihood difference heuristic predict selection of the LM test, while only probability gain predicts selection of the DE test. We expected Experiment 2 to be more difficult than Experiment 1, because of the addition of the probability of certainty heuristic as a model that disagrees with the correct probability gain prediction. In a similar information search task, the probability of certainty heuristic accounted for 42% of search decisions when subjects were given the standard probability format, but only 3% of search decisions for subjects assigned to experience-based learning (Experiments 1 and 2, Condition 4; Nelson et al., 2010). This finding suggests that when the probabilistic structure of the environment is not learned through experience, the possibility of certainty is an important factor in how queries are selected.

Results

Information search decisions. Overall, only 36% of participants chose the correct query, compared with 61% in Experiment 1, with all formats having a comparatively lower proportion of correct responses (Figure 4A). As in Experiment 1, the posterior icon array and posterior bar graph had the highest proportion of correct responses (48% for both). However, there were no reliable differences between formats in terms of search behavior, $\chi^2(13, N = 683) = 17.86$, $p = .16$, which may be because of floor effects. A logistic regression analysis found that only spatial extent was a statistically reliable predictor for correct search decisions (see Table 7). Remarkably, no other design features, numeracy skill, or probability judgment error were reliable predictors for search behavior, which is a strong result given the large sample size of 683 subjects. It seems that the introduction of a certain outcome made it substantially harder for participants to identify the more useful query, and the query with the possibility of certainty was valued beyond its contribution to classification accuracy. We address this possibility in Experiment 3, where we isolate the prediction of the probability of certainty heuristic from all other models.

Probability judgments. Consistent with Experiment 1, probability judgment error was not significantly affected by presentation format, $F(13, 670) = 1.26$, $p = .24$ (Figure 4B), nor was probability judgment error correlated with search decisions ($r_{pb} = .06$; Figure 4C).

Numeracy. Again, there was no correlation between numeracy and the proportion of probability gain search decisions ($r_{pb} = -.06$; Figure 4D), although higher numeracy was correlated with lower error on the probability judgment task (Pearson $r = -.45$; Figure 4E). Both of these results were consistent with the previous experiment.

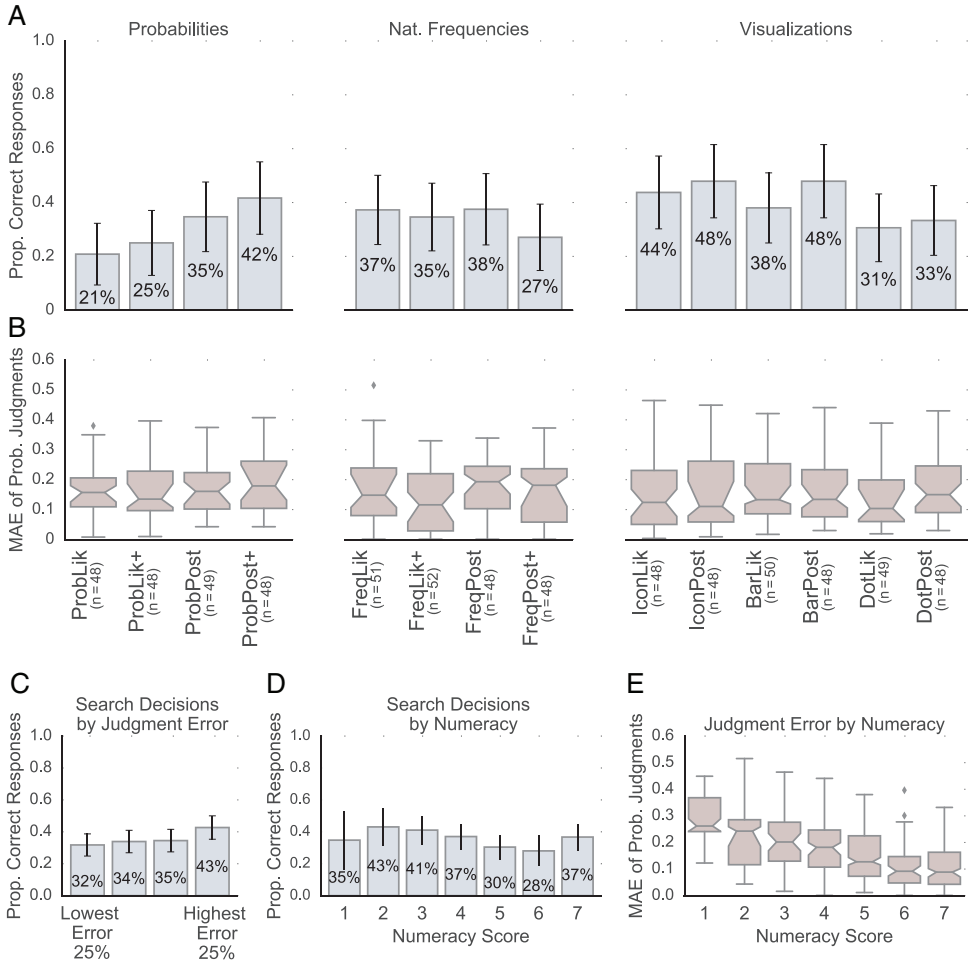


Figure 4. Experiment 2 results. (A) Across all format there was a lower proportion of correct (accuracy-maximizing) search queries compared to Experiment 1. (B) Performance on the probability judgment task replicated Experiment 1 results, with neither judgment error; (C) nor numeracy; and (D) being predictors for search decisions. (E) As in Experiment 1, there was a negative correlation between judgment error and numeracy. See the online article for the color version of this figure.

Discussion

Experiment 2 produced results that resemble Experiment 1, but with the proportion of probability gain choices reduced substantially across all formats. No format led to consistent selection of the accuracy-maximizing query, although formats using spatial extent led to relatively better search performance. The lowered performance across all formats in Experiment 2 indicates that introducing certain outcomes may have contributed to a search problem where the accuracy-maximizing query is substantially less likely to be selected, irrespective of format. Why does the possibility of certainty create such challenging environment when the query that can lead to a certain outcome (LM test) is in opposition to the accuracy-maximizing query (DE test)? To examine how certainty influences query selection by itself, we conducted Experiment 3 to examine search behavior in an environment where the probability of certainty prediction contradicted all other models.

Experiment 3

We conducted Experiment 3 to try to isolate the extent to which the possibility of a certain outcome influences search behavior, across the various presentation formats. As in Experiment 2, the LM test had a chance of yielding an M result, with the implication that the turtle is certainly Bayosian. But again, it was a risky choice, because the alternative L outcome resulted in maximal uncertainty. The main difference in the environmental probabilities, compared with the previous study, was that in Experiment 3, probability gain, information gain, impact, and the likelihood difference heuristic predicted selection of the DE test, while only the probability of certainty heuristic predicted selection of the LM test.

Results

Information search decisions. Overall, 64% of participants chose the correct query, while the remaining 36% of search

decisions were consistent with only the probability of certainty heuristic. The proportion of correct responses varied between 47 and 73% (Figure 5A), but there was no overall difference between formats, $\chi^2(13, N = 681) = 17.91$, $p = .16$. The logistic regression analysis found no statistically reliable predictors for search behavior from the set of design features (see Table 7). Thus, none of the design features nor individual presentation formats we tested yielded a systematic advantage in this particular search environment.

Probability judgments. There were differences between formats on the probability judgment task, $F(13, 668) = 2.32$, $p = .005$; however, the descriptive statistics indicate that the overall effect is mainly because of the likelihood variants of the natural frequency formats (FreqLik, FreqLik+) having lower error than the other formats (Figure 5B). If these two conditions are excluded, an ANOVA shows no significant differences, $F(11, 569) = 1.2$, $p = .28$. No correlation was found between judgment error and search decisions ($r_{pb} = -.07$; Figure 5C), which is consistent with the regression analysis (see Table 7). Together with the previous findings, this suggests that probability judgment error had no bearing on search decisions.

Numeracy. In contrast to Experiments 1 and 2, we found a small correlation between numeracy and search decisions ($r_{pb} = .12$; Figure 5D). This suggests that numeracy may be a predictor for search behavior when the more useful query is relatively obvious, which could be the case in Experiment 3 where probability gain, information gain, impact, and the likelihood difference heuristic all make the same prediction. Indeed, numeracy skill was the only reliable predictor for search decisions in the regression model (see Table 7). On the other hand, as in all previous experiments, numeracy was strongly correlated with performance on the probability estimation task

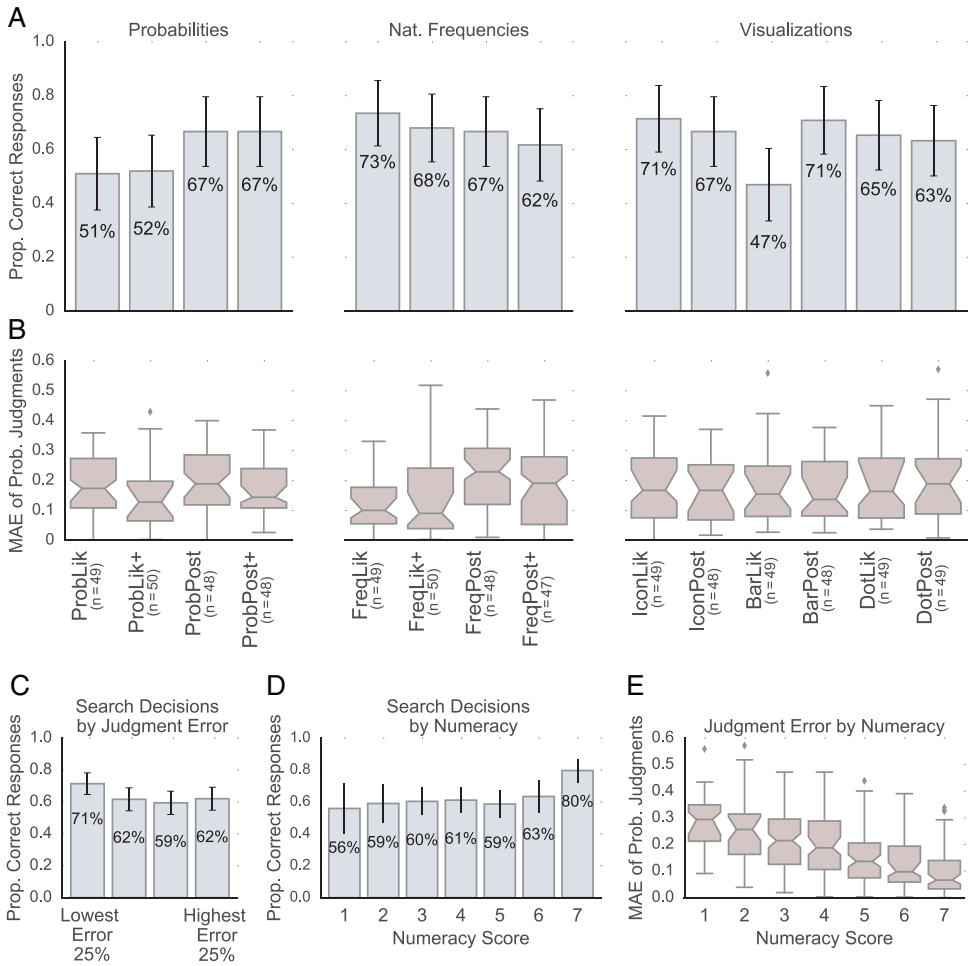


Figure 5. Experiment 3 results. (A) Sixty-four percent of subjects (aggregated across formats) chose the correct accuracy-maximizing query, with the remaining choices consistent with only the probability of certainty heuristic, of the models in consideration. There were no differences in search behavior across formats, although we found that the natural frequency formats (FreqLik and FreqLik+) had lower probability judgment error than the other formats (B). Judgment error was not a predictor for search decisions (C), although higher numeracy was weakly related to better search choices ($r_{pb} = .12$; D). The negative correlation between judgment error and numeracy was replicated (E). See the online article for the color version of this figure.

(Pearson $r = -.47$; Figure 5E). The correlation between numeracy and probability judgment accuracy is a robust finding, which appears to be independent of the task environment.

Discussion

Sixty-four percent of subjects correctly selected the DE test, which was predicted by all of the OED and heuristic models with the sole exception of the probability of certainty heuristic. This suggests that by itself, the possibility of obtaining a certain outcome influenced search decisions a great deal. To understand more precisely how much the possibility of certainty influences behavior, it is important to have an idea of the highest performance that can be obtained. We address this in our final experiment.

Experiment 4

We conducted one further experiment to assess the upper range of performance with respect to the experimental methods and subject population, in which there was no disagreement among model predictions. Experiment 4 used an environment with similar model predictions as Experiment 3, but without any certain outcomes. Probability gain, information gain, impact, and the likelihood difference heuristic all made the same prediction (DE test), while the probability of certainty heuristic made no prediction. This experiment was unique in that no model disagreed with the probability gain prediction.

Results

Information search decisions. Participants in all formats reliably chose the correct query (80% overall, range: 73–88%), which was the highest proportion out of all experiments (Figure 6A). In comparison with Experiment 3 (64% correct choices), which had a similar relationship between models, but without the possibility of a certain outcome, we can infer that a strategy corresponding to the probability of certainty heuristic accounted for a difference of about 16 percentage points (95% confidence interval [12, 21], Wilson score interval). There were no differences in search choice between formats,

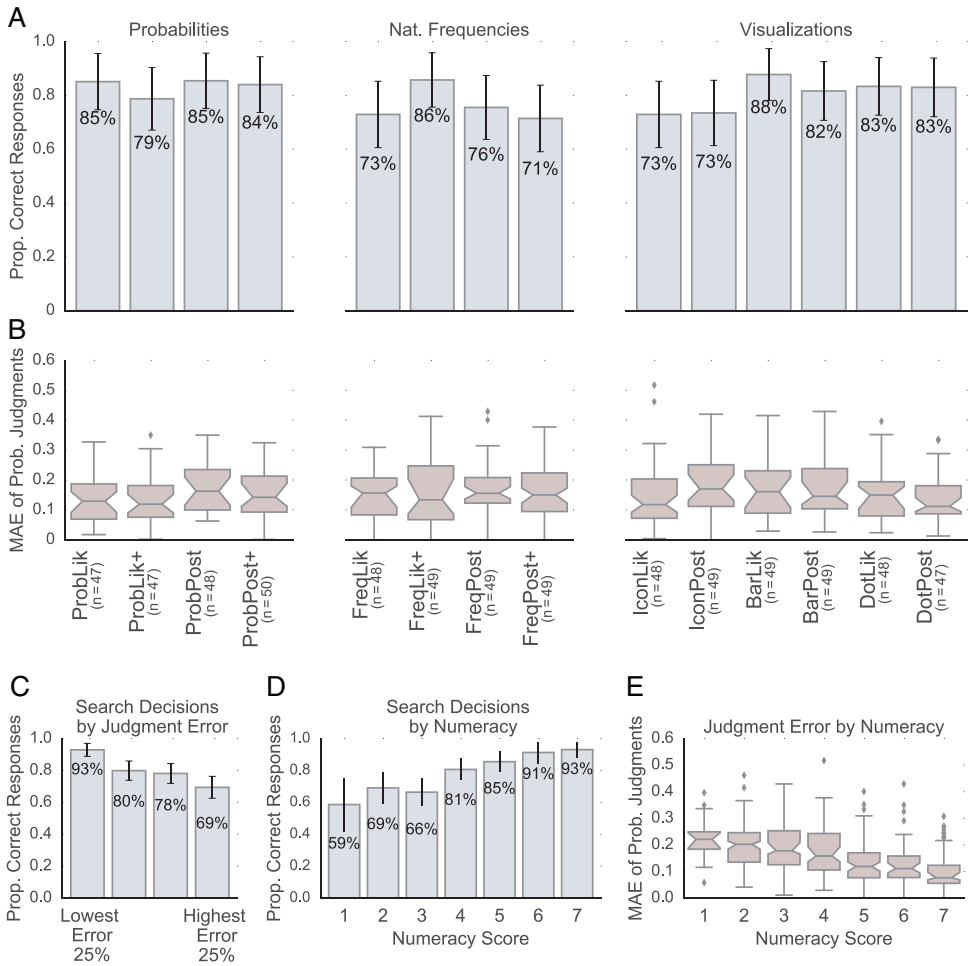


Figure 6. Experiment 4 results. (A) With no divergence between model predictions, the majority of subjects, regardless of format, chose the correct accuracy-maximizing query (80% aggregated across formats). (B) There were no substantial differences in judgment error across formats. (C) In this experiment, there were correlation between (C) search decisions and judgment error ($r_{pb} = -.16$) and (D) between search decisions and numeracy ($r_{pb} = .25$). (E) A negative correlation between judgment error and numeracy was found, as in all previous experiments, suggesting that this is a robust result. See the online article for the color version of this figure.

$\chi^2(13, N = 677) = 12.92, p = .45$. The logistic regression analysis found that from the set of design features, only countability was a reliable predictor; however, countability had a negative effect on performance (see Table 7).

Probability judgments. Format did not have an effect on probability judgments, $F(13, 664) = 1.18, p = .29$ (Figure 6B). Taken together with the previous experiments, our results suggest that no single format has an overall advantage on the probability judgment task, although there were differences for specific subsets of questions. For the first time, we found a weak negative correlation ($r_{pb} = -.16$) between probability judgment error and correct search behavior (Figure 6C); however, when controlling for differences in design features and numeracy skill, our regression model indicates that judgment error is not a reliable predictor for search decisions (see Table 7). In the context of all previous experiments, people who are able to make more accurate probability judgments do not seem to perform any better in the search task.

Numeracy. Higher numeracy scores were correlated with a higher proportion of accuracy-maximizing search decisions ($r_{pb} = .25$; Figure 6D). This was a stronger correlation than in Experiment 3, whereas no correlations were found for Experiments 1 and 2. These results are consistent with the regression models, which found significant effects of numeracy on search behavior in Experiments 3 and 4 (see Table 7). One explanation for this result is that with an easier task (i.e., less disagreement among the models considered), numeracy may be a predictor for search decisions. Numeracy was also correlated with lower error on the probability judgment task (Pearson $r = -.42$; Figure 6E), consistent with all previous experiments.

Discussion

In Experiment 4, where all models predicted selection of the DE test (with the exception of the probability of certainty heuristic that made no prediction), we measured the highest proportion of correct query selections out of all the experiments (80% overall). Countability was the only statistically reliable predictor for search behavior among the design features, but it had a negative contribution toward correct choices. Numeracy was a reliable predictor for search choice, as in Experiment 3. Therefore, one might conclude that high numeracy is helpful in simple environments in which there is no disagreement among the heuristic and OED models' predictions. However, not even the highest level of numeracy was adequate for more difficult probabilistic environments.

General Discussion

Our aims in this study were both practical and theoretical: to explore how search decisions are influenced by presentation formats and design features, while addressing the theoretical question of whether search behavior is mediated by probabilistic reasoning, numeracy skill, or both. We studied how search behavior and probability estimation varied across 14 different numerical and visual presentation formats, in four different search environments. Table 8 presents an overview of statistical tests conducted in each experiment; Figure 7 shows the results aggregated over all experiments.

We extended the logistic regression analysis by aggregating over all four experiments (see Table 7), with Model 5 using the same predictors as in the individual experiments and with

Table 8
Overview of Experiments

Experiment	Correct search queries	Influence of Format on		Probability judgment error and search decisions	Relationship between numeracy and	
		Search decision	Probability judgment error		Search decision	Probability judgment error
1 (n = 817)	61%, range: [27–86%]	$\chi^2(13, N = 817) = 78.57, p < .001$	$F(13,804) = 1.15, p = .31$	$r_{pb} = -.04$	$r_{pb} = .003$	$r = -.40$
2 (n = 683)	36%, range: [21–48%]	$\chi^2(13, N = 683) = 17.86, p = .16$	$F(13,670) = 1.26, p = .24$	$r_{pb} = -.06$	$r_{pb} = -.06$	$r = -.45$
3 (n = 681)	64%, range: [47–73%]	$\chi^2(13, N = 681) = 17.91, p = .16$	$F(13,668) = 2.32, p = .005$	$r_{pb} = -.07$	$r_{pb} = .12$	$r = -.47$
4 (n = 677)	80%, range: [71–88%]	$\chi^2(13, N = 677) = 12.92, p = .45$	$F(13,664) = 1.18, p = .29$	$r_{pb} = -.16$	$r_{pb} = .25$	$r = -.42$
Overall (n = 2,858)	60%, range: [45–71%]	$\chi^2(13, N = 2,845) = 52.66, p < .001$	$F(13,2845) = 2.87, p = .001$	$r_{pb} = -.02$	$r_{pb} = .05$	$r = -.43$

Note. Performance on the probability judgment task is measured in terms of mean absolute error across all 11 questions. Thus, a negative correlation between numeracy and probability judgments signifies that higher numeracy skill is a predictor for lower error.

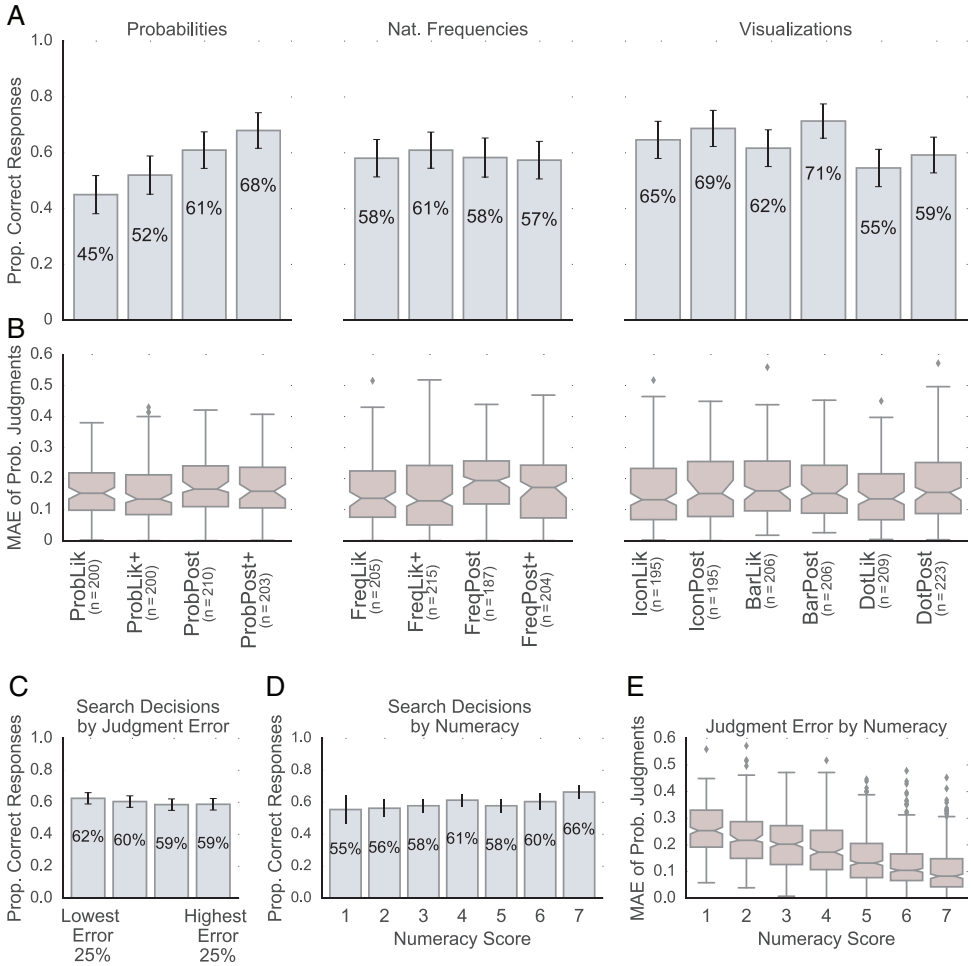


Figure 7. Summary of all experiments. (A) Aggregated over all experiments, the choice of format led to noticeable differences in the proportion of correct search decisions, but not in judgment error (B). Probability judgment error (C) was not correlated with search behavior, while each incremental change in numeracy had a small influence on search decisions (D). High numeracy skill was a robust predictor for lower judgment error (E). See the online article for the color version of this figure.

Model 6 adding two environmental factors (that vary across the individual experiments) as binary predictors. *Certainty* denotes environments where there is the possibility of a query outcome leading to certainty about the true species of a turtle, which in our case was always in opposition to the correct probability gain query (Experiments 2 and 3). *OED Disagreement* denotes environments where there is disagreement between OED models, specifically when the probability gain prediction is in opposition to the prediction made by information gain, impact, and the likelihood difference heuristic (Experiments 1 and 2). These two environmental factors are able to describe all four experiments as a 2 × 2 factorial design. Both aggregate models yield similar estimates of the log odds for shared predictors, although Model 6 achieves higher classification accuracy and a lower AIC by including environmental factors.

The Role of the Environment

Environments where the probability gain prediction disagreed with the probability of certainty heuristic or with other OED models were more difficult, with similar reductions to the log odds of making a correct search decision (see Table 7). Why did these factors contribute toward a substantially more difficult search task? Hogarth and colleagues (Hogarth, Lejarraga, & Soyer, 2015) have introduced the notion of *kind* and *wicked* environments, where wicked environments represent a disjoint between learning and test environments. Are the environments we tested wicked in this sense? We unfortunately do not know the true distribution of search environments in the world. As a first step to address this question, we conducted simulations over 10 million randomly generated environments,[9] and considered the relationship of the predictions made by each of the OED and heuristic models in a pairwise manner. In cases where both competing models

[9] We randomly generated 10 million sets of the five probabilistic variables (prior probability and test likelihoods) that fully describe a search environment, like the ones presented in this study. For each environment, we conducted pairwise comparisons between model predictions. Simulation code is available at https://github.com/charleywu/AskingBetterQuestions. For simulations of a larger variety of OED models, see supplemental material from Nelson (2005).

made a prediction, information gain and probability gain made the same prediction 92% of the time; the probability of certainty heuristic and probability gain made the same prediction 77% of the time; and information gain and the probability of certainty heuristic made the same prediction 83% of the time.

This illustrates a potential tension between using OED principles to test environments where we can disentangle overlapping model predictions, and environments that are representative of search problems experienced in the wild. The use of heuristics, such as the probability of certainty, may represent an ecologically valid strategy because of a correspondence with the correct probability gain query in most environments, trading off computational complexity for accuracy. However, this still poses a challenge to the design of presentation formats for conveying statistical information about search problems, because we are seeking formats that are robust across different types of environment. Indeed, in Nelson et al. (2010) the possibility of certainty did not influence participants assigned to the experience-based learning conditions, who preferentially selected the higher probability gain query in every statistical environment that was tested. We have not identified numeric or graphical formats that consistently lead to as high of rates of selection of the probability gain test as experience-based learning. However, across all our experiments, we found that some formats and design features were better than others in soliciting correct search decisions.

Presentation Formats and Search Decisions

Aggregated over all experiments (Figure 7A), the posterior bar graph and posterior icon array had the highest performance (69 and 71% correct choices, respectively), and the standard probability format (ProbLik) had the lowest performance (45%). The logistic regression analysis shows that from the set of design features, spatial extent and posterior information were the only statistically reliable predictors for search behavior (Model 6, Table 7). Therefore, the present findings suggest that bar graphs or icon arrays are the most helpful for information search problems, where the underlying natural frequency information is conveyed using spatial extent, and the information is grouped by primarily by query outcome, as in our posterior formats.

In contrast to results from the Bayesian reasoning literature and our own initial hypotheses, the other design features did not consistently influence search behavior across environments. Given the large ($N = 2{,}858$) sample size, we take the lack of statistically significant results for natural frequencies, complement information, countability, and part-to-whole information as strong evidence against our hypotheses that these design features (by themselves) facilitate adaptive information search.

Among the classes of presentation formats (conditional probabilities, natural frequencies, and graphical visualizations), the largest variability in performance was found within the four conditional probability formats. Presenting information in terms of posteriors and including complement information led to enhanced performance for the conditional probability formats, although these same design features did not substantially influence the natural frequency formats (Figure 7A). Aggregating across the experiments, the posterior probability with complement format (ProbPost+) led to a similar proportion of correct search choices (68%) as the posterior icon array (69%) and posterior bar graph formats (71%). An interesting finding was that the posterior probability format with complement (ProbPost+) led to a higher proportion of correct choices (68%) than any of the numeric natural frequency formats (that ranged from 57 to 61%). This, combined with natural frequencies failing to be a predictor in the regression model (see Table 7), suggests that numeric natural frequencies are not necessarily more advantageous for information search problems than conditional probabilities—if they are presented in the right way (i.e., not the standard probability format).

Posterior bar graphs, despite not being countable, led to as good of performance as the posterior icon arrays, and to better performance than any of the numeric natural frequency formats. This suggests that countability is not necessary for communicating probabilistic information for information search. Furthermore, out of all the formats we studied, icon arrays and bar graphs were the only formats that represented natural frequencies using spatial extent, a perceptual attribute. Spatial extent was found to be one of the best predictors for correct test selection in the regression analysis, suggesting that it may be an important design feature for communicating probabilistic information in search tasks.

Not all visual formats were successful, with the dot diagrams leading to fewer accuracy-maximizing search decisions than icon arrays or bar graphs. Thus, the increased accessibility of part-to-whole information did not compensate for the lack of spatial extent, which were the two design features differentiating the dot diagrams from icon arrays and bar graphs. Future studies should test whether providing part-to-whole information could be helpful if combined with spatial extent (e.g., using stacked icon arrays).

Information Search and Probabilistic Reasoning

In contrast to our initial hypothesis, the ability to make accurate probability judgments was not related to search behavior (Figure 7C), nor was it a predictor in the logistic regression model, when controlling for differences in design features and numeracy skill. This suggests that presentation formats that improve explicit Bayesian reasoning accuracy do not necessarily transfer to more complex decision-making tasks, such as information search, even though information search can be modeled "as if" it is based on Bayesian reasoning processes. Of course, improving explicit probability judgments may be useful in some situations (e.g., where doctors need to explicitly explain the meaning of a test result in terms of posterior probabilities to a patient). However, people can sometimes make decisions that correspond to the correct application of Bayes's rule without calculating the correct probability estimates (Domurat, Kowalczuk, Idzikowska, Borzymowska, & Nowak-Przygodzka, 2015). In our studies, the ability to reason about information search decisions seems to be largely independent from the ability to explicitly estimate probabilities.

Numeracy was robustly correlated with probability judgment accuracy in all experiments (Figure 7E), and may also be helpful in information search tasks (Figure 7D). We found weak correlations between numeracy and search decisions in Experiments 3 and 4; however, these relationships were not found in Experiments 1 and 2, which, judging by the disagreement among model predictions and the overall rates of probability gain queries, were substantially more difficult. Aggregated across all experiments, numeracy was a statistically reliable predictor for correct search decisions (see Table 7). Using Model 6 from the regression anal-

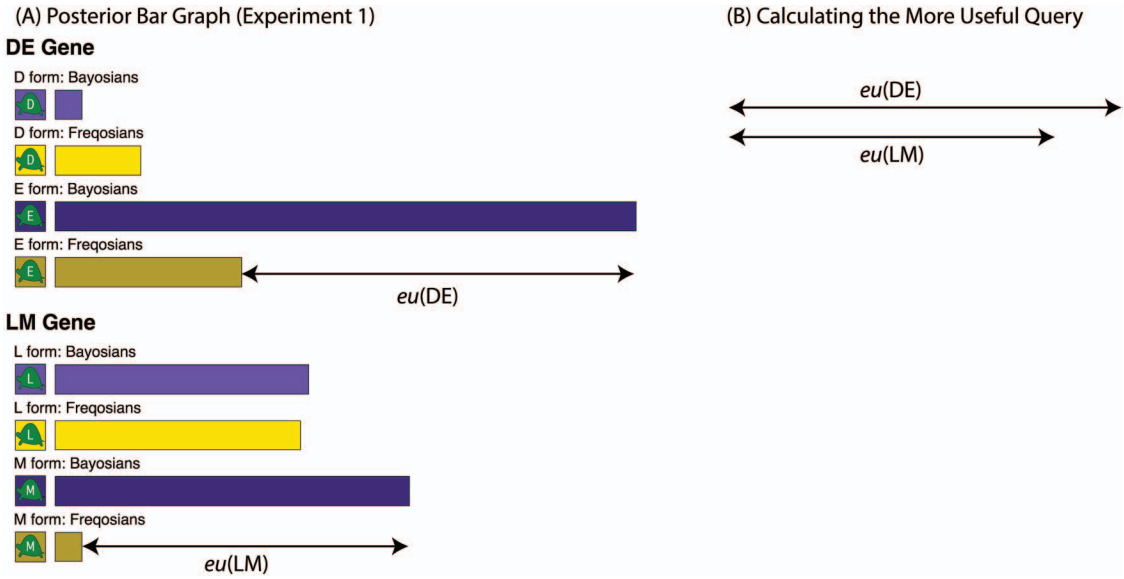


Figure 8. An example of the take-the-difference (TTD) heuristic, applied to the posterior bar graph format from Experiment 1. TTD is a simple heuristic strategy that is capable of identifying the accuracy-maximizing (i.e., probability gain) query, without explicit computation of posterior probabilities. (A) The expected utility of each test, $eu(\cdot)$, can be described as the absolute difference between natural frequencies for the outcome that has the largest difference. In the case of the DE test, the E gene has the largest difference, which can be determined visually by comparing lengths of the "D and Bayosian" bar to the "D and Freqosian" bar, and the "E and Bayosian" bar to the "E and Freqosian" bar. The expected utility of the DE test can be described using a line spanning the distance between the lengths of "E and Bayosian" bar and the "E and Freqosian" bar, since the length of each bar is a visual representation of a natural frequency. In panel (B), the relative usefulness of both tests are compared, and the accuracy-maximizing query can be identified by picking the test with the longest line. See the online article for the color version of this figure.

ysis, which includes environmental factors as predictors, we see that the difference between the lowest and the highest possible numeracy scores (a difference of six levels) can be represented by a sixfold increase in the log odds ($0.101 * 6 = 0.606$), which is similar to the combined influence of spatial extent and posteriors ($0.378 + 0.306 = 0.684$). Thus, an individual with the lowest numeracy skill, given a format with both spatial extent and posterior information, is about as likely to identify the correct search query as an individual with the highest numeracy, but given a format lacking these two design features. The right format may help compensate for low numeracy in information search, though this effect may be limited by individual differences in statistical reasoning. Just as the performance-boosting benefits of numeric presentation formats (i.e., natural frequencies) require a sufficient level of numeracy (Chapman & Liu, 2009), so do visual formats require a sufficient level of graph literacy to yield the correct interpretation (Galesic et al., 2009; Okan & Garcia-Retamero, 2012).

Take-The-Difference Heuristic

Why were some presentation formats more helpful than others, particularly the posterior icon arrays and posterior bar graphs? We discovered a simple heuristic strategy that identifies the higher probability gain query in our tasks, without requiring explicit computation of posterior probabilities.[10] The *take-the-difference* (TTD) heuristic is a relevant strategy for all natural frequency formats, although we believe it is especially well suited for visualizations of natural frequencies using spatial extent. This new heuristic can be executed by choosing the query with the largest absolute difference of natural frequencies, when comparing each query outcome (see Figure 8). We can describe the TTD heuristic using the OED framework, where the expected usefulness of a query is measured as the absolute difference between the joint frequencies $N(c_1 \wedge q_j)$ and $N(c_2 \wedge q_j)$ for the outcome q_j where the absolute difference is largest:

$$eu_{TTD}(Q) = \max_j |N(c_1 \wedge q_j) - N(c_2 \wedge q_j)| \quad (9)$$

This heuristic strategy is most salient for the posterior variations of the natural frequency and visual formats, because $N(c_1 \wedge q_j)$ and $N(c_2 \wedge q_j)$ are grouped together and easy to compare; this may explain why the posterior variations generally yielded more correct responses. Formats using spatial extent to visualize natural frequencies have the added advantage that a visual comparison of the lengths of two bars or arrays of icons is equivalent to calculating the absolute difference between two natural frequencies specified by the TTD heuristic. TTD also offers a potential explanation for

[10] Through simulations over 10 million randomly generated environments, we found that TTD and probability gain never made contradictory predictions. Thus, we conjecture that in every environment where both models make a prediction about query selection, TTD always agrees with probability gain. Simulation code is available at https://github.com/charleywu/AskingBetterQuestions

the poor performance of the dot diagrams in the search task. The presentation of the dots—placed in a random uniform distribution within a container of fixed area—makes the task of comparing absolute differences much less intuitive than comparing relative differences (through a density estimate). Comparing relative differences is undesirable because it renormalizes the natural frequencies as absolute frequencies, and loses the additional information about base rates and the marginals that are contained within the natural frequency representation. Comparing relative differences does not implement the probability gain model and does not consistently yield the same search predictions.

Conclusion

We systematically investigated a wide range of presentation formats across different search environments. Differences in search environments played a large role in influencing search behavior across experiments, where the possibility of certainty and disagreements between OED models contributed toward substantially more difficult tasks. Aggregated over all search environments, formats using spatial extent (icon arrays and bar graphs) and presenting information in terms of posteriors were the most effective design features for improving search behavior, while numeracy skill also made a positive contribution. Because some environments still proved difficult for all formats, future studies may want to examine didactic methods such as story boarding (Ottley et al., 2016) or adding interactivity to visualizations (Tsai, Miller, & Kirlik, 2011) as a manipulation orthogonal to the choice of presentation format, to further improve search performance.

Although many of the presentation formats we tested are useful approximations of how humans might learn about a search problem through experience, there is a notable incongruence between *description* and *experience* (Hertwig & Erev, 2009). In Experiments 2 and 3, each format gave the descriptive information that for the entire *population* of turtles, those with the M gene are invariably Bayosian. This may be qualitatively different from learning about a probabilistic environment through direct experience, where single events are *sampled* from the population. This may explain why certain outcomes did not prove challenging for experience-based learners in a similar information search task (Nelson et al., 2010). This incongruence of representing certain outcomes through descriptive versus experiential information has not played an important role in Bayesian reasoning tasks, since certainty makes deriving posterior probabilities easier, rather than harder. However, this is not the case in information search, specifically when the query with a certain outcome does not lead to the highest *expected* classification accuracy. We recommend future research to examine behavioral differences in judgment and decision making tasks when conveying statistical information about certain or near certain outcomes, across descriptive presentation formats and experience-based learning, and manipulating whether the information is presented as a sample or as representative of the whole population, to examine these issues.

We identified a new information search heuristic, TTD, which chooses the accuracy-maximizing query (i.e., higher probability gain query) without explicitly computing posterior probabilities. The TTD heuristic offers a potential explanation for differences in search behavior across formats, as well as the lack of correlation between probability judgments and search decisions. Thus, we recommend that future studies investigate the effectiveness of the TTD heuristic by providing explicit instruction in its use, together with appropriate visualizations of natural frequencies using spatial extent.

The lack of correlation between probability judgment accuracy and search behavior has the important additional implication that improving explicit probability judgment accuracy is not necessarily the same as improving performance on more complex probabilistic decision-making tasks, such as information search. Thus, we recommend future research on the use of presentation formats for communicating risk to not only address the ability to make explicit probability judgments, but also how formats influence more complex decisions that people make about the world.

References

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association, 13,* 608–618. http://dx.doi.org/10.1197/jamia.M2115

Barbey, A. K., & Sloman, S. a. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and Brain Sciences, 30,* 241–254. http://doi.org/10.1017/S0140525X07001653

Baron, J. (1985). *Rationality and Intelligence*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571275

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning. *Organizational Behavior and Human Decision Processes, 42,* 88–110. http://dx.doi.org/10.1016/0749-5978(88)90021-0

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information: An empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology, 6,* 1186.

Bodemer, N., Meder, B., & Gigerenzer, G. (2014). Communicating relative risk changes with baseline risk: Presentation format and numeracy matter. *Medical Decision Making, 34,* 615–626. http://doi.org/10.1177/0272989X14526305

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 708–731. http://dx.doi.org/10.1037/xlm0000061

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology, 23,* 369–381. http://dx.doi.org/10.1002/acp.1460

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology, 6,* 340.

Brown, S. M., Culver, J. O., Osann, K. E., MacDonald, D. J., Sand, S., Thornton, A. A., . . . Weitzel, J. N. (2011). Health literacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient Education and Counseling, 83,* 92–98. http://dx.doi.org/10.1016/j.pec.2010.04.027

Burns, K. (2004, May 25–28). Painting pictures to augment advice. In *Proceedings of AVI '04, Gallipoli, Italy*, 344–349. http://dx.doi.org/10.1145/989863.989921

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making, 4,* 34–40.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *American Association for the Advancement of Science, 229,* 828–833. http://dx.doi.org/10.1126/science.229.4716.828

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7,* 25–47.

Cole, W. (1989). Understanding Bayesian reasoning via graphical displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 20,* 381–386. http://dx.doi.org/10.1145/67449.67522

Cole, W., & Davidson, J. (1989). Graphic representation can lead to fast and accurate Bayesian reasoning. *Proceedings of the Annual Symposium on Computer Application in Medical Care, 8,* 227–231.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1–73. http://dx.doi.org/10.1016/0010-0277(95)00664-8

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2016). *Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search.* [Manuscript in preparation].

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology, 6,* 1194.

Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, A.-S., Broschinski, S., & Politi, M. C. (2012). Numbers can be worth a thousand pictures: Individual differences in understanding graphical and numerical representations of health-related information. *Health Psychology, 31,* 286–296. http://dx.doi.org/10.1037/a0024850

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association, 28,* 210–216. http://dx.doi.org/10.1037/a0014474

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine, 83,* 27–33. http://dx.doi.org/10.1016/j.socscimed.2013.01.034

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning. *Behavioral and Brain Sciences, 30,* 264–267. http://dx.doi.org/10.1017/S0140525X07001756

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7,* 464–481. http://dx.doi.org/10.1177/1745691612454304

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. *In Proceedings of the 28th annual chi conference on human factors in computing systems* (pp. 203–212). New York, NY: ACM. http://dx.doi.org/10.1145/1753326.1753357

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004

Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 65,* 2343–2368. http://dx.doi.org/10.1080/17470218.2012.687004

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84,* 343–352. http://dx.doi.org/10.1016/S0010-0277(02)00050-1

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24,* 379–385. http://dx.doi.org/10.1177/0963721415591878

Jarecki, J. B., Meder, B., & Nelson, J. D. (2016). *Naïve and robust: Class-conditional independence in human classification learning.* [Manuscript submitted for publication].

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106,* 62–88. http://dx.doi.org/10.1037/0033-295X.106.1.62

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211

Kleiter, G. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York, NY: Springer.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1. http://dx.doi.org/10.1017/S0140525X00041157

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*. Advance online publication. http://dx.doi.org/10.1214/aoms/1177729694

Lagae, A., & Dutré, P. (2008). A comparison of methods for generating poisson disk distributions. *Computer Graphics Forum, 27,* 114–129. http://dx.doi.org/10.1111/j.1467-8659.2007.01100.x

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review, 104,* 524–553. http://dx.doi.org/10.1037/0033-295X.104.3.524

Lindeman, S. T., van den Brink, W. P., & Hoogstraten, J. (1988). Effect of feedback on base-rate utilization. *Perceptual and Motor Skills, 67,* 343–350. http://dx.doi.org/10.2466/pms.1988.67.2.343

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics, 27,* 986–1005. http://dx.doi.org/10.1214/aoms/1177728069

Markant, D., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In N. Miyake, D. Peebles, and R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 719–724). Austin, TX: Cognitive Science Society.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Martignon, L., & Krauss, S. (2003). Can L'Homme Eclaire be fast and frugal? Reconciling Bayesianism and bounded rationality. In S. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 108–122). Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511609978.006

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International, 240,* 61–68. http://dx.doi.org/10.1016/j.forsciint.2014.04.005

McDowell, M., & Jacobs, P. (2016). *Meta-analysis of the effect of natural frequencies on Bayesian reasoning.* [Manuscript submitted for publication].

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review, 22,* 258–264. http://dx.doi.org/10.3758/s13423-014-0645-y

Meder, B., & Gigerenzer, G. (2014). Statistical thinking: No one left behind. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking* (pp. 127–148). The Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-7155-0_8

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

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psy-*

chology: General, 117, 68–85. http://dx.doi.org/10.1037/0096-3445.117.1.68

Micallef, L., Dragicevic, P., & Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics, 18,* 2536–2545. http://dx.doi.org/10.1109/TVCG.2012.199

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434,* 387–391. http://dx.doi.org/10.1038/nature03390

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision, 8,* 4.1–414. http://dx.doi.org/10.1167/8.3.4

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review, 118,* 120–134. http://dx.doi.org/10.1037/a0021110

Neace, W., Michaud, S., & Bolling, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making, 3,* 140–152.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112,* 979–999. http://dx.doi.org/10.1037/0033-295X.112.4.979

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 143–164). Oxford, United Kingdom: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0007

Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing, 70,* 2256–2272. http://dx.doi.org/10.1016/j.neucom.2006.02.026

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition, 130,* 74–80. http://dx.doi.org/10.1016/j.cognition.2013.09.007

Nelson, J. D., McKenzie, C. R., Cottrell, G. W. G., & Sejnowski, T. J. T. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science, 21,* 960–969. http://dx.doi.org/10.1177/0956797610372637

Nelson, J. D., Meder, B., & Jones, M. (2016). *Optimal experimental design, heuristics, and sequential search*. [Manuscript submitted for publication].

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103,* 381–391. http://dx.doi.org/10.1037/0033-295X.103.2.381

Okan, Y., & Garcia-Retamero, R. (2012). When higher bars are not larger quantities: On individual differences in the use of spatial information in graph comprehension. *Spatial Cognition & Computation, 12,* 195–218.

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., . . . Chang, R. (2016). Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Transactions on Visualization and Computer Graphics, 22,* 529–538. http://dx.doi.org/10.1109/TVCG.2015.2467758

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17,* 407–413. http://dx.doi.org/10.1111/j.1467-9280.2006.01720.x

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Cambridge, MA: Division of Research, Graduate School of Business Administration, Harvard University.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision, 7,* 6. http://dx.doi.org/10.1167/7.3.6

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135,* 943–973. http://dx.doi.org/10.1037/a0017327

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition, 143,* 203–216. http://dx.doi.org/10.1016/j.cognition.2015.07.004

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology, 52,* 2159–2173. http://dx.doi.org/10.1037/dev0000240

Rusconi, P., & McKenzie, C. R. M. (2013). Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 66,* 2443–2464. http://dx.doi.org/10.1080/17470218.2013.793732

Savage, L. (1954). *The foundations of statistics. Journal of Consulting Psychology*. New York, NY: Wiley.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127,* 966–972. http://dx.doi.org/10.7326/0003-4819-127-11-199712010-00003

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130,* 380–400. http://dx.doi.org/10.1037/0096-3445.130.3.380

Simon, H. A. (1978). On the forms of mental representation. *Perception and Cognition: Issues in the Foundations of Psychology, 9,* 3–18.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22,* 93–121. http://dx.doi.org/10.1016/0022-1031(86)90031-4

Sloman, S., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91,* 296–309. http://dx.doi.org/10.1016/S0749-5978(03)00021-9

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20,* 392–405. http://dx.doi.org/10.3758/BF03210923

Stone, E. R., Sieck, W. R., Bull, B. E., Frank Yates, J., Parks, S. C., & Rush, C. J. (2003). Foreground:background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes, 90,* 19–36. http://dx.doi.org/10.1016/S0749-5978(03)00003-7

Tsai, J., Miller, S., & Kirlik, A. (2011). Interactive visualizations to improve Bayesian reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55,* 385–389. http://dx.doi.org/10.1177/1071181311551079

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics, 52,* 479–487. http://dx.doi.org/10.1007/BF01016429

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776–784.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition, 98,* 287–308. http://dx.doi.org/10.1016/j.cognition.2004.12.003

Received May 31, 2016
Revision received October 24, 2016
Accepted November 5, 2016 ■