Experience matters:
Information acquisition optimizes probability gain

Jonathan D. Nelson
Max Planck Institute for Human Development

Craig R. M. McKenzie
University of California, San Diego

Garrison W. Cottrell
University of California, San Diego

Terrence J. Sejnowski
Howard Hughes Medical Institute, Salk Institute for Biological Studies
and University of California, San Diego

Wednesday, September 23, 2009

In press, *Psychological Science*.

Ideas and correspondence are welcomed. Please address correspondence to:

Jonathan D Nelson
nelson@mpib-berlin.mpg.de or jonathan.d.nelson@gmail.com
Adaptive Behavior and Cognition Group
Max Planck Institute for Human Development
Lentzeallee 94
14195 Berlin
Germany

Abstract

Deciding which piece of information to acquire or attend to is fundamental to perception, categorization, medical diagnosis, and scientific inference. Four statistical theories of the value of information—information gain, Kullback-Liebler distance, probability gain (error minimization), and impact—are equally consistent with extant data on human information acquisition (Nelson, 2005; 2008). Three experiments, designed via computer optimization to be maximally informative, tested which of these theories best describes human information search. Experiment 1, which used natural sampling and experience-based learning to convey environmental probabilities, found that probability gain explained participants' information search better than the other statistical theories or the probability of certainty heuristic. Experiments 1 and 2 found that participants behaved differently when the standard method of verbally-presented summary statistics was used to convey environmental probabilities. Experiment 3 found that participants' preference for probability gain is robust, suggesting that other models contribute little to participants' search behavior.

Many situations require careful selection of information. Appropriate medical tests can improve diagnosis and treatment. Carefully designed experiments can facilitate choosing between competing scientific theories. Visual perception also requires careful selection of eye movements to informative parts of a visual scene. Intuitively, useful experiments are those for which plausible competing theories make the most contradictory predictions. A Bayesian optimal experimental design (OED) framework provides a mathematical scheme for calculating which query (experiment, medical test, or eye movement) is expected to be most useful. Mathematically, it is a special case of Bayesian decision theory (Savage, 1954). Note that a single theory is not tested in this framework, but rather multiple theories. The usefulness of an experiment is a function of the probabilities of the hypotheses under consideration, the explicit (and perhaps probabilistic) predictions that those hypotheses entail, and which utility function is being used.

In situations where different queries cost different amounts, and different kinds of mistakes have different costs, those constraints should be used to determine the best queries to make, rather than general purpose criteria for the value of information. This article, however, deals with situations where information gathering is the only goal. Specifically, we focus on situations in which the goal is to categorize an object by selecting useful features to view. Querying a feature, to obtain information about the probability of a stimulus belonging to a particular category, corresponds to an "experiment" in the OED framework, and will generally change one's belief about the probability the stimulus belongs to each of several categories. For instance, in environments where a higher proportion of men than women have beards, learning that a particular individual has a beard increases the probability that they are male. The various OED models differ in terms of how they calculate the usefulness of looking at particular features. All of the models

use Bayes's theorem to update beliefs about the probability of each category c_i when a particular feature value f is observed:

$$P(c_i \mid f) = \frac{P(f \mid c_i)\ P(c_i)}{P(f)} \tag{1}$$

where

$$P(f) = \sum_i P(f \mid c_i)\ P(c_i) \tag{2}$$

For updating to be possible, the probability distribution of the features and categories must be known. A practical difficulty is conveying a particular set of environmental probabilities to participants, an issue we address subsequently.

Several researchers have offered specific OED models (utility functions) for quantifying experiments' usefulness in probabilistic environments (e.g. Good, 1950; Fedorov, 1972; Lindley, 1956). Some prominent OED models from the literature are described below. They disagree with each other in important cases about which potential experiment is expected to be most useful (Nelson, 2005, 2008).

OED models of the usefulness of experiments

We use F (a random variable) to represent the experiment of looking at feature F, before its specific form f_j is known. Each OED model quantifies F's expected usefulness as the average of the usefulness of the possible f_j, weighted according to their probability:

$$E_{P(f)}[u(f)] = \sum_j P(f_j)\ u(f_j) \tag{3}$$

where $u(f_j)$ is the usefulness (utility) of observing f_j, according to a particular utility function. Each OED model's calculation of the usefulness of observing a feature value f_j, $u(f_j)$, is given below.

Probability gain (error minimization: Baron, 1981/1985) defines a datum's usefulness as the extent to which it increases the probability of correctly guessing the category of a randomly selected item:

$$u_{PG}(f) = \max_i(P(c_i \mid f)) - \max_i(P(c_i)) \tag{4}$$

Probability gain is by definition optimal where correct decisions are equally rewarded, and incorrect decisions are equally penalized.

Information gain (Lindley, 1956) defines a datum's usefulness as the extent to which it reduces uncertainty (Shannon entropy) about the probabilities of the individual categories c_i:

$$u_{IG}(f) = \sum_i P(c_i) \log \frac{1}{P(c_i)} - \sum_i P(c_i \mid f) \log \frac{1}{P(c_i \mid f)} \tag{5}$$

KL distance defines a datum's usefulness as the extent to which it changes beliefs about the possible hypotheses c_i, where belief change is measured with Kullback-Liebler (Kullback & Liebler, 1951) distance:

$$u_{KL}(f) = \sum_i P(c_i \mid f) \log \frac{P(c_i \mid f)}{P(c_i)} \tag{6}$$

Expected KL distance and expected information gain are always identical (Oaksford & Chater, 1996), e.g. $E_{P(f)}[\, u_{KL}(f) \,] = E_{P(f)}\, [\, u_{IG}(f) \,]$, making those measures equivalent for present purposes.

Impact (Klayman & Ha, 1987, pp. 219–220; Nelson, 2005, 2008; Wells & Lindsay, 1980) defines a datum's usefulness as the sum absolute change from prior to posterior beliefs (perhaps multiplied by a positive constant), over all hypotheses:

$$u_{Imp}(f) = \sum_i \left| (P(c_i) - P(c_i \mid f)) \right| \tag{7}$$

Impact and probability gain are equivalent if prior probabilities of the categories are equal. These utility functions can be viewed as candidate descriptive models of attention for categorization.

Bayesian diagnosticity (Good, 1950) and log diagnosticity, two additional measures, appear to contradict participants' behavior (Nelson, 2005), so we do not consider them here.[1]

Statistical models and human information acquisition

Which, if any, of the OED models describe human behavior? Wason's research in the 1960s, and several subsequent articles, suggest that there are biases in human information acquisition (Baron, Beattie, & Hershey, 1988; Klayman, 1995; Nickerson, 1998; Wason, 1960, 1966, Wason & Johnson-Laird, 1972; but see Peterson & Beach, 1967, pp. 37-38). Since about 1980, however, several authors have suggested that OED principles provide a good account of human information acquisition (McKenzie, 2004; Nelson, 2005, 2008; Trope & Bassok, 1982), even on Wason's original tasks (Ginzburg & Sejnowski, 1996; McKenzie, 2004; Nelson, Tenenbaum & Movellan, 2001; Oaksford & Chater, 1994). OED principles have been used to design experiments on human memory (Cavagnaro, Myung, Pitt & Kujala, in press), to explain eye movements as perceptual experiments (Butko & Movellan, 2008; Rehder & Hoffman, 2005; Nelson & Cottrell, 2007), to control eye movements in oculomotor robots (Denzler & Brown, 2002), and to predict individual neurons' responses (Nakamura, 2006).

In some cases, claims that human information acquisition is suboptimal because it follows ostensibly suboptimal heuristic strategies are questionable, because the heuristic strategies themselves correspond to OED models. Consider the feature difference heuristic (Slowiaczek, Klayman, Sherman & Skov, 1992). This heuristic, which applies in categorization tasks with two categories (c_1 and c_2) and two-valued features, entails looking at the feature for which abs($P(f_1 \mid c_1) - P(f_1 \mid c_2)$) is maximized. This heuristic exactly implements impact, an OED

model, irrespective of the prior probabilities of c_1 and c_2, and irrespective of the specific feature likelihoods (proof in Nelson, 2005, footnote 2; Nelson, 2009). This heuristic, therefore, is not suboptimal at all. In another case, Baron et al. (1988) found that participants exhibited information bias—valuing queries that change beliefs but do not improve probability of correct guess— on a medical diagnosis information-acquisition task. Yet information gain and impact, alternate OED models, also exhibit that bias (Nelson, 2005), suggesting that the choice of model may be central to whether or not a bias is found.

Which OED model best describes people's choices about which questions to ask? Nelson (2005) found that existing experimental data in the literature were unable to distinguish between the candidate models. Nelson's new experimental results strongly contradicted Bayesian diagnosticity and log diagnosticity, but were unable to differentiate between other OED models as descriptions of human behavior.

Here we address whether information gain/KL distance, impact, or probability gain best explains participants' evidence-acquisition behavior. We also test the possibility that participants may use a non-OED heuristic strategy of maximizing the probability of learning the true hypothesis (or category) with certainty (Baron et al., 1988). Mathematically, this model states that a datum (e.g. a specific observed feature value, or other experiment outcome) has utility 1 if it reveals the true category with certainty, and utility 0 otherwise.

We use computer search techniques to find statistical environments in which two models maximally disagree about which of two features is most useful for categorization, and then test those environments with human participants. A major limitation of most previous work in this area is that the participants have been told probabilities verbally. Yet verbal description and experience-based learning result in different behavior on several psychological tasks (Hertwig,

Barron, Weber, & Erev, 2004; McKenzie, 2006). We therefore designed an experiment using experience-based learning, with natural sampling (random according to environmental probabilities) and immediate feedback to convey the underlying probabilities. We also use a within-subjects manipulation to compare how experience in the statistical environment versus verbal statistics-based transmission of the probabilities influences information acquisition. The results are dramatically different.

Experiment 1: Pitting OED theories against one another using experience-based learning

This experiment involved classifying the species of simulated plankton (copepod) specimens as species *a* or *b* (here *a* and *b* play the role of c_1 and c_2) where the species was a probabilistic function of two two-valued features, *F* and *G*. Participants first learned environmental probabilities in a learning phase, where both features were visible, and then completed an information-acquisition phase, in which only one of the features could be selected and viewed on each trial.

In the learning phase, participants learned the underlying environmental probabilities by classifying the species of each plankton specimen, with immediate feedback. On each trial, a stimulus was chosen randomly according to the probabilities governing categories and features. One form of each feature was always present. The learning phase continued until a subject mastered the underlying probabilities. Figure 1 gives illustrative examples of the plankton stimuli and the probabilistic nature of the categorization task.

In the subsequent information-acquisition phase participants continued to classify the plankton specimens. However, the features were obscured, and only one feature (selected by the participant) could be viewed on each trial. The feature likelihoods in each condition were designed so that two competing theories of the value of information strongly disagreed about

which of the two features was more useful. In this way, participants' choice of which feature to view also provided information about which theoretical model best describes their intuitions about the usefulness of information. We pitted the four different OED models and the heuristic against each other in four conditions, as shown in Table 1.

Finally, each participant completed a verbal summary statistic-based questionnaire on the usefulness of several features in an alien categorization task. The questionnaire employed the same probabilities that the participant had just learned experientially on the plankton task. This enabled within-subjects comparison of how the different means of conveying environmental probabilities affect information-acquisition behavior.

Participants

Participants were 129 students in social science classes at UCSD, who received partial or extra course credit for participation. Participants were run in small groups of up to 5 people, over 1.5 to 2 hours. Participants were assigned at random to one of the four conditions in Table 1, while keeping approximately equal numbers who reached criterion learning performance in each condition.

Optimizing experimental probabilities

We used computational search techniques to identify the feature likelihoods in each condition such that a pair of theories maximally disagreed about which feature (*F* or *G*) was more useful (see supplementary material). This automatic procedure found scenarios with strong (and often non-obvious) cases of disagreement between theories. Note that a prior probability distribution is specified by five numbers: the prior probability of category *a*, $P(a)$, and four feature likelihoods, $P(f_1|a)$, $P(f_1|b)$, $P(g_1|a)$, and $P(g_1|b)$. We set $P(a)$ to 70%, as suggested by Nelson's (2005) optimizations. The program first finds a case at random in which the two models disagree,

and then modifies the four feature likelihoods to make that disagreement as large as possible (Figure 2). Table 1 gives the feature likelihoods for the conditions of Experiment 1 obtained by the optimizations, for each pair of models compared.

We defined the preference strength of a model m for feature F, $PStr_m$, as the difference between the two features' expected usefulness, e.g. $eu_m(F) - eu_m(G)$, where each term is defined by Equation 1, scaled by the maximum possible difference in features' usefulness according to model m, $maxPStr_m$, multiplied by 100:

$$PStr_m = 100 * (eu_m(F) - eu_m(G)) / maxPStr_m \quad (8)$$

The (typically unique) maximum possible preference strength, for all the OED models and the probability of certainty heuristic, is obtained where the categories are equally probable a priori, one feature is definitive, and the other feature is useless, e.g. where $P(a) = P(b) = 0.50$, $P(f_1|a) = 0$, $P(f_1|b) = 1$, and $P(g_1|a) = P(g_1|b)$.

We then defined the pairwise disagreement strength (*DStr*) as the geometric mean of the opposed models' respective absolute preference strengths ($PStr_{m1}$ and $PStr_{m2}$), when model 1 and model 2 disagree:

$$DStr_{m1\ vs.\ m2} = (|PStr_{m1}| * |PStr_{m2}|)^{0.5}, \text{ if } PStr_{m1} * PStr_{m2} \leq 0 \quad (9)$$

If, however, the models agree about which feature is most useful, *DStr* is zero:

$$DStr_{m1\ vs.\ m2} = 0, \text{ if } PStr_{m1} * PStr_{m2} \geq 0. \quad (10)$$

An example calculation is provided in the supplementary material.

Design and procedure of behavioral experiment

Software was programmed to conduct the experiment. Participants were familiarized with the features in advance, to ensure that they perceived the two variants of each feature (Figure S3).

The physical features (eye, claw, and tail) were adjusted during pilot research to minimize any salience differences. Each participant was randomly assigned to one of 96 possible randomizations of each condition to guard against any residual bias among the physical features, the two variants of each feature, or the species names.

The learning phase of the experiment was similar to probabilistic category learning experiments (Knowlton, Squire & Gluck, 1994; Kruschke & Johansen, 1999). In each trial a plankton stimulus was randomly sampled from the environmental probabilities, and presented to the participant: the category was chosen according to the prior probabilities $P(a)$ and $P(b)$, and the features were generated according to the feature likelihoods, $P(f_1|a)$, $P(f_1|b)$, $P(g_1|a)$, and $P(g_1|b)$. There were no symmetries or other class-conditional feature dependencies. The participant classified the specimen as species *a* or *b* and was given immediate feedback on whether the classification was correct according to which category had been generated. Note that the optimal decision (corresponding to the category with highest posterior probability, given the observed features) was frequently given negative feedback, because certain combinations of features were observed in both species (cf. Figure 1). Participants were also given the running percent of trials in which their classifications were correct.

Subjects vary by more than a factor of 10 in the number of trials needed to learn. The learning phase continued until criterion performance was reached, or the available time (~2h) elapsed. Criterion performance was defined as either (1) making at least 99% optimal (not necessarily correct) responses in the last 200 trials, *irrespective of the specific stimuli* in those trials; or (2) making at least 95% optimal responses in the last 20 trials *of every single stimulus type*. The goal was to ensure that participants achieved high mastery of the environmental probabilities before beginning the information-acquisition test phase.[2]

Experience-based learning results

A median of 933, 734, 1082, and 690 trials on the experience-based learning phase were required to achieve criterion performance in Conditions 1-4, respectively. 113 of 129 participants achieved criterion performance, and were given the information-acquisition task.

The most striking information-acquisition result is that in all conditions, irrespective of what theoretical models were being compared, the feature with higher probability gain was preferred by a majority of participants (Figure 3). Moreover, the preference to view the higher-probability gain feature is quite strong. Aggregating all conditions, the median participant viewed the higher-probability gain feature 99% of the time (in 100 of 101 trials).[3] Between 82% and 97% of participants preferentially viewed the higher probability gain feature (*F*) in each condition (Table 2; all p's < 0.001). In Conditions 1 and 2, all models except probability gain preferred *G*, making participants' preference for *F* especially striking. In Condition 3, 27 of 28 participants preferred *F*, which optimized information gain, probability gain, and probability of certainty, rather than impact. In Condition 4, 28 of 29 participants preferred to optimize the OED models, including probability gain, rather than the probability of certainty heuristic.

Summary statistics-based task

After completion of the experience-based learning and information-acquisition phases of the probabilistic plankton categorization task, participants were given an equivalent task, in which environmental probabilities (prior probabilities and feature likelihoods) were presented verbally via summary statistics. (Gigerenzer & Hoffrage, 1995, called this the *standard probability format*.) This task used the Planet Vuma scenario (Skov & Sherman, 1986), in which the goal is to classify the species of invisible aliens ("glom" or "fizo") by asking about features that the different species have in varying proportion (such as wearing a hula hoop, or gurgling a lot). The prior

probability of each species, e.g. $P(glom) = 70\%$, and the likelihoods of each feature, e.g. $P(hula \mid glom) = 0\%$, and $P(hula \mid fizo) = 29\%$, exactly matched the plankton task the participant had just completed (though this was not disclosed to subjects). An uninformative third feature (present in 100% of both species or in 0% of both species) was also included to ensure that participants read and understood the given information. Participants were asked to rate, from most to least useful, which of the features would be most helpful to enable them to categorize an alien as a glom or fizo.

Summary statistics-based results; comparison with experience-based learning

Statistics-based results were much less clear than experience-based results. Interestingly, the trend in every condition was for the feature with higher *information gain* (not probability gain) to be preferred. Were participants consistent? We performed chi-square tests in each condition across the two tasks, to assess whether individual preferences in experience-based learning predict preferences in summary statistics-based learning. All tests were nonsignificant, providing no evidence for within-subject consistency, inconsistency, or any relationship whatsoever between the modalities. This suggests that data from summary statistics-based information-acquisition experiments in the literature may fail to predict naturalistic information-acquisition tasks (e.g. eye movements in natural scenes) where people have personal experience with environmental probabilities.

Experiment 2: summary statistics-based information acquisition

Confidence intervals for participants' preferences between features from the statistics-based task (in which participants gave a rank order only) were much broader than those from the comparatively data-rich experience-based task (in which there were 101 information-acquisition trials). We therefore obtained statistics-based-task data from 106 additional UCSD students.

Participants were randomly assigned to one of the same four conditions as in Experiment 1. Participants were assigned at random to either an alien or plankton categorization scenario. Each participant was randomly assigned to one of 96 possible randomizations of their condition's probabilities. Results in both scenarios were consistent with Experiment 1's statistics-based results. We therefore aggregate all statistics-based results below.

Experience- vs. statistics-based learning results

Table 2 compares experience- and statistics-based information-acquisition results. The proportion of participants preferring *F* was different, in every condition, between the types of learning. Experience-based learning led to preferring the feature with higher probability gain in every condition. Statistics-based learning led to a modal preference to maximize information gain in each condition. However, statistics-based results are indistinguishable from chance in Conditions 3 and 4, and less clear in all conditions.

Experiment 3: How robust is the preference for probability gain?

In Experiment 3, we explore possible limits in the circumstances where participants maximize probability gain, as outlined below.

Experiment 3, Condition 1. Would information gain or the possibility of a certain result "break the tie" when probability gain is indifferent? We tested this in one scenario where both *F* and *G* have probability gain 0.25, yet *F* has higher information gain and is the only feature to offer the possibility of a certain result: $P(a) = 0.50$, $P(f_1|a) = 0$, $P(f_1|b) = 0.50$, $P(g_1|a) = 0.25$, and $P(g_1|b) = 0.75$. Surprisingly, only about half of subjects (12/22= 55%) preferred F, even though its greater information and the possibility of certainty had zero cost in terms of probability gain.

Experiment 3, Condition 2. Another approach is to modify Experiment 1, Conditions 1 and

2, so that probability gain has a relatively marginal preference for F, while the other models have increased preference for G. We tested one such scenario, where $P(a) = 0.70$, $P(f_1|a) = 0$, $P(f_1|b) = 0.15$, $P(g_1|a) = 0.57$, and $P(g_1|b) = 0$. Here, probability gain marginally prefers *F* (*PStr*= 9), whereas the other models more strongly prefer *G* (*PStr* for information gain, impact, and probability of certainty, respectively, are: -20, -35, and -35). Eight of nine learners maximized probability gain.

Experiment 3, Condition 3. Yet another approach is to modify Experiment 1, Conditions 1 and 2, so that the *F* feature, taken alone, can never give a certain result. We evaluated this where $P(a)$=0.70, $P(f_1|a)$=0.04, $P(f_1|b)$=0.37, $P(g_1|a)$=0.57, and $P(g_1|b)$=0. *F* has higher probability gain. Yet *G* is the only feature to offer the possibility of a certain result, and has higher information gain and impact. Here, 6 of 20 subjects (30%) preferred *G*. This environment is the first we have identified in which a nontrivial minority of subjects optimize something besides probability gain.

Taken together, our data strongly point to probability gain (or a substantially similar model) as the primary basis for the subjective value of information in categorization tasks.

General Discussion

This article reports the first information-acquisition experiment in which both:

(1) environmental probabilities were designed to maximally differentiate theoretical predictions of competing models, and

(2) experience-based learning was used to convey environmental probabilities.

Previous studies did not distinguish between several models of information-acquisition behavior. Yet we obtained very clear results pointing to probability gain as the primary basis for

the subjective value of information for categorization. Our within-subjects comparison of traditional summary statistics-based presentation of environmental probabilities with experience-based learning is another contribution. The convincing lack of relationship between the two types of tasks is remarkable and should be explored further. For instance, the visual system may code statistics and contingencies more effectively than linguistic parts of the brain. As a practical matter, experience-based learning might be speeded by simultaneous presentation of multiple examples (Corter & Matsuka, 2007). Verbal-based information search might be facilitated by natural frequency formats or explicit instruction in Bayesian reasoning (Gigerenzer & Hoffrage, 1995; Krauss, Martignon, & Hoffrage, 1999; Sedlmeier & Gigerenzer, 2001).

Treating evidence acquisition as an experimental design problem broadens the "statistical man" approach, which originally focused on inferences people make given preselected data (Peterson & Beach, 1967). Key current questions include:

-- Does information acquisition in medical diagnosis, scientific hypothesis testing, and word learning optimize probability gain?

-- Does the visual system optimize probability gain when directing the eyes' gaze?

-- Can people optimize criteria besides probability gain when necessary?

Theories of the statistical human should aim to address these issues in a unified account of cognitive and perceptual learning and information acquisition.

Sources cited

Baron, J. (1985). *Rationality and Intelligence*. Cambridge, England: Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, 42, 88–110.

Butko, N. J. & Movellan, J. R. (2008). I-POMDP: An infomax model of eye movement. *Proceedings of the 7th IEEE International Conference on Development and Learning*, 139-144. doi:10.1109/DEVLRN.2008.4640819

Cavagnaro, D. R., Myung, J. I., Pitt, M. A. & Kujala, J. (in press). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*.

Denzler, J., & Brown, C. M. (2002). Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24, 145-157.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats, *Psychological Review*, l02(4), 684-704.

Ginzburg, I., & Sejnowski, T. J. (1996). Dynamics of rule induction by making queries: Transition between strategies. In G. W. Cottrell (Ed.), *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 121–125). Mahwah, NJ: Erlbaum.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. New York: Griffin.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8), 534-539.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, 42, 385–418.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning and Memory*, 1, 106-120.

Krauss, S., Martignon, L. & Hoffrage, U. (1999) Simplifying Bayesian inference: the general case. In L. Magnani, N. Nersessian & P. Thagard (Eds), *Model-Based Reasoning in Scientific Discovery*. New York: Kluwer Academic/Plenum Publishers

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(5), 1083-1119.

Kullback, S., & Liebler, R. A. (1951). Information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

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

Matsuka, T. & Corter, J. E. (2007). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology,* 61(7), 1067 – 1097.

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200-219). Oxford: Blackwell.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition*, 34(3), 577-588.

Nakamura, K (2006). Neural representation of information measure in the primate premotor cortex. *Journal of Neurophysiology*, 96, 478-485.

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

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In Oaksford, M. & Chater, N. (Eds.), *The Probabilistic Mind: Prospects for Rational Models of Cognition*., pp. 143-163. Oxford: Oxford University Press.

Nelson, J. D. (2009). Naïve optimality: Subjects' heuristics can be better-motivated than experimenters' optimal models. *Behavioral and Brain Sciences, 32,* 94-95.

Nelson, J. D. & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, 70, 2256-2272.

Nelson, J. D., Tenenbaum, J. B., & Movellan, J. R. (2001). Active inference in concept learning. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Conference of the Cognitive Science Society*, 692-697. Mahwah, NJ: Erlbaum.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175-220.

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

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, 103, 381–391.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29-46.

Rehder, B. & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1-41.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3), 380–400.

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.

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.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, 43, 22–34.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129–140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Harmondsworth, England: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

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

Author note

NIH T32 MH020002-04 (TJS, P.I.), NIH MH57075 (GWC, P.I.), NSF SBE-0542013 (Temporal Dynamics of Learning Center, GWC, P.I.), and NSF SES-0551225 (CRMM, P.I.) supported this research. We thank Björn Meder, Gudny Gudmundsdottir, and Javier Movellan for helpful ideas; Gregor Caregnato, Tiana Zhang and Stephanie Buck for help with experiments; Paula Parpart for translation help; and the participants who conscientiously completed the experiments. We thank Jorge Rey and Sheila O'Connell (University of Florida, FMELe) for allowing us to base our artificial plankton stimuli on their copepod photographs. Additional data and analyses are available from JDN and in the Supplementary Material.

Footnotes

1. The diagnosticity measures are also flawed as theoretical models (Nelson 2005, 2008). For instance, they prefer a query that offers 1 in 10^{100} probability of a certain result, but is otherwise useless, to a query that will always provide 99% certainty.

2. In some conditions, one could in principle learn only the *F* feature and trigger the performance criterion. However, error data during learning (Figures S1 and S2), debriefing of subjects following the experiment, explicit tests of knowledge in a replication of Experiment 1, Condition 1, and subsequent experiments show that participants learned configurations of features.

3. We separately tested the extent to which participants will view an individual feature, if two features are statistically identical, where $P(a) = P(b) = 0.50$, $P(f_1 \mid a) = 0$, $P(f_1 \mid b) = 0.5$, $P(g_1 \mid a) = 0$, and $P(g_1 \mid b) = 0.5$. For each participant, the median percent of views to their more-frequently-viewed feature was 64%. This suggests that if the vast majority of participants view a particular feature in the vast majority of trials, that behavior should be taken to reflect a real preference between features, and not simply habit or perseveration.

Tables

Table 1

Feature likelihoods to best differentiate competing theoretical models of the value of information

Cond.	$P(f_1 \mid a)$	$P(f_1 \mid b)$	$P(g_1 \mid a)$	$P(g_1 \mid b)$	*DStr*	Model preferring *F*	$PStr_{m1}$	$eu_{m1}(F)$	$eu_{m1}(G)$	Model preferring *G*	$PStr_{m2}$	$eu_{m2}(F)$	$eu_{m2}(G)$
1	0	0.24	0.57	0	14.5	Probability gain	14.4	0.072	0.000	Information gain (impact, prob. certainty)	-14.5	0.135	0.280
2	0	0.29	0.57	0	20.2	Probability gain	17.4	0.087	0.000	Impact (information gain, prob. certainty)	-23.5	0.122	0.239
3	0	0.40	0.73	0.22	8.2	Information gain (probability gain, prob. certainty)	7.2	0.238	0.166	Impact	-9.2	0.168	0.214
4	0.05	0.95	0.57	0	37.9	Probability gain, information gain, impact	36.0	[#]	[#]	Prob. certainty	39.9	0.000	0.399

Note. In all conditions, we set $P(a) = 0.70$, and $P(b) = 0.30$. *F* denotes the feature with higher probability gain. Disagreement strength (*DStr*) scales between 0 (none) to 100 (maximal). $PStr_{m1}$ denotes Model 1's preference strength for *F*, versus *G*. $PStr_{m2}$ denotes Model 2's preference strength between *F* and *G*; this is negative because Model 2 prefers *G*. $eu_{m1}(F)$ is the expected utility of *F*, according to Model 1. Models in parentheses were not optimized in the condition per se, but also prefer the feature in their respective column.

[#]In Condition 4, $PStr_{m1}$ is based on the geometric mean of the individual Preference Strengths of probability gain (50), information gain (34), and impact (28).

Table 2

Information-acquisition results, Experiments 1 and 2

Condition	Proportion of participants preferring higher-probability gain feature (*F*):			Views to *F* in experience-based task:	
	Experience-based	Statistics-based	Experience = Statistics?	Median participant	Mean, over all participants
1	82 % ***	27 % **	no ****	97 %	77 %
2	82 % ***	30 % *	no ****	97 %	75 %
3	96 % ****	65 % ^{ns}	no **	99 %	89 %
4	97 % ****	58 % ^{ns}	no ***	100 %	94 %

Note. *F* denotes the feature with higher probability gain. In all conditions, $P(a) = 0.70$, and $P(b) = 0.30$. Table 1 gives the feature likelihoods in each condition. Two-tail binomial tests were used to test whether the number of participants favoring *F* was different from chance in each condition. Two-tail difference of proportions tests were used to test whether equivalent proportions of participants preferred *F* in the experience-based and summary statistics-based tasks. P-values are reported as follows: $p < 0.0001$, ****; $p < 0.001$, ***; $p < 0.01$, **; $p < 0.05$, *; $p \geq 0.05$, ^{ns}. There were 28-29 experience-based, and 43-45 statistics-based, participants in each condition.

Figures and Captions

Figure 1.

Illustrative plankton specimens (see supplemental figures S3-S5 for actual stimuli – these have been altered to make the differences clearer in print). The plankton in the left half are species *a*, and those on the right are species *b*. Note that only the eye (which can be yellow or black) and claw (which can be dark or light green) vary between the specimens. Because of the probabilistic distribution of the features within each species, most specimens cannot be identified as species *a* or species *b* with certainty. (The same images, e.g., with black eye and light green claw, occur in each category.) In this case (assuming the observed specimens match underlying probabilities), P(species *a* | yellow eye) = 1, P(species *b* | black eye) = 13/16, P(species *a* | light green claw) = 7/8, and P(species *b* | dark green claw) = 7/8. Information gain, impact, and probability gain agree that the claw is more useful than the eye, but only the eye offers the possibility of certainty.

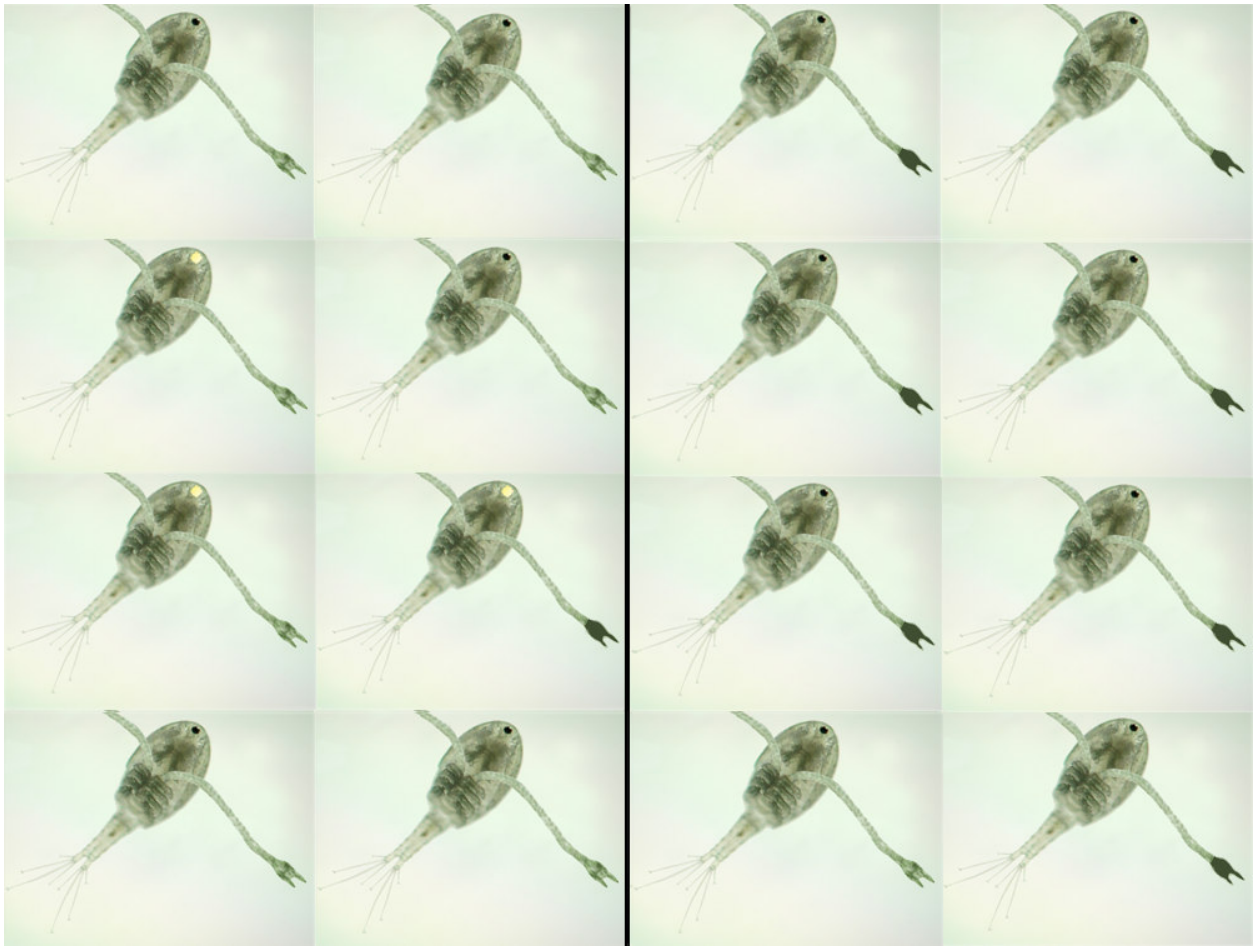


Figure 2.

Finding maximally informative features (*F* and *G*) to differentiate the predictions of competing theoretical models of the value of information (Model 1 and Model 2). The goal of each optimization is to maximize Disagreement Strength (*DStr*)—which is based on the geometric mean of the two models' absolute preference strengths—between the models.

--A: Model 1 considers *F* to be slightly more useful than *G*, and Model 2 considers *G* to be slightly more useful than *F*. The shallow slopes of the connecting lines illustrate that the models' (contradictory) preferences are weak, and *DStr* is low. Generating feature likelihoods at random, the first step in the optimizations to maximally differentiate competing theoretical models of the value of information, typically only finds weak disagreement.

--B: the ideal scenario for experimental test, where *DStr* is maximal. Model 1 holds that *F* is much more useful than *G*; Model 2 has opposite and equally strong preferences.

--C: Model 2 strongly prefers *G* to *F*, and Model 1 marginally prefers *F* to *G*. This is not an ideal case to test experimentally. Because Model 1 is close to indifferent, *DStr* is low even though Model 2 has a strong preference.

--D: *DStr* is higher in this scenario than in Panel C, because the models both have moderate (and contradictory) preferences.

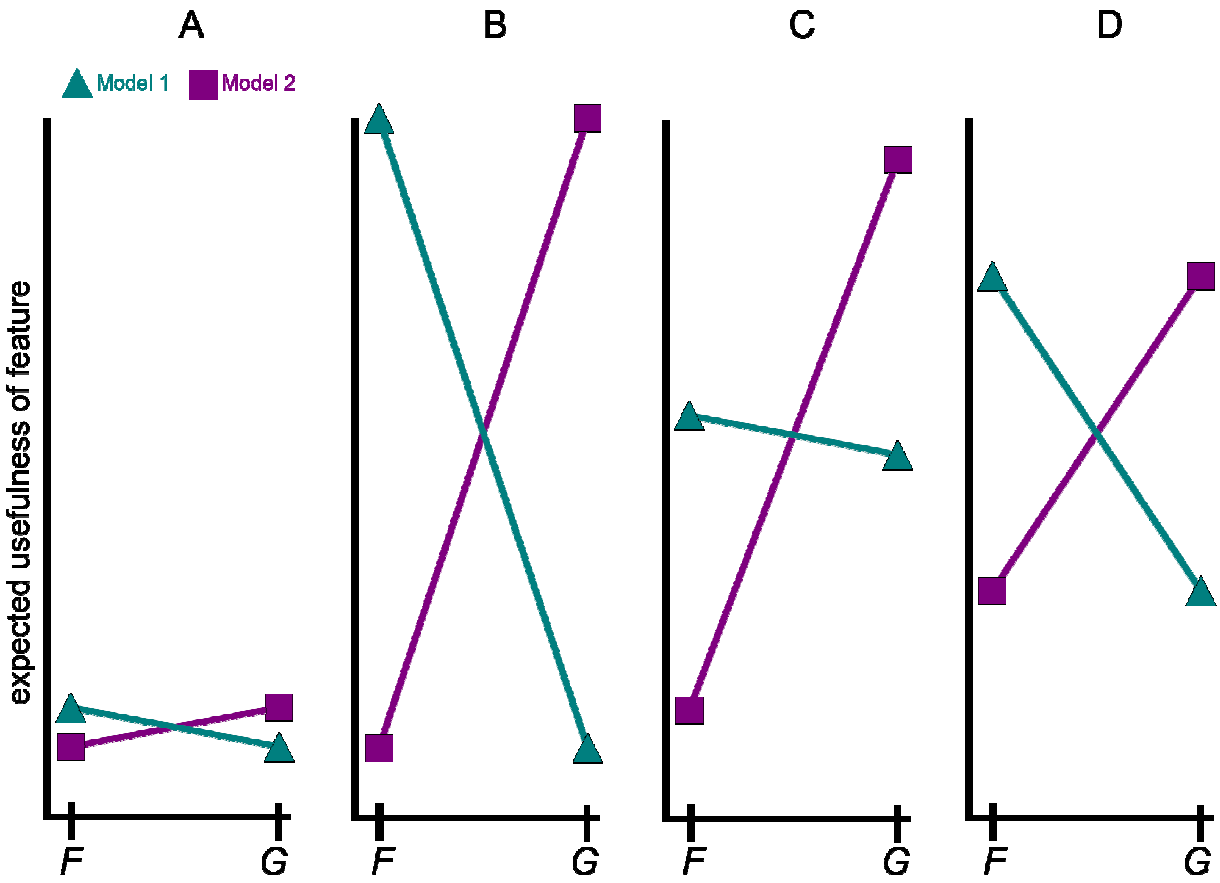
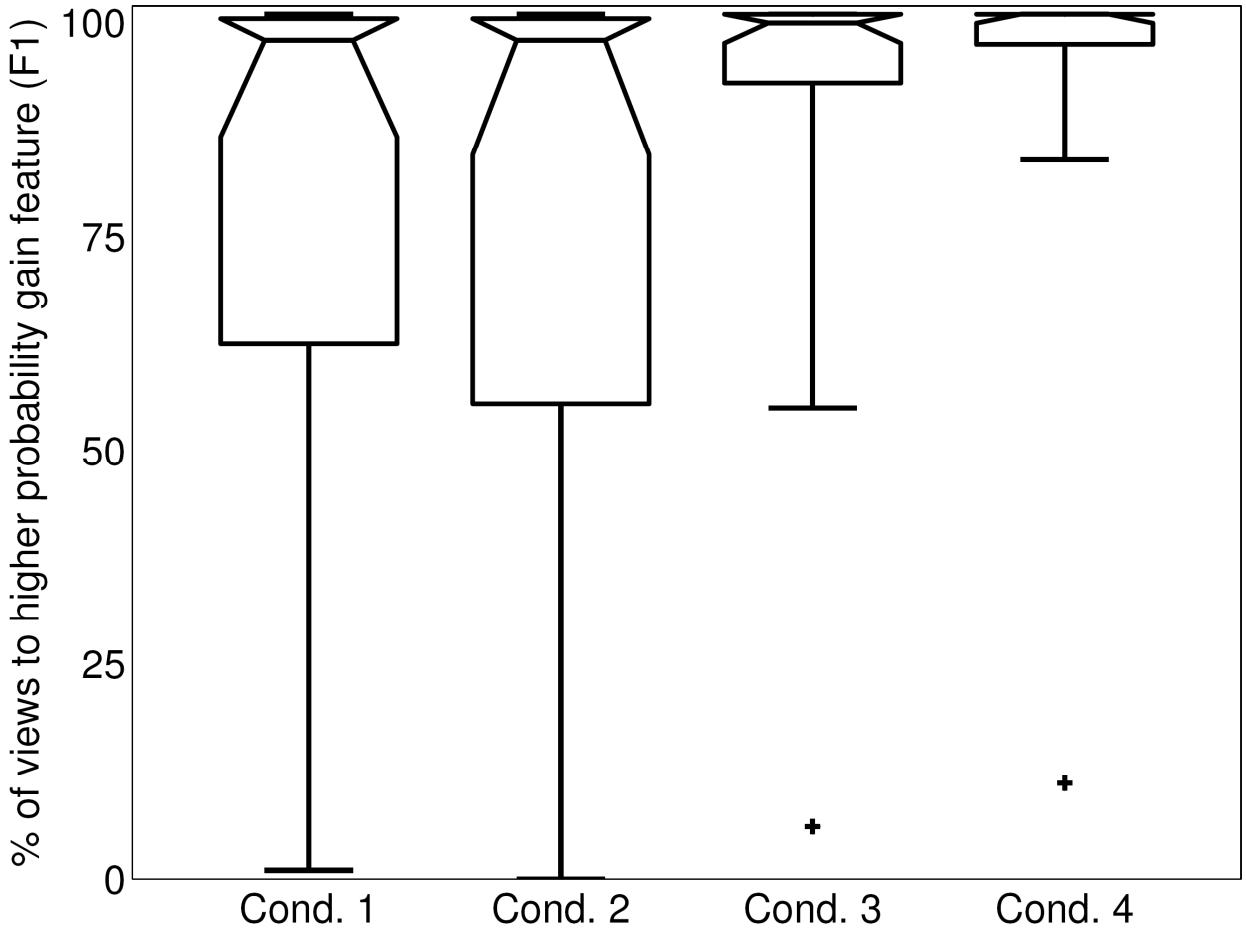

A
B
C
D
Model 1
Model 2
expected usefulness of feature
F
G
F
G
F
G
F
G

Figure 3.

Experience-based-learning participants almost exclusively viewed the higher-probability gain feature (*F*). The median participant viewed *F* 97%, 97%, 99%, and 100% of the time in Conditions 1 through 4, respectively. (Chance = 50% in each condition. All p's < 0.001; see Table 2.) The boxes give the interquartile range, with notches denoting the median. The outermost bars depict the extent of the data, with the exception of outliers (+) which are more than 10 times beyond the interquartile range.



Supplemental material

Optimization notes

An example illustrates calculation of *DStr*, for Condition 1:

$$PStr_{PG} = 100 * (eu_{PG}(F) - eu_{PG}(G)) / maxPStr_{PG}$$
$$= 100 * (0.072 - 0) / 0.50 = 14.4$$

$$PStr_{IG} = 100 * (eu_{IG}(F) - eu_{IG}(G)) / maxPStr_{IG}$$
$$= 100 * (0.134 \text{ bits} - 0.280 \text{ bits}) / 1 \text{ bit} = -14.6$$

$$DStr = (| 14.4 | * | -14.6 |)^{0.5} = 14.5, \text{ because } PStr_{IG} * PStr_{PG} < 0.$$

In each optimization, obtained feature likelihoods were rounded to the nearest 0.01 for use in the experiments. In Condition 1 (information gain versus probability gain), the original optimizations produced values such as $P(f_1|a) = 0.04$, $P(f_1|b) = 0.38$, $P(g_1|a) = 0.57$, and $P(g_1|b) = 0$. These values confounded the possibility of knowing for sure with the desired comparison of information gain and probability gain. (Whereas our desired test was between information gain and probability gain, only *G* offered the possibility of a certain result. If participants wished to maximize probability of a certain result, and hence preferred *G*, this could have been misinterpreted as a preference to optimize information gain.) We therefore repeated the optimization, requiring $P(f_1|a) = 0$, just as $P(g_1|b) = 0$. This removed that confound while having negligible effect on strength of disagreement. The same confound appeared in Condition 2, and was also remedied by requiring $P(f_1|a) = 0$. In Experiment 3 an environment along these lines where $P(f_1|a) = 0.04$ was tested; results continue to favor probability gain.

Pairwise optimizations of each OED model vs. the probability of certainty heuristic resulted in virtually identical feature likelihoods. In Condition 4, we therefore optimized the disagreement strength of probability of certainty versus the joint preference of all three OED models. (We defined the joint preference of the OED models as the geometric mean of their individual preference strengths.) A further note is that this optimization produced features for which $P(f_1|a) = \varepsilon$, and $P(f_1|b) = 1- \varepsilon$, where $\varepsilon \approx 0.0001$. Unfortunately, the difference between $P(f_1|a) = 0$ and $P(f_1|a) = 0.0001$, though important for the probability of certainty model, is not learnable in two hours of experience-based training with natural sampling. We therefore redid this optimization, fixing *F* such that $P(f_1|a) = 0.05$, and $P(f_1|b) = 0.95$.

In the optimizations (see Table 1 in the article), a feature where $P(f_1|a) = 4/7 \approx 0.57$, and $P(f_1|b) = 0$, occurred frequently. This may be because, holding $P(a) = 0.70$ and $P(f|b) = 0$ constant, $P(f|a) = 4/7$ is the highest feature likelihood such that the feature has zero probability gain. In Condition 1 and Condition 2, *F* is rarely f_1 (7% or 9% of the time); but if $F=f_1$, the probability of species *b* changes from 30% to 100%. If $F = f_2$, the probability of species *a* increases (from 70% to 75% or 77%). If $G=g_1$, it is species *a* for sure. However, if $G=g_2$, it is a 50/50 chance whether the species is *a* or *b*. These possibilities cancel each other out, such that the overall probability of correct guess is not improved by querying *G*, despite *G*'s higher information gain and impact. In Condition 3, *F* is f_1 12% of the time; if $F=f_1$ uncertainty is eliminated; information gain prefers *F*. If $F=f_2$ the probability of species *a* goes from 70% to 80%, which also reduces uncertainty. Impact depends on the absolute difference in feature

likelihoods, which favors *G* (0.73 - 0.22 = 0.51) over *F* (0.40 – 0 = 0.40). In Condition 4, all the OED models, which were jointly optimized versus probability of certainty, prefer *F*, which leads to always knowing the true category with high probability, but never for sure. *G* leads to knowing the true category for sure 40% of the time, but to lower overall probability correct, to higher uncertainty, and to lesser absolute change in beliefs

Experiment notes

Between 6% and 22% of participants did not reach criterion performance in each condition of Experiment 1. Condition 1 had 13% nonlearners (4/32); Condition 2, 7% (2/30); Condition 3, 22% (8/36); and Condition 4, 6% (2/31). Condition 3 was difficult because one of its stimulus items, which occurred less than 1/3 of the time, led to only 57% posterior probability of the most-probable category, and thus took a great deal of experience to learn.

Did subjects learn both features *F* and *G*, as intended, or only marginal probabilities involving a single feature? In some conditions, it is theoretically possible to only learn *F*, and yet to achieve the performance criterion. We therefore analyzed the proportion of optimal responses for each configuration of features. (Optimal is choosing the more-probable species, irrespective of how close the posterior probability is to 50%, given a particular configuration. This is true irrespective of which utility a person wishes to optimize in the information-acquisition phase.) We present data for Experiment 1, Condition 1, below; this is representative of the conditions where it is theoretically possible to only learn the *F* feature.

If subjects only learned the *F* feature, then the green line ('certain-a config,' f2,g1) and the blue line ('uncertain-a config,' f2,g2) would be overlaid, except for random jitter, throughout learning, as these configurations differ only along the *F* feature. The results, however, show that subjects differentiated these configurations, quickly mastering the certain-a configuration, yet struggling with the uncertain-a configuration until very late (e.g. the last 4% of learning trials) in the learning process.

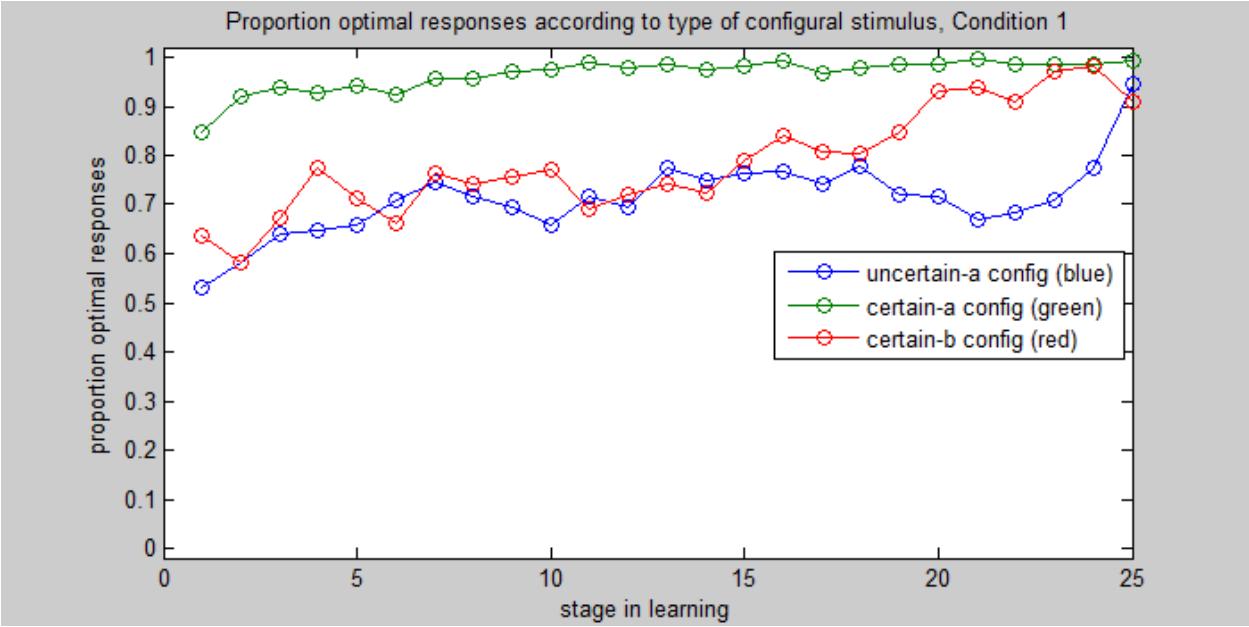


Figure S1. Aggregate learning data for Experiment 1, Condition 1.

The difference between the green line (top), for the certain-a configuration (f2,g1), and the blue line (bottom), for the uncertain-a configuration (f2,g2), demonstrate that subjects learned configurally. The red line depicts the certain-b (f1,g2) configuration.

Because different subjects learned in different numbers of trials, and because different configurations of stimuli occurred with different frequencies, the data below are normalized so that the first 1/25th (4%) of trials on a particular configuration is plotted first, the second 4% of trials on a particular configuration is plotted second, etc., for each subject. In this way, rare stimuli and frequent stimuli, and subjects who learned quickly and slowly, contribute equally to the proportion of optimal responses denoted at each point in learning. (Note that the figure requires color.)

What do individual subjects data show? Figure S2 shows every learning trial for each subject in Experiment 1, Condition 1. Each of the 28 rows represents a single subject.

Note the greatly higher rates of suboptimal responding to the uncertain-a configuration (left column), versus the certain-a configuration (middle column), which differ only according to the *G* feature. This demonstrates that individual subjects separately (configurally) learned each stimulus item, and did not only learn marginal probabilities associated with the *F* feature. Some subjects vacillate between periods of correct and incorrect responding on the uncertain-a configuration, further evidence that they perceive the difference between the configurations.

Could the subjects, once they learned probabilities involving both features and each configuration of features, have forgotten those configural probabilities late in learning, before the information-

acquisition phase?[1] It was possible to debrief the vast majority of subjects following the experiment; the vast majority of these subjects showed high familiarity with environmental probabilities, including the fact that various configurations (though both pointing to species a, for instance) had widely varying levels of certainty.

To more systematically evaluate this qualitative result, we subsequently obtained data from an additional 13 subjects in the Experiment 1, Condition 1, environment. (There was one additional nonlearner.) Eleven of thirteen subjects preferentially viewed the *F* feature, consistent with earlier information-acquisition results. This replication experiment included a new knowledge test page (following the information-acquisition phase) in which subjects were explicitly asked, for each kind of specimen that appeared, the percent of instances in which it had been species a and b. Subjects were also asked which percent of specimens, overall, were species a and b. Analysis of individual subjects' results (Table S1) shows that the vast majority of subjects were qualitatively very close in their beliefs, identifying the more probable species overall, the more probable species given each configuration of features, and the approximate certainty induced by each configuration of features. Thus, subjects preferred the *F* feature given their knowledge of configural environmental probabilities, not because it was the only feature that they learned.

Additional data, describing corresponding analyses of other conditions, are available from the first author. These data show configural learning throughout.

---

[1] Note that this concern is not a theoretical possibility in some conditions, in which responding optimally to all configurations unequivocally implies that a subject effectively differentiates the two features, and not just a single feature. This a theoretical possibility in Experiment 1, Conditions 1 and 2—though it is implausible: note from Fig. S1 that such forgetting would have to have occurred in the last 4% or so of learning trials.

Figure S2
(at right; notes below)
Data for learning phase from Experiment 1, Condition 1, from each of 28 individual subjects who obtained criterion performance.
Key: trials are ordered from top to bottom, and left to right, in each rectangle.

Each subject appears on one row; each configuration in one column. Optimal responses are depicted in white; suboptimal responses are depicted in black.

Left column:
uncertain-a (f2,g2; 56.9% are Species a);

Middle column:
certain-a (f2,g1; 100% are Species a);

Right column:
certain-b (f1,g2; 100% are Species b).

The f1,g1 configuration does not occur in this environment.

The higher suboptimal response rates for the uncertain-a configuration (left) than for the certain-a configuration (middle) show that subjects learned configurations of features, and not merely the higher probability gain feature. Suboptimal response rates are statistically greater for the uncertain-a configuration than the certain-a configuration in 26 of 28 subjects, by both difference-of-proportions and bootstrap tests.

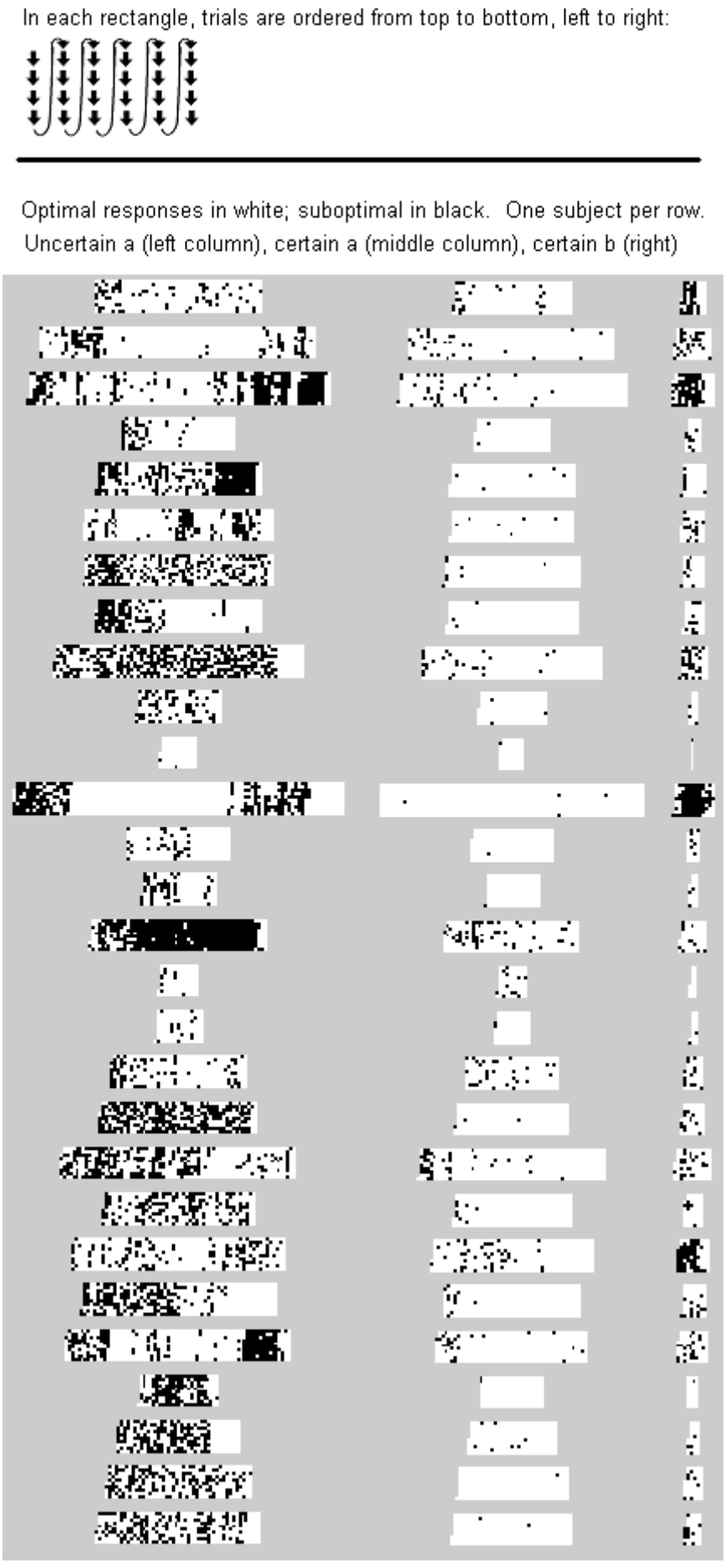


Table S1. Subjects show high calibration to the environmental probabilities.

Item	True percent	Median rating	Mean rating	Individual subjects' probability ratings: #1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
P(a|f2,g2)	57	65	67	50	55	99	80	65	54	55	75	75	67	77	55	65
P(a|f2,g1)	100	100	100	100	100	100	100	95	100	100	100	100	100	100	100	100
P(a|f1,g2)	0	0	9	0	0	0	0	20	0	0	0	0	0	95	0	0
P(a)	70	82	80	48	73	90	90	90	79	85	82	75	62	79	94	92

Note. The item being judged is in the left column; its true percent next; and the median and mean of subjects' estimated percentages next. Individual subjects (columns #1 to #13, at right) in most cases showed very good learning of environmental probabilities. Whether species a or b was more probable was randomized across subjects. In this table, 'a' denotes whichever species was more probable in a particular subject's randomization.

Plankton stimuli

The actual plankton stimuli appear below. Our plankton stimuli, though hopefully naturalistic in appearance, should not be confused with real copepods. (For instance, the claw feature did not occur in the original images.) The stimuli were designed to have three subtly-varying two-valued features (tail, eye, claw), roughly equidistant from each other. We thank Profs. Jorge Rey and Sheila O'Connell (University of Florida, Medical Etymology Laboratory), for allowing us to base our artificial plankton stimuli on their photographs of real copepod plankton specimens.

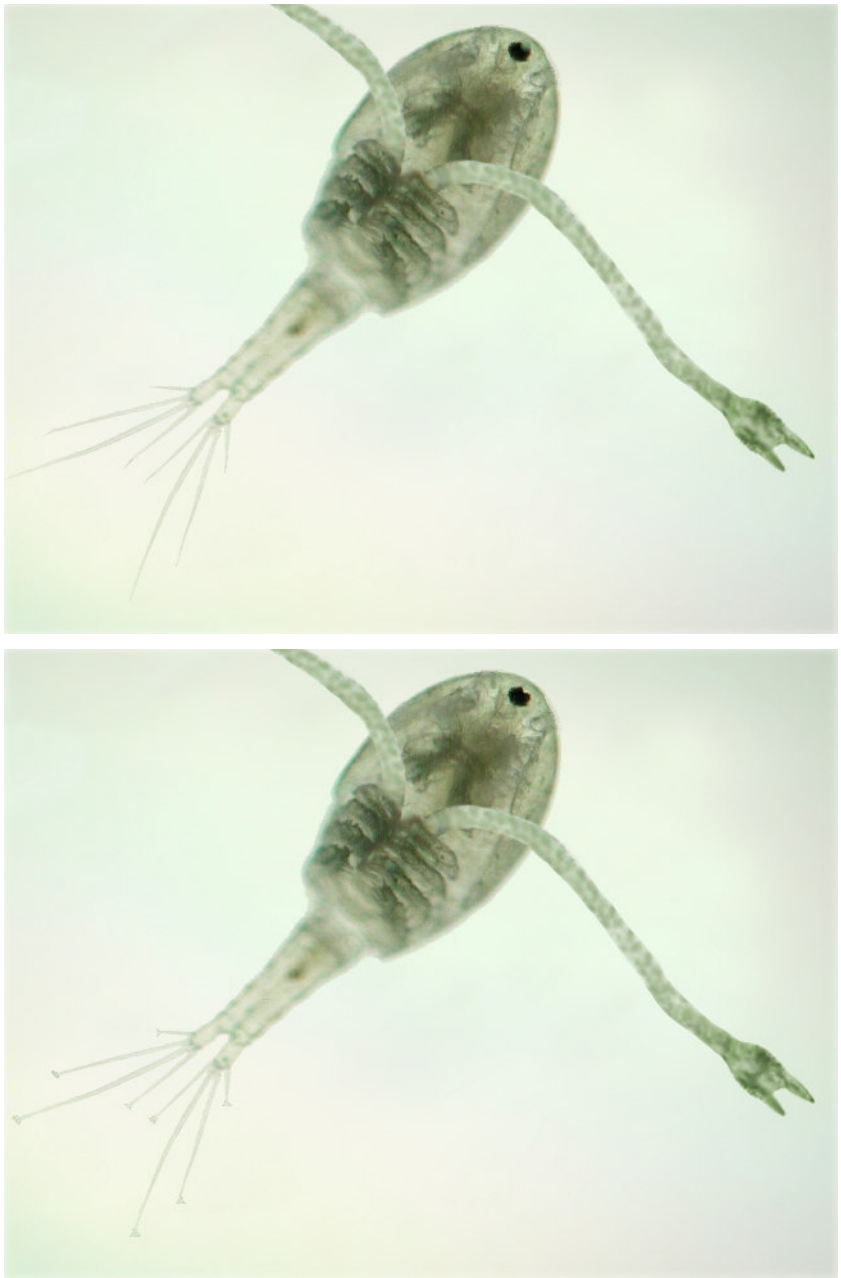


Figure S3. Example plankton stimuli, from learning phase. Specimen at top has fine tail, blurry eye, and unconnected claw. Specimen at bottom has blunt tail, dotted eye, and connected claw

Figure S4. Example plankton stimulus, from information-acquisition phase, with eye and claw obscured.

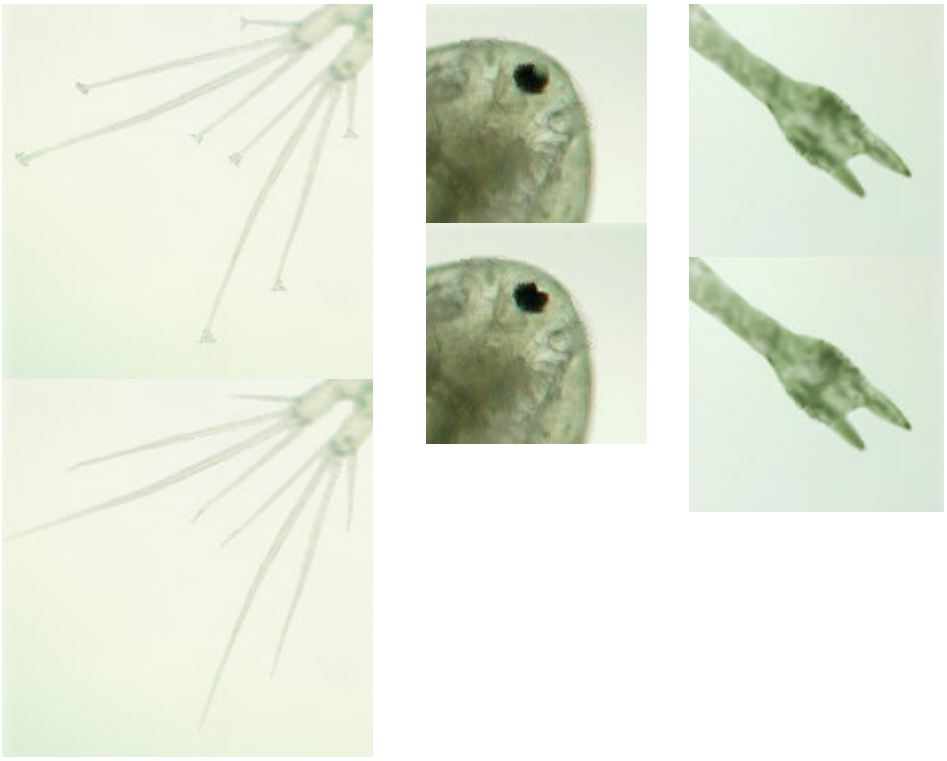


Figure S5. The two versions of each plankton feature: blunt or fine claw (left); blurry or dotted eye (middle), and unconnected or connected claw (right)