Finding useful questions:

on Bayesian diagnosticity, probability, impact, and information gain

Jonathan D. Nelson

University of California at San Diego

Note. This is the final text as approved for publication in *Psychological Review* (anticipated July or October, 2005). An additional file (at http://www.jonathandnelson.com/) has supplementary material. Feedback, ideas, corrections, and criticism are welcome.

Direct correspondence to
Jonathan D Nelson
jnelson@cogsci.ucsd.edu
UCSD Cognitive Science, Dept. 0515
La Jolla, CA 92093-0515
USA

Abstract

Several norms for how people should assess a question's usefulness have been proposed, notably Bayesian diagnosticity, information gain (mutual information), Kullback-Liebler distance, probability gain (error minimization), and impact (absolute change). Several probabilistic models of previous experiments on categorization, covariation assessment, medical diagnosis and the selection task are shown to not discriminate between these norms as descriptive models of human intuitions and behavior. Computational optimization found situations in which information gain, probability gain, and impact strongly contradict Bayesian diagnosticity. In these situations, diagnosticity’s claims are normatively inferior. Results of a new experiment strongly contradict the predictions of Bayesian diagnosticity. Normative theoretical concerns also argue against use of diagnosticity. It is concluded that Bayesian diagnosticity is normatively flawed and empirically unjustified.

Finding useful questions:

on Bayesian diagnosticity, probability, impact, and information gain

When learning the meaning of a new word, diagnosing a patient’s illness, interviewing job candidates, or testing scientific hypotheses, choice of questions (experiments, tests, queries) is critical. Whether people can identify useful questions, and what exactly constitutes a useful question, are central problems in theories of human cognition. Popper’s (1959) falsificationist philosophy of science and Piaget and Inhelder's (1955/1958) experimental research inspired Wason's (1960, 1966) 2-4-6 task, which was designed to be a miniature scientific problem. Wason suggested that success on that task required "a willingness to attempt to falsify hypotheses” (p. 139). However, many subjects had difficulty devising tests to falsify their working hypotheses, a pattern sometimes called "confirmation bias" (Wason & Johnson-Laird, 1972, cited in Mynatt, Doherty, & Tweney, 1977; reviewed by Klayman, 1995). Many recent researchers have distanced themselves from the falsificationist view, and suggested that *differentiation* of plausible hypotheses is normatively a better goal than *falsification* of the working hypothesis. Several of these researchers have described evidence-acquisition situations in a probabilistic framework (Trope & Bassok, 1982, 1983; Fischoff & Beyth-Marom, 1983; Baron, 1981, cited in Baron, 1985, pp. 130-167; Skov & Sherman, 1986; Oaksford & Chater 1994, 2003).

An advantage of the probabilistic approach is the ability to differentiate the following components:

1. A *probabilistic belief model* with a set of hypotheses, a prior probability of each, and a set of possible questions (or experiments, queries, or tests) to differentiate between them;
2. A *sampling norm* to quantify the expected usefulness of each possible question,

relative to a probabilistic belief model;

3. A method to update beliefs according to a test's outcome.

In models of sequential tasks, in which the first question's answer is known before the second question is asked, an additional component, to check whether a stopping criterion has been reached, is also needed. (Similar accounts appear in Box & Hill, 1967; Fischoff & Beyth-Marom, 1983; Over & Jessop, 1998; and Zimmerman, 2000.) Most experimental research in this area (Table 1) assesses the adaptiveness of information-gathering behavior by considering whether people choose highly useful queries, as identified by a particular sampling norm. The goal of the present paper is to illuminate differences between sampling norms (component 2) by holding constant both probability belief models (component 1) and the method of belief updating (component 3).

In this paper, the intuitive scientist metaphor as outlined above is used as a descriptive model of human cognition, and Bayes' (1763) theorem is used to update beliefs. Use of normative models as descriptive models, or to facilitate development of descriptive models, is a familiar research strategy (Anderson's 1990, 1991, rational analysis; Brunswik's, 1952, molar analysis; Peterson & Beach, 1967; Marr, 1982; Viswanathan, et al., 1999; Oaksford & Chater, 2001; McKenzie, 2003; Baron, 2004). However, Kuhn (1989, 2000) discusses both the usefulness and limitations of the normative intuitive scientist metaphor as a descriptive model of human cognition. It has also been suggested that humans change their beliefs to a lesser extent than predicted by Bayes' (1763) theorem, perhaps by using a *conservative* Bayesian form of belief revision (Edwards, 1968). Setting these issues aside enables the present paper to focus on the relative normative and descriptive justification of several sampling norms in the psychological literature, each of which explicitly describes what makes a question (experiment, query, or test) useful.

From a mathematical standpoint, method of belief revision (Bayesian or other) and choice of sampling norm (to assess possible questions' expected usefulness) are completely independent of each other. This notwithstanding, some researchers have implied that Bayesian diagnosticity is the only theoretically (normatively) defensible, or the only Bayesian, sampling norm. For instance, Slowiaczek et al. (1992) stated "according to Bayes' theorem, the diagnosticity [usefulness] of an answer or datum, D, depends on the likelihood ratio," (p. 393), e.g. on that answer's Bayesian diagnosticity. (Similar statements appear in Beyth-Marom & Fischoff, 1983, p. 1193; Fischoff & Beyth-Marom, 1983, pp. 240-241; Bassok & Trope, 1983-1984, p. 200; and Doherty, et al., 1996, p. 644.) Evans and Over (1996) stated log diagnosticity would be "much more satisfactory as a normative standard" than information gain (p. 358). Good (1975) simply said diagnosticity "was central to my first book (1947/50) and also occurred in at least 32 other publications … . What I say thirty-three times is true" (pp. 52-53).

Yet several norms have been proposed for evaluating the usefulness of a question (or test or experiment) in probabilistic evidence-gathering situations. Prominent proposals include information gain, Kullback-Liebler distance, impact (absolute change), probability gain (minimal error), Bayesian diagnosticity, and log diagnosticity (Table 1). The literature to date, however, does not make clear (1) what norms best describe human behavior, (2) when the norms disagree, or (3) whether some norms are theoretically (normatively) better motivated than others.

The rest of this paper is structured as follows. Each sampling norm is explicitly defined below. (Appendix 1 provides a more intuitive scenario and example calculations.) Prior experimental evidence-acquisition research is reanalyzed to examine the extent to which earlier researchers' conclusions depend on the sampling norm used. New simulations demonstrate that the sampling norms can disagree with each other, and that this disagreement occurs in a variety of statistical environments. Further simulations identify particularly strong cases of disagreement between norms. Those limiting cases are used to design a definitive experiment, to address whether diagnosticity and log diagnosticity are plausible descriptive psychological models of human intuitions. Finally, theoretical objections to diagnosticity and log diagnosticity are discussed, and important issues for future research are outlined.

[insert Table 1 about here]

The sampling norms

There is a conceptual distinction between *disinterested* utility functions for evidence acquisition, and *situation-specific* utility functions, for situations with unique reward structures (Lindley, 1956; Box & Hill, 1967; Baron & Hershey, 1988; Kirby, 1994; Chater, Crocker, & Pickering, 1998; Chater & Oaksford, 1999). Disinterested utility functions are useful for information gathering, where no immediate action is required; Chater, Crocker, and Pickering liken them to pure scientific research. Situation-specific utility functions are appropriate when making the best decision is more important than believing the correct hypothesis. However, the same mathematical framework can be used in both cases (Savage, 1954, pp. 105-119; Raiffa, 1968). Each norm discussed in this paper—diagnosticity, log diagnosticity, information gain, KL distance, probability gain, and impact—can be thought of as a subjective utility function for evidence acquisition. The norms give different definitions of an individual *answer's* usefulness. However, all of the norms define a *question's* usefulness as the *expected usefulness* of its possible answers, averaged according to the probability of each possible answer occurring.

A technical definition of each norm is given below. Appendix A provides a more intuitive treatment, in the context of the Vuma probability model, with example calculations and further discussion. In this paper's notation capital letters represent random variables; lowercase letters represent specific values that those random variables can take. Q is a question, query, test, or experiment, whose results are unknown; q_j are specific answers, or experiment results; C is the unknown category, or hypothesis; and c_i are particular categories, or hypotheses.

Bayesian diagnosticity

Good (1950, 1975, 1983) introduced the concept of diagnosticity. He called an answer's diagnosticity the "weight of evidence" and a question's diagnosticity the "expected weight of evidence." Explicit use of the term *diagnosticity* in this context appeared at least as early as Edwards (1968, pp. 25-27).[1] Diagnosticity relates to the likelihood ratio of particular data given two categories:

$$\mathit{diagnosticity}(q_j) = \max\left(\frac{P(q_j \mid c_1)}{P(q_j \mid c_0)}, \frac{P(q_j \mid c_0)}{P(q_j \mid c_1)}\right), \text{ and}$$

$$\mathit{diagnosticity}(Q) = \sum_{q_j} P(q_j) * \mathit{diagnosticity}(q_j).$$

(Max(a,b) denotes the larger of a and b. If equal pick randomly.)

Log$_{10}$ diagnosticity

The set of researchers using log diagnosticity is virtually the same as the set of researchers using diagnosticity. Some researchers first introduce diagnosticity, then use log diagnosticity for their calculations, without addressing whether those sampling norms may disagree, or giving a rationale for preferring one or the other (Evans & Over, 1996; McKenzie & Mikkelsen, in press). But their expectations are not equivalent: a question's diagnosticity cannot be derived from its log diagnosticity, or vice versa. Base 10 ("log$_{10}$ diagnosticity") will be used in this paper; constant positive multiples convert between bases. Specifically:

$$log_{10}\ diagnosticity\left(q_j\right) = \log_{10} \max\left(\frac{P\left(q_j \mid c_1\right)}{P\left(q_j \mid c_0\right)}, \frac{P\left(q_j \mid c_0\right)}{P\left(q_j \mid c_1\right)}\right) = abs\left(\log_{10} \frac{P\left(q_j \mid c_1\right)}{P\left(q_j \mid c_0\right)}\right) \text{, and}$$

$$log_{10}\ diagnosticity(Q) = \sum_{q_j} P\left(q_j\right) * log_{10} diagnosticity\left(q_j\right).$$

It should be emphasized that the two above formulations of an answer q_j's $\log_{10}$ diagnosticity, one taking the maximum likelihood ratio, the other taking an absolute log likelihood ratio, are identical. Diagnosticity and log diagnosticity are both infinite for an answer q_j that eliminates a category (hypothesis) by setting its posterior probability to 0. The descriptive psychological claim is that such an answer, or a question Q with some probability of it, is infinitely useful.

Information gain

One way to measure a question's usefulness is by quantifying its expected reduction in uncertainty, or *information gain,* with respect to the true hypothesis, or category. Lindley (1956), Box (1967), and Fedorov (1972) quantified this idea explicitly, using Shannon's (1948) entropy to measure uncertainty. Good (1950, pp. 74-75) also alluded to the possibility of using a sampling norm based on Shannon entropy. Baron (1985, pp. 150-151; citing Marschak, 1974) suggested that using information gain would be appropriate in rare situations only, such as when maximizing data transmitted through a telephone line. Outside the realm of cognitive tasks, information gain has been employed to model development of visual neurons (Ruderman, 1994; Ullman, Vidal-Naquet, & Sali, 2002) and auditory neurons (Lewicki, 2002), and to set camera parameters in computer vision (Denzler & Brown, 2002). This paper measures information with base 2 logarithms (*bits*); one bit equals $\log_e 2 = 0.6931$ *nats* of information. The information gain in Q is the mutual information between C and Q (Cover & Thomas, 1991):

$$I(C,Q) = H(C) - H(C \mid Q) \text{, where}$$

$$H(C)=\sum_{c_i} P(c_i) \; * \log_2 \frac{1}{P(c_i)}$$, the initial entropy in *C*,

$$H(C \mid Q)=\sum_{q_j} P(q_j) * H(C \mid q_j)$$, the conditional entropy in *C* given *Q*, and

$$H(C \mid q_j)=\sum_{c_i} P(c_i \mid q_j) * log_2 \frac{1}{P(c_i \mid q_j)}$$, the entropy in *C* given a particular answer *q*$_j$.

Kullback-Liebler distance

A question's usefulness could also be quantified as the amount that its answer is expected to *change* one's beliefs. Kullback-Liebler distance (Kullback & Liebler, 1951; Cover & Thomas, 1991) provides one means to measure the change from prior beliefs about the true category, *C*, to posterior beliefs after a particular question is answered:

$$KL\ distance(q_j)=\sum_{c_i} P(c_i \mid q_j) * \log_2 \frac{P(c_i \mid q_j)}{P(c_i)}$$, and

$$KL\ distance(Q)=\sum_{q_j} P(q_j) * \sum_{c_i} P(c_i \mid q_j) * \log_2 \frac{P(c_i \mid q_j)}{P(c_i)}.$$

In this paper, KL distance and information gain are equivalent, because they give identical measures of a *question's* usefulness (Oaksford & Chater, 1996). (Table A1 illustrates that they give different statements of particular *answers'* usefulness.)

Probability gain

Baron (1981, cited in Baron, 1985) suggested this norm as a special case of Savage's (1954, chap. 6, pp. 105-119) analysis of the value of observations, in which the inquirer assigns the same utility to any correct guess, and lower, equal utility to any incorrect guess. Assuming that the most probable category is chosen after a question's answer is obtained, it's possible to calculate how much asking a particular question improves the expected probability of making a correct guess, the question's *probability gain*. Maximizing probability gain is equivalent to minimizing probability of error, a common criterion in computer science, as well as to maximizing probability of making a correct decision. Specifically:

$$probabilityGain(q_j) = \max_{c_i} P(c_i \mid q_j) \quad - \quad \max_{c_i} P(c_i), \text{ and}$$

$$probabilityGain(Q) = \left(\sum_{q_j} P(q_j) * \max_{c_i} P(c_i \mid q_j) \right) \quad - \quad \max_{c_i} P(c_i).$$

Impact (absolute change)

Impact is based on the idea that answers q_j that change beliefs are useful. In this respect impact is similar to KL distance; however, it is based on a different measure of belief change. Klayman and Ha (1987, pp. 219-220) and Nickerson (1996), discussing belief models with two hypotheses, suggested measuring a particular answer's *impact* on an individual hypothesis as abs(P(hypothesis | answer) – P(hypothesis)); or, in present notation, abs(P(c_1 | q_j) – P(c_1)). If a belief model contains exactly two hypotheses, as Nickerson noted, then a particular answer has the same impact on each hypothesis. The present paper generalizes impact to situations with multiple categories or hypotheses:

$$impact(q_j) = \frac{1}{n} \sum_{c_i} abs(P(c_i \mid q_j) - P(c_i)), \text{ where } n \text{ is the number of categories } c_i, \text{ and}$$

$$impact(Q) = \sum_{q_j} P(q_j) \frac{1}{n} \sum_{c_i} abs(P(c_i \mid q_j) - P(c_i)).$$

Slowiaczek et al. (1992, p. 402) reported that many subjects used the *heuristic* strategy of asking about the feature with the maximal difference in feature probabilities, abs(P(feature | c_1)- P(feature | c_2)). It is presently shown that this strategy is not merely heuristic, but exactly implements impact.[2] If prior probabilities of two hypotheses are equal, impact and probability gain are identical. Several models in the literature meet this condition (Slowiaczek et al., 1986; Skov & Sherman, 1992; Oaksford & Chater, 1994, 1998, 2003; Nickerson, 1996). (After an answer has been obtained, posterior probabilities of two hypotheses—which become priors for purposes of evaluating successive questions—will in general no longer be equal.)

Properties of the sampling norms

The psychological plausibility of claiming that subjective utility is infinite is questionable, so it is of interest to note whether each norm is finite. Similarly, it has been argued (Evans & Over, 1996) that it is psychologically implausible that an answer that changes beliefs can have negative utility. It is thus useful to note which norms are nonnegative. Finally, is the usefulness of obtaining two pieces of data (answers q_1 and q_2) simultaneously the sum of the usefulness of obtaining each datum separately? Intuitively, it seems that if two data in effect cancel each other out, such that posterior beliefs after obtaining both of them are the same as prior beliefs, those data were useless.[3] Under which sampling norms is this the case? Table 2 lists which of these properties each sampling norm satisfies.

[insert Table 2 about here]

Prior research and the descriptive plausibility of each norm

Most research that assesses people's faculties at identifying useful questions has used a single sampling norm to calculate each question's usefulness. This raises the possibility that using other normative models would result in different conclusions. To what degree do earlier researchers' conclusions about people's sensitivity to questions' usefulness depend on the specific sampling norm used? This section considers several belief models, each with specific hypotheses, prior probabilities, and available questions. For each belief model, the usefulness of each question is computed using diagnosticity, information gain-KL distance, probability gain, and impact. Because "the complexity of any real decision problem defies complete explicit description" (Savage, 1954, p. 107), the probability models below reflect simplified experimental tasks. For instance, although Baron, Beattie, and Hershey (1988) studied a simple case of medical diagnosis, extension of that analysis to actual medical decision making is not straightforward (Cohen, 1996; Baron, 1996). Nor is it straightforward to predict how subjects will interpret an experimental task (McKenzie, Wixted, & Noelle, 2004). On Wason's selection task, there are several proposed probability models. An evenhanded approach, for the present paper, is to exactly implement the original researchers' belief model, in each experiment considered.

Two-category, binary-feature tasks. Skov and Sherman (1986) and Slowiaczek et al. (1992, experiments 3a and 3b) designed a task to be a case of miniature scientific inference, but in which appropriateness of a particular belief model was clear. Subjects were told the distribution of gloms and fizos, the two creatures on planet Vuma, and the distribution of various binary features within gloms and fizos. Subjects were asked to indicate which features they would ask about, to determine whether a novel creature were a glom or fizo. Skov and Sherman and Slowiaczek et al. used diagnosticity or log diagnosticity to measure the usefulness of questions. In Skov and Sherman's experiment, as many as 5 high diagnosticity questions could be chosen: 68% of the 66 subjects chose 5 high diagnosticity questions; an additional 18% chose 4 high diagnosticity questions. In Slowiaczek et al.'s experiment 3a, a single question was selected; 98% of subjects (196 of 199) chose a high diagnosticity question.

Do Skov and Sherman's (1986) or Slowiaczek et al.'s (1992, experiment 3a) results depend on using diagnosticity or log diagnosticity to measure the usefulness of questions? In the present analysis, each question's usefulness was calculated using information gain-KL distance, probability gain, impact, diagnosticity and log diagnosticity. All norms agree with Skov and Sherman's high>medium>low ordering, and with Slowiaczek et al's high>low ordering, of the usefulness of each question (Table 3). This corroborates the earlier findings that people tend to select reasonable questions. But neither experiment's behavioral results differentiate the relative plausibility of the sampling norms under consideration as descriptive models.

[insert Table 3 about here]

Slowiaczek et al. (1992, experiment 3b) also sought to address whether people prefer questions with extreme feature probabilities; e.g. wearing a hula hoop (Hula), present in 90% of gloms and 55% of fizos; versus drinking iced tea (Drink), found in 65% of gloms and 30% of fizos. The intent was to hold usefulness constant while modifying the extremity of feature probabilities (Table 4) in a pair of questions. Unfortunately, the stimuli confounded extremity with several sampling norms, such that the questions with more extreme feature probabilities also had higher diagnosticity, log diagnosticity, and information gain-KL distance. Impact and probability gain were indifferent between the two questions in each pair. Slowiaczek et al.'s behavioral results were that on average, the feature with higher diagnosticity, log diagnosticity, and information gain-KL distance was chosen around 60% of the time (chance would be 50%). It is difficult to make strong inferences about the sampling norms from this result.

[insert Table 4 about here]

Medical diagnosis. Baron, Beattie, and Hershey (1988), experiments 4, 5, and 6, asked subjects to rate the usefulness of several medical tests. The task instructions described a situation in which probability gain would be the most obviously justifiable measure.[4] Only one test could be conducted before diagnosing and treating the disease. Diseases were described abstractly, as diseases A, B, and C; their prior probabilities were 0.64, 0.24, and 0.12, respectively. Thus the diseases were presumably equally problematic, if untreated, and equally treatable. Subjects were given the conditional probabilities that each of several tests would come out positive or negative, given each disease.[5]

Subjects rated each test's usefulness on a 0-100 scale "where 0 means the test is worthless and should not be done and 100 means the test would remove all doubt about which disease the patient has" (p. 101). Baron et al. found that subjects were generally sensitive to the relative usefulness of each test, as measured with probability gain. However, subjects consistently gave positive ratings to some tests that were useless, as calculated by probability gain. Baron et al. referred to this tendency as information bias.

Could the idea that subjects were making use of a sampling norm besides probability gain better explain Baron et al.'s (1988) data? To address this, the probability gain, impact, and information gain-KL distance of each test were computed (Table 5 and Table 6). Impact and information gain-KL distance agree with subjects that some zero probability gain tests are useful (tests 3-8 in experiment 4). In other words, if impact or information gain were to be deemed appropriate normative models of this task, information bias would largely disappear. Overall, however, while subjects' ratings correlate highly with each sampling norm, there is no clear pattern wherein a particular sampling norm best accounts for responses (Table 7).

[insert Table 5 about here]

[insert Table 6 about here]

[insert Table 7 about here]

The abstract selection task. In a typical version of this task, introduced by Wason (1966, 1968), a subject is shown the top faces of four cards, showing A, 2, K, and 3. The subject is asked what cards would need to be flipped to falsify the rule that *if a card has an A on one side,* *it has a 2 on the other side*. Wason intended the selection task to be a deductive logical task, in which the 2 card and the K card would be useless. But very few subjects (seldom 10%) select just the A and 3 card (Stanovich & West, 1998); most subjects do select the 2 card, which Wason took to be a mistake. Over several decades of subsequent study involving around 1000 experimental subjects, the ordering of the most frequently selected cards has been A>2>3>K (Oaksford & Chater, 1994).

Oaksford and Chater (1994, 1998, 2003) introduced probabilistic, rather than logical, models of the selection task and proposed that subjects choose cards to maximize information gain[6] with respect to their beliefs, rather than to falsify a particular hypothesis. There has been extensive debate about this approach (Evans & Over, 1996; Laming, 1996; Almor & Sloman, 1996; Oaksford & Chater, 1996, 1998, 1999; Over & Jessop, 1998), and other probabilistic models have also been proposed (e.g. Kirby, 1994; Klauer, 1999; Hattori, 2002). However, novel predictions have been tested behaviorally (e.g. Oaksford, Chater, & Grainger, 1999), in turn improving the model. Oaksford & Chater's (2003) belief model includes two hypotheses: a *dependence hypothesis*, that every card with an A on one side does have a 2 on the other side; and an *independence hypothesis*, that A, K, 2, and 3 are assigned independently, with the constraint that each card has a letter on one side and a number on the other side. The model requires four parameters: probability of the dependence hypothesis, overall P(A), overall P(2), and probability of an error under the dependence hypothesis, under which A is paired with 3. The general constraints specified by Oaksford & Chater are that A's and 2's are rare and that the combination of parameters lead to a valid probability distribution under the dependence hypothesis (Over & Jessop, 1998, explicitly specify these constraints). Among other results, Oaksford and Chater found that the ordering of the information gain of each card, A>2>3>K, matched card selection frequencies.

Would sampling norms other than information gain provide similar results? Oaksford and Chater's (2003) model was implemented here, fixing P(dependence hypothesis)=0.5, P(error)=0.1, P(A)=0.22 and P(2)=0.27. Each norm gave the same ordering (Table 8), suggesting that at this level of analysis, the model does not differentiate the sampling norms under consideration.[7] Would other parameter settings strongly differentiate the norms? If P(error)=0, both diagnosticity and log diagnosticity rate the A and 3 cards as infinitely useful. This is because if P(error)=0, each of these cards offers a chance of eliminating the dependence hypothesis, which diagnosticity and log diagnosticity consider infinitely useful. Diagnosticity and log diagnosticity are indifferent to the relative probability of eliminating a hypothesis when selecting the A or 3 card in this case. If P(error)=0.01, diagnosticity rates the A and 3 cards as most useful; log diagnosticity rates the A and 2 cards as most useful. To summarize: if subjects believe P(error) is very low, diagnosticity and/or log diagnosticity contradict abstract selection task data. Information gain and impact-probability gain maintain the A>2>3>K ordering when P(error)=0.01 or 0. A further note is that Evans and Over (1996) objected to information gain per se, but not to Oaksford and Chater's probability model. Present results show that a sampling norm as intuitive as probability gain or impact could be used.

[insert Table 8 about here]

Covariation assessment. How are variables related to each other? Inhelder and Piaget's studies (1955/1958) set a foundation for research in several areas.[8] This section focuses on covariation assessment (Inhelder & Piaget, 1955/1958; Smedslund, 1963; Peterson & Beach, 1967; McKenzie, 1994). A typical task involves two binary variables: *X* (e.g., glom or not) and *Y* (hulaWorn or not). Each individual observation falls in one cell of a matrix with four cells: cell

A, glom wearing a hula hoop; B, glom not wearing a hula hoop; C, fizo wearing a hula hoop; or D, fizo not wearing a hula hoop. Subjects frequently are shown a matrix with counts of the number of individuals in each of those four cells. Which cell's observations are the most informative with respect to the goal of determining whether the variables *X* and *Y* covary? Most normative models treat the four cells as equally useful. Yet over several experimental manipulations (reviewed in McKenzie & Mikkelsen, in press), subjects treat A as most useful, and D as least useful, with B and C in between: A>B≈C>D. This differential evaluation has been considered suboptimal.

McKenzie and Mikkelsen (in press) proposed that subjects may be approaching covariation assessment tasks as inferential tasks, and using their prior beliefs to interpret the tasks. For instance, subjects may have the goal of finding out which of two hypotheses is true: h_1, that there is a moderate correlation between *X* and *Y*, or h_0, that *X* and *Y* are independent. Each hypothesis specifies the probability that an observation will fall in each of the cells A through D. Presence of *X* (glom) and *Y* (hulaWorn) are each rare (10% probability under both h_1 and h_0) in the model, corresponding to several researchers' findings that subjects usually assume rarity in related tasks (Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2000; McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable, 2001; Oaksford & Chater, 2003). McKenzie and Mikkelsen calculated the $\log_2$ diagnosticity of an observation in each cell, relative to their probability model. McKenzie and Mikkelsen's model gave the ordering A>B=C>D, providing a rational explanation of one of the main covariation assessment research findings. In the present analysis, McKenzie and Mikkelsen's probability model was implemented,[9] and each cell's usefulness was calculated, relative to information gain, KL distance, probability gain, impact, diagnosticity, and log diagnosticity. All sampling norms agree on the A>B=C>D ordering. This result bolsters

McKenzie and Mikkelsen’s conclusion that subjects’ behavior is justifiable, but does not provide evidence against any sampling norm as a descriptive model of subjects’ behavior.

[insert Table 9 about here]

Do the sampling norms ever disagree?

The previous section showed that the sampling norms under consideration behave similarly with respect to several probability belief models. This result corroborates previous researchers’ frequent finding that participants are sensitive to questions' and answers' usefulness. However, this result does not directly address when the sampling norms make contradictory claims. Simulations were therefore conducted to address the frequency, and pervasiveness in different environments, of disagreements between the sampling norms.

Simulation 1.0, Multiple features. How often do any of the norms disagree with each other, and does the number of features relate to the frequency of disagreement? To address this, the Vuma scenario (described in the above review of Skov & Sherman, 1986, and Slowiaczek et al., 1992, and in Appendix A) was simulated, with P(glom)=P(fizo)=0.5, and random feature probabilities, such as P(hulaWorn | glom). (“Random probabilities” denotes pseudorandom numbers independently sampled from a uniform distribution between [0,1].) Ten thousand random trials were run for each number of features between 2 and 20. Each trial was analyzed to determine whether there was disagreement between diagnosticity, log diagnosticity, information gain-KL distance, and probability gain-impact on the relative usefulness of each feature. Number of disagreements increased monotonically with the number of features. Disagreements occurred in 7% of 2 feature trials, a majority of 6 feature trials, and more than 99% of 15 feature trials.

Simulations 1.1-1.3, Several environments.[10] Is the existence of disagreements between sampling norms restricted to a particular environment? In each simulation, 1,000,000 random trials were generated. Simulation 1.1 used random feature probabilities and random prior probabilities. Simulation 1.2 used equal prior probabilities, and random feature probabilities. Simulation 1.3 considered an environment where both features were “rare” (feature probabilities between 0 and 0.5), as subjects in hypothesis-testing tasks usually assume (Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2000; McKenzie et al., 2001; Oaksford & Chater, 2003), but was otherwise identical to Simulation 1.1. Results showed that in each simulation, every possible type of pairwise disagreement (where one norm prefers Hula and another norm prefers Drink) occurred. Likewise, each simulation produced cases of disagreement with and without extreme feature values: having a feature probability close to 0 or 1 is not necessary for a disagreement to occur. In simulations 1.1 and 1.3, where P(glom) was random, sometimes both questions had zero probability gain. Simulation 1.3, with rare features, had results similar to Simulation 1.1. Together, these simulations show that in a variety of environments, using one norm leads to asking different questions than using another norm. Some cases of disagreement were qualitatively stronger than others, an issue addressed below.

Simulation 2: Cases of strongest disagreement

Could identification of cases of strong disagreement elucidate the differences between the sampling norms? This section describes a simulation to search for limiting cases.

Methods. A simulation automatically searched for cases of high pairwise disagreement strength (*DStr*, defined in Appendix B) between diagnosticity, log diagnosticity, information gain-KL distance, probability gain, and impact. Each optimization used fixed prior probabilities, specified by P(glom), and began with random feature probabilities, in which the two norms being compared disagreed about which feature was more useful. Optimizations to maximize the disagreement between all ten pairs (5 choose 2 = 10) of norms were run, for each P(glom) between 0.50, 0.505, 0.51, … 0.995. The optimization procedure was allowed to find feature probabilities of 0, any number between 0.0001 and 0.9999, or 1. Each trial was repeated 10 times; feature probabilities in the trial with the highest DStr were recorded.

Results and discussion. Figure 1 shows the maximum obtained DStr as a function of P(glom), for each pair of norms. A surprising result, considering that some researchers have used diagnosticity and log diagnosticity interchangeably, is the consistently high disagreement between diagnosticity and log diagnosticity (dashed line at top of figure). In each of the trials in this particular optimization, diagnosticity's claims were strongly suboptimal with respect to all other sampling norms; log diagnosticity agreed with information gain-KL distance, probability gain, and impact. In all other optimizations where very high disagreement (DStr near 100) was observed, diagnosticity and log diagnosticity agreed with each other in each trial, but were suboptimal with respect to all other norms. In these trials diagnosticity and log diagnosticity were unduly influenced by the occasional presence of a certainty-inducing answer, which occurred because a feature probability was 1 or 0. Information gain-KL distance, probability gain, and impact had only moderate degrees of disagreement with each other, suggesting that they may be more closely related to each other (three lines in middle and bottom of figure).

Representative cases of disagreement are discussed below; Appendix B shows how patterns change as a function of P(glom). (Additional results are included in the supplementary material posted online.) In most of the individual cases discussed below, P(glom)=0.70; this prior led to high DStr in most optimizations (Figure 1, Appendix B). Where multiple optimizations, for example probability gain vs. diagnosticity, and probability gain vs. log diagnosticity, produced essentially identical feature probabilities, those optimizations are discussed simultaneously.

[insert Figure 1 about here]

Log_{10} diagnosticity vs. diagnosticity (Table B1). Illustrative trial where P(glom)=0.70: P(hulaWorn | glom)=0.9987, P(hulaWorn | fizo)=0.0013, P(drinksTea | glom)=0.0001, and P(drinksTea | fizo)=0.6202; DStr=99.68. Diagnosticity rates the Drink feature as much more useful than Hula; $\log_{10}$ diagnosticity rates Hula as much more useful than Drink. Testing the Hula feature leads to probability 0.9987 of correct guess (remaining uncertainty 0.0137 bit); testing the Drink feature leads to probability 0.8860 of correct guess (remaining uncertainty 0.4763 bit). Similar examples were observed in other limiting cases of disagreement between diagnosticity and $\log_{10}$ diagnosticity. In all of these cases, using diagnosticity to select questions is strongly suboptimal with respect to probability of correct guess (probability gain), reduction in uncertainty (information gain), and absolute change in beliefs (impact).

Information gain vs. diagnosticity; information gain vs. log_{10} diagnosticity; impact vs. diagnosticity; impact vs. log_{10} diagnosticity (Table B2). If a question has any possibility of completely eliminating a hypothesis, then even if that possibility is extremely remote, diagnosticity and log diagnosticity regard that question as infinitely useful. Illustrative trial for information gain vs. diagnosticity: P(glom)=0.70, P(hulaWorn | glom)=0.0001, P(hulaWorn | fizo)=0.9999, P(drinksTea | glom)=0.0001, and P(drinksTea | fizo)=0.0000. Hula leads to 99.99% probability of correct guess (probability gain 0.2999), and to uncertainty 0.0014 bit (information gain 0.8799 bit). Drink gives no improvement in probability of correct guess (probability gain 0), and almost no reduction in uncertainty (information gain 0.00004 bit). Yet the Drink question, with probability 7 in 100,000, results in the drinksTea answer, which provides conclusive evidence that the creature is a glom. Diagnosticity and log diagnosticity therefore consider the Drink question to be infinitely[11] useful. Probability gain, information gain-KL distance, and impact all strongly prefer the Hula question. Essentially identical results occurred throughout the optimizations in which information gain or impact was contrasted with diagnosticity or log diagnosticity.

Probability gain vs. diagnosticity; probability gain vs. log_{10} diagnosticity (Table B3). Illustrative trial for probability gain vs. diagnosticity: P(glom)=0.70, P(hulaWorn | glom)=0.0001, P(hulaWorn | fizo)=0.9999, P(drinksTea | glom)=0.0914, and P(drinksTea | fizo)=0.0000; DStr= 97.48. Hula has probability gain 0.2999; Drink, 0. Hula has information gain 0.8799; Drink, 0.0343. Yet the Drink feature has infinite diagnosticity and $\log_{10}$ diagnosticity. In all cases of these optimizations, diagnosticity and log diagnosticity agreed with each other, and were suboptimal with respect to information gain, impact, and probability gain. These optimizations' results are similar in some respects to the previous example comparing information gain and diagnosticity. However, probability gain is zero for a wider range of feature probabilities than information gain and impact. In the present optimization that tends to result in more variable feature probabilities in the question that diagnosticity and $\log_{10}$ diagnosticity prefer. (Had the present example contrasted information gain with diagnosticity, P(drinksTea | glom) would have been expected to be 0.0001, rather than 0.0914.)

Information gain vs. probability gain (Table B4). Recall that information gain generally prefers features with an extreme feature probability, especially features with an extreme feature probability given the working hypothesis. By contrast, probability gain prefers features with an extreme feature probability given the working hypothesis, but not extreme feature probabilities in general. In this optimization, information gain preferred a feature with an extreme feature probability given the *alternate* hypothesis, e.g. P(hulaWorn | fizo)≈0 or 1. Probability gain preferred a feature in which the more extreme of the two feature probabilities was paired with the working hypothesis (glom). Illustrative example where P(glom)=0.70, P(hulaWorn | glom)=0.5714, P(hulaWorn | fizo)=0.0000, P(drinksTea | glom)=0.9525, and P(drinksTea | fizo)=0.5587; DStr=58.04. Hula had probability gain zero; Drink had probability gain 0.0991. Information gain, however, was greater for Hula (0.2813 bit) than for Drink (0.1577 bit). Impact agreed with probability gain when P(glom) was between 0.50 and 0.58, and with information gain for more extreme values of P(glom). In most cases the feature preferred by information gain offered the possibility of a certain result, and was favored by diagnosticity and log diagnosticity.

Information gain vs. impact (Table B5). Information gain has a preference for extreme feature probabilities, especially extreme feature probabilities given the working hypothesis (glom where P(glom)>0.50). If feature difference is held constant, then impact has no preference between features with extreme feature probabilities given the working or alternate hypothesis, or without extreme feature probabilities altogether. This optimization gave each feature preferred by information gain (Hula) probability of 1 or 0 given the working hypothesis. Compared with the Hula features, the Drink features were given less extreme feature probabilities overall. To further minimize their information gain, the Drink features are asymmetric in the sense that the relatively extreme feature probability was paired with the alternate hypothesis (fizo). Illustrative example, where P(glom)=0.70: P(hulaWorn | glom)=1, P(hulaWorn | fizo)=0.6247, P(drinksTea | glom)=0.2770, and P(drinksTea | fizo)=0.7791; DStr=39.57. Information gain was 0.2213 for Hula, and 0.1604 for Drink; impact was 0.1576 for Hula, and 0.2109 for Drink. Probability gain was 0.1126 for Hula, and 0.0398 for Drink. In this optimization probability gain agreed with impact when P(glom) was between 0.5 and 0.58, and with information gain for more extreme P(glom) values.

Probability gain vs. impact (Table B6). Probability gain and impact are identical if

P(glom)=P(fizo)=0.5, but otherwise can disagree. Impact is a constant multiple of feature difference, for fixed P(glom), and always prefers the question with maximal difference in feature probabilities. For impact's preferred feature (Hula), the optimization appears to have maximized feature difference, subject to the constraint that the feature have probability gain approximately zero. This was achieved by making the most extreme feature probability conditional on the alternate hypothesis, e.g. P(hulaWorn | fizo)≈0 or 1. For the feature that probability gain prefers, the optimization found features with lower feature difference than the corresponding feature preferred by impact, but in which an extreme feature probability was conditioned on the working hypothesis, e.g. P(drinksTea | glom)≈0 or 1, so as to maximize probability gain. Illustrative example where P(glom)=0.70: P(hulaWorn | glom)=0.5714, P(hulaWorn | fizo)=0, P(drinksTea | glom)=1, P(drinksTea | fizo)=0.6966; DStr=55.25. Hula has impact 0.2400; Drink, 0.1274. Hula has probability gain 0; Drink, 0.0910. In each trial of this optimization, information gain agreed with impact.

An experiment to separate sampling norms

Given that people may choose queries in a noisy manner (Hattori, 2002), an experiment to identify what norms best predict human queries should use cases of strong disagreement between norms. This section reports an experiment whose design was based on cases of maximal possible disagreement between norms, given equal priors, as identified by Simulation 2 (Appendix B).

Method. Subjects were undergraduate students in an introduction to cognitive science class at UCSD (*N*=151), who participated as part of class requirements. All subjects gave informed consent. A planet Vuma scenario was used; subjects were asked to rate the possible questions from most to least useful for determining whether a novel Vumian was a glom or fizo. Prior probabilities of glom and fizo were equal.[12] Table 10 gives the feature probabilities of each question, and each norm's calculation of each question's usefulness. A sample stimulus is given in Appendix C. The order of features with particular feature probabilities, and the order of features' labels (Drink, Gurgle, etc.), both varied across subjects.

[insert Table 10 about here]

Results and discussion. Three subjects were excluded for not ranking all features. One subject gave all queries equal rank. Correlations between that subject's data and each sampling norm are set to 0 in the analyses. The most common pattern (32% of subjects) was an exact match to information gain-KL distance. The next most common pattern (27%) was an exact match to probability gain-impact. No responses matched diagnosticity or log diagnosticity. Table 11 summarizes frequent ordering patterns, along with each pattern's Spearman rank correlation with information gain, probability gain, diagnosticity, and log diagnosticity.

[insert Table 11 about here]

Two analyses were conducted. The first analysis addressed whether the number of responses consistent with each norm was consistent with assigning rank orders 1 through 4 at random. The probability of exactly matching a particular norm by chance, under this null hypothesis, is 1/4!= 1/16. The number of subjects matching a particular norm, under this null hypothesis, is approximately normally distributed with mean 9.25, standard deviation 2.94. (Standard deviation is derived from the binomial variable with parameter 1/16, and n=148.) Significantly more subjects gave rankings in accord with information gain ($t(147)$=13.2, p<0.0001) and probability gain ($t(147)$=10.4, p<0.0001) than expected under the null hypothesis. No subjects responded according to diagnosticity or log diagnosticity, less than expected under the null hypothesis (each $t(147)$= -3.1, p<0.01). (All stated p values are two tailed, and have been multiplied by 4 to correct for multiple tests.) The second analysis consisted of all 148 responses' Spearman rank correlations with each sampling norm (bottom row, Table 11). Average correlation with information gain-KL distance was 0.78; probability gain-impact, 0.69; log diagnosticity, -0.22; and diagnosticity, -0.41. These results strongly contradict claims that people choose queries according to diagnosticity or log diagnosticity.

The other norms make relatively similar predictions, and are not strongly differentiated from each other here. However, some points may be noted. Information gain-KL distance had the largest number of exact matches. Probability gain and impact, which made the same predictions (due to equal priors), were second. Subjects were not asked to explain their choices in this experiment. However, several pilot subjects reported using the feature difference strategy, in a scenario with unequal priors, which corresponds to impact. (Slowiaczek et al., 1992 found a similar result in a scenario with equal priors.) If the feature difference strategy is consistently used, then impact is a necessary component of a descriptive theory of human questions.

Theoretical problems with Bayesian diagnosticity and log diagnosticity

It is possible that an inquirer's subjective sense of queries' usefulness might correspond to diagnosticity or log diagnosticity. But that seems unlikely and normatively ill-advised, in part for the following reasons.

Disregard for priors if features are symmetric. Intuitively, it seems that if the inquirer already knows the true category or hypothesis, for instance because all creatures are known to be gloms, no question is useful (Lindley, 1956, p. 987), but if the inquirer is highly uncertain, many questions are useful. Consider the symmetric feature probability case, where P(hulaWorn | glom) = 1-P(hulaWorn | fizo). Diagnosticity and $\log_{10}$ diagnosticity rate the Hula question as equally useful, irrespective of whether gloms comprise 1%, 20%, 50%, or 99.9999% of the creatures on Vuma. The other sampling norms recognize that if the true category is known in advance, every question is useless. This disregard for prior knowledge, in cases with symmetric feature probabilities, is an undesirable yet unavoidable consequence of using diagnosticity or log diagnosticity.

Oversensitivity to occasional certainty: Features with multiple values. This paper focused on a situation with binary features. What if answers (features) could take three or more values q_j? Perhaps a single fizo has a hard time hearing "Do you drink iced tea?" because of an injury to its auditory system, and answers "maybe" if a question is not clear. If occurrence of the maybe answer makes it certain that the creature is a fizo, then diagnosticity and log diagnosticity rate the question as infinitely useful. This holds even if the maybe answer is rare, occurring with probability 1/1,000,000. The other norms are not unduly influenced by a certainty-inducing answer with rare occurrence.[13]

Log diagnosticity can be normatively inferior to diagnosticity. Let P(glom)=0.99, P(hulaWorn | glom)=0.99, P(hulaWorn | fizo)=0.50, P(drinksTea | glom)=0.97, and P(drinksTea | fizo)=0.50. Probability gain is zero for both questions. Log diagnosticity prefers Drink, which is suboptimal with respect to diagnosticity, information gain-KL distance, and impact. There are a range of cases like this in which diagnosticity and log diagnosticity contradict each other, and in which log diagnosticity makes the normatively inferior (as judged by information gain-KL distance, impact, and diagnosticity) claim.

Multiple hypotheses/categories. Fischoff and Beyth-Marom (1983, p. 243) stated that diagnosticity was defined in situations with multiple hypotheses, although they did not provide an operational definition; Oaksford and Chater (2003, p. 309) stated the contrary. A literature review conducted for the present paper failed to produce a single example calculation of a query's diagnosticity, in a situation with more than two hypotheses. However, for the sake of argument, an operational definition is provided here. (Generalization to cases with more than three categories, and definition of $\log_{10}$ diagnosticity^{++}, are straightforward.) Note that what has been called the diagnosticity of a question, Hula, could just as easily be called the *pairwise diagnosticity* of Hula with respect to whether a creature is a glom or fizo, and denoted diagnosticity$_{glom\&fizo}$(Hula). Now suppose that jevas, in addition to gloms and fizos, reside on Vuma. Define diagnosticity^{++} as the average of each pairwise diagnosticity:

$$\text{diagnosticity}^{++}\text{ (Hula)} = 1/3 * (\ \text{diagnosticity}_{\text{glom\&fizo}}\text{(Hula)} + \text{diagnosticity}_{\text{glom\&jeva}}\text{(Hula)} + \text{diagnosticity}_{\text{fizo\&jeva}}\text{(Hula)}\ ).$$

How would diagnosticity^{++} perform? To explore this, the diagnosticity^{++} of the medical tests from experiments 4-6 in Baron et al. (1988) was computed. Results showed that with the exception of those tests that can never change beliefs, irrespective of their outcome (and that all norms agree are useless), every test has infinite diagnosticity^{++}. Diagnosticity^{++} is useless in this case, and in every case where each test has non-zero probability of eliminating a hypothesis.

Does it matter? In the real world, would using diagnosticity or log diagnosticity actually be problematic? Suppose one's task were to identify the gender of a passerby, by inquiring about one of several features of interest: whether they have a beard, are wearing a dress, are wearing earrings, etc. Statistics were gathered from one natural environment, the UCSD campus in the afternoon, with each of about 500 passersby classified according to their gender (51% were male; 49% female) and several features (Table 12). What features are most useful, according to each sampling norm? The Hair length feature has maximal information gain-KL distance, probability gain, and impact. Asking about Hair length leads to 93% probability of correctly categorizing gender. The Skirt and Beard features have infinite diagnosticity and log diagnosticity, because 0% of males wore skirts and 0% of females had beards. Asking about Skirt or Beard leads, respectively, to only 52% or 59% probability of correct guess. Although the wearsSkirt and hasBeard answers provide 100% certainty of the person's gender in this population, those answers are infrequent. For a person, or seeing robot, to use diagnosticity or log diagnosticity to select queries in this natural environment would be exceptionally inefficient, with respect to all the other sampling norms. The other norms make reasonable claims, and differ from each other only slightly.

[insert Table 12 about here]

General discussion

Diagnosticity and log diagnosticity lack several useful properties that the other norms each possess, including (1) sensitivity to prior probabilities: if there is minimal uncertainty, no question has high usefulness; (2) being finite; and (3) equal applicability in situations with 2, 3, or 1,000,000 hypotheses or categories. Because diagnosticity and log diagnosticity lack these important properties, contradict this paper's experimental results, and appear unnecessary to explain other empirical data, there appears to be no further purpose for them in normative or descriptive theories of evidence acquisition.

In many evidence-gathering situations, where there is no particular external coercion shaping behavior, more than one sampling norm might reasonably apply. Several researchers explicitly make this point (Baron, 1985; Klayman & Ha, 1987; Klayman, 1987; Slowiaczek et al., 1992; Over & Jessop, 1998; Oaksford & Chater, 2003). Theoretical claims that human evidence seeking is adaptive (or that it is biased) would be bolstered by showing that multiple sampling norms agree on what questions are most useful.

A further issue is whether people actually use normative, versus heuristic, utilities. Various heuristic *confirmatory* or *positive test* strategies have been proposed (Skov & Sherman, 1986; Klayman & Ha, 1987, 1989; Gorman, Stafford & Gorman, 1987; Devine, Hirt, & Gehrke, 1990;

Slowiaczek et al., 1992; reviewed in Klayman, 1995). In some situations, strategies that reduce memory load (Costa-Gomes, Crawford, & Broseta, 2001) might also be used. These and other possibilities, including the sampling norms of the present paper, could be combined to form very flexible heuristic models. Most empirical data might then appear to be better fit by a heuristic model than by a sampling norm, because of the heuristic model's greater complexity. To avoid overfitting, future research comparing simple sampling norms with flexible heuristic models could explicitly balance model complexity with descriptive accuracy (Pitt, Myung, & Zhang, 2002).

Geman and Jedynak (2001) describe situations in which looking only one step into the future, as done in this paper, results in having to ask more questions on average than would be required if an optimal sequence of questions were planned in advance (also see Chernoff, 1959, 1972; Klauer, 1999). Unfortunately, planning an optimal series of questions requires knowledge of the individual features’ conditional dependence on each other, given the true category. For example, if a particular glom drinks gasoline, is he more likely to breathe fire than the average glom? Most tasks have not specified this information, or have not obtained sequences of questions, and are therefore unable to address whether subjects’ queries are sensitive to class-conditional dependencies. In a review of work on perceptual information integration (not involving active sampling), Movellan and McClelland (2002) found that people made appropriate use of features' level of class-conditional dependence. Whether people's queries make use of these intricate statistical relationships is an important issue for future research.

References

Allan, L. G. (1993). Human contingency judgments: rule based or associative? *Psychological Bulletin, 114(3)*, 435-448.

Almor, A., & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review, 103(2)*, 374-380.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98(3)*, 409-429.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition, 23(4)*, 510-524.

Baron, J. (1981). *An analysis of confirmation bias*. Paper presented at 1981 Psychonomic Society meeting.

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

Baron, J. (1996). Why expected utility theory is normative, but not prescriptive. *Medical Decision Making, 16*, 7-9.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*, pp. 19-36. London: Blackwell.

Baron, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: I. Priors, error costs, and test accuracy. *Organizational Behavior and Human Decision Processes, 41*, 259-279.

Baron, J., Beattie, J., & Hershey, J. C. (1998). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes, 42*, 88-110.

Bassok, M., & Trope, Y. (1983-1984). People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition, 2(3)*, 199-216.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

Beyth-Marom, R., & Fischoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45(6)*, 1185-1195.

Box, G., & Hill, W. (1967). Discrimination among mechanistic models. *Technometrics, 9*, 57-71.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: The University of Chicago Press.

Card, W. I., & Good, I. J. (1974). A logical analysis of medicine. In R. Passmore, & J. S. Robson, (Eds.), *A Companion to Medical Studies, Vol. 3*, (Chapter 60). Oxford: Blackwell.

Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: the case for parsing. In N. Chater, & M. Oaksford (Eds.), *Rational Models of Cognition* (pp. 441-468). Oxford: Oxford University Press.

Chater, N., Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38(2)*, 191-258.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics, 30(3)*, 755-770.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. Philadelphia: Society for Industrial and Applied Mathematics.

Cohen, B. J. (1996). Is expected utility theory normative for medical decision making? *Medical Decision Making, 16*, 1-6.

Costa-Gomes, M., Crawford V. P., Broseta, B. (2001). Cognition and behavior in normal-form games: an experimental study. *Econometrica, 69(5)*, 1193-1235.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Denzler, J., & Brown, C. M. (2002). Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2)*, 145- 157.

Devine, P. G., Hirt, E. W., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58(6)*, 952-963.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory & Cognition, 24(5)*, 644-654.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica, 43(2)*, 111-121.

Doherty, M. E., Schiavo, M. B., Tweney, R. D., Mynatt, C. R. (1981). The influence of feedback and diagnostic data on pseudodiagnosticity. *Bulletin of the Psychonomic Society, 18(4)*, 191-194.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment*, (pp. 17-52). New York: John Wiley.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103(2)*, 356-363.

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

Fischoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90(3)*, 239-260.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty.

Psychological Science, 14, 195-200.

Geman, D., & Jedynak, B. (2001). Model-based classification trees. *IEEE Transactions on Information Theory, 47(3)*, 1075-1982.

Ginzburg, I., & Sejnowski, T. J. (1996). Dynamics of rule induction by making queries: transition between strategies. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 121-125.

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

Good, I. J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese, 30*, 39-73.

Good, I. J. (1983). *Good Thinking*. Minneapolis, MN: University of Minnesota.

Good, I. J., & Card, W. I. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine, 10*, 176-188.

Gorman, M. E., Stafford, A., & Gorman, M. E. (1987). Disconfirmation and dual hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 39(1)*, 1-28.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason’s selection task. *The Quarterly Journal of Experimental Psychology, A, 55(4)*, 1241-1272.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures* (A. Parsons & S. Milgram, Trans.). New York: Basic Books. (Original work published 1955 as *De la logique de l'enfant à la logique de l'adolescent: Essai sur la construction des structures opératoires formelles*. Paris: Presses Universitaires de France.)

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79(1, Whole No. 594).*

Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning, 49*, 209-232.

Kirby, K. N. (1994). Probabilities and fictional outcomes on the selection task. *Cognition, 51*, 1-28.

Klahr, D. (2000). Exploring Science: the cognition and development of discovery processes. Cambridge, MA: MIT Press.

Klauer, K. C. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review, 106*, 215-222.

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

Klayman, J. (1987). *An information theory analysis of the value of information in hypothesis testing*. Working paper, Center for Decision Research, University of Chicago Graduate School of Business. Available online at http://www.chicagocdr.org/cdrpubs/

Klayman, J. (1995). Varieties of confirmation bias. In J. R. Busemeyer, R. Hasties, D. L. Medin (Eds.), *Decision Making from a Cognitive Perspective*. New York: Academic Press.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4)*, 596-604.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96(4),* 674-689.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9(5)*, 178-181.

Kuhn, D. (2002). What is scientific reasoning and how does it develop? In U. Goswami (Ed.), Handbook of Childhood Cognitive Development (pp. 371-393). Oxford, UK: Blackwell.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills*. San Diego, CA: Academic Press.

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

Laming, D. (1996). On the analysis of irrational data selection: a critique of Oaksford and Chater (1994). *Psychological Review, 103(2)*, 364-373.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience, 5(4)*, 356-363.

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

Maggi, J., Butera, F., Legrenzi, P., & Mugny, G. (1998). Relevance of information and social influence in the pseudodiagnosticity bias. *Swiss Journal of Psychology, 57(3)*, 188-199.

Marr, D. C. (1982). *Vision: A Computation Investigation into the Human Representational System and Processing of Visual Information*. San Francisco: Freeman.

Marschak, J. (1974). *Economic information, decision, and prediction*. D. Reidel: Dordrecht, Holland.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies. *Cognitive Psychology, 26*, 209-239.

McKenzie, C. R. M. (2003). Rational models as theories—not standards—of behavior. *Trends in Cognitive Sciences, 7(9)*, 403-406.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review, 7*, 360-366.

McKenzie, C. R. M., & Mikkelsen, L. A. (in press). A Bayesian view of covariation assessment. *Cognitive Psychology*.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organizational Behavior and Human Decision Processes, 85*, 291-309.

McKenzie, C. R. M., Wixted, J. T., & Noelle, D. C. (2004). Explaining purportedly irrational behavior by modeling skepticism in task parameters: an example examining confidence in forced-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(5),* 947–959.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: implications for models of perception. *Psychological Review, 108*, 113-148.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29(1)*, 85-95.

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.

Nickerson, R. S. (1996). Hempel’s paradox and Wason’s selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2*, 1-32.

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(2)*, 381-391.

Oaksford, M., & Chater, N. (1998). A revised rational analysis of the selection task: exceptions and sequential sampling. In Oaksford, M., & Chater, N. (Eds.), *Rational Models of Cognition* (pp. 372-393). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (1999). Information gain and decision-theoretic approaches to data selection: Response to Klauer (1999). *Psychological Review, 106(1)*, 223-227.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences, 5(8)*, 349-357.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10(2)*, 289-318.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5(3)*, 193-243.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(2)*, 441-458.

Over, D. & Jessop, A. (1998). Rational analysis of causal conditionals and the selection task. In Oaksford, M., & Chater, N. (Eds.), *Rational Models of Cognition* (pp. 399-414). Oxford: Oxford University Press.

Peterson, C. R., Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68(1)*, 29-46.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109(3)*, 472-491.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Raiffa, H. (1968). *Decision Analysis*. Menlo Park, CA: Addison-Wesley.

Ruderman, D. L. (1994). Designing receptive fields for highest fidelity. *Network: Computation in Neural Systems, 5*, 147-155.

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

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A Re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233-250.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423, 623–656.

Skov, R. B. & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22(2)*, 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 and Cognition, 20(4)*, 392-405.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165-173.

Stanovich, K. E. & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning, 4(3)*, 193-230.

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

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

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis testing. *Journal of Experimental and Social Psychology, 19*, 560-576.

Ullman, S., Vidal-Naquet, M., Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5(7)*, 682-687.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S, da Luz, M. G., Raposo, E. P., Stanley, H. E. (1999). Optimizing the success of random searches. *Nature, 401(6756)*, 911-914.

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

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. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology. 20(3)*, 273-281.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99-149.

Author note

Predoctoral fellowships from NIH grant 5T32MH020002-04 to T. Sejnowski and NSF grant DGE-0333451 to G. Cottrell, and a Pew fellowship supported JDN during this research. Jonathan Baron, Nick Chater, Craig McKenzie, Javier Movellan, Mike Oaksford, and an anonymous reviewer provided thorough feedback on versions of this article, and a number of key ideas that contributed to the research. Gary Cottrell, Flavia Filimon, Carrie Joyce, Laura Kemmer, Irene Merzlyak, Alan Robinson, Eric Wiewiora, and Lingyun Zhang suggested specific ideas and references, proofread this paper, or made other contributions to this research. The natural environment example was presented at the Cognitive Science Society conference in Chicago, Illinois, August 2004.

Correspondence concerning this article should be addressed to Jonathan D. Nelson, UCSD Cognitive Science Department, 9500 Gilman Dr., MC 0515, La Jolla, CA 92093-0515, USA. E-mail: jnelson@cogsci.ucsd.edu.

Appendix A: Example calculations of each sampling norm

Each sampling norm will be discussed in the context of the planet Vuma scenario (Skov & Sherman, 1986; Slowiaczek et al., 1992). The task is to discover whether a novel creature is a *glom* or *fizo*, by asking whether it possesses a particular binary feature, such as wearing a hula hoop or not, or drinking iced tea or not. The probability belief model is specified with five parameters. These parameters are listed here, together with specific example values: the *prior probability* that an individual creature is a glom, P(glom)=0.70, and four *feature probabilities*, P(hulaWorn | glom)=0.1, P(hulaWorn | fizo)=0.9, P(drinksTea | glom)=0.3, and P(drinksTea | fizo)=0.5. Questions are answered truthfully; Bayes' theorem is used to update beliefs. For instance, if a hula hoop is worn:

$$P(glom \mid hulaWorn) = \frac{P(hulaWorn \mid glom) * P(glom)}{P(hulaWorn)} = \frac{0.1 * 0.7}{0.34} \approx 0.21, \text{ where}$$

$$\begin{aligned} P(hulaWorn) &= P(hulaWorn \mid glom) * P(glom) \quad + \quad P(hulaWorn \mid fizo) * P(fizo) \\ &= \; 0.1*0.7 \; + \; 0.9*0.3 \; = \; 0.34, \text{ by the Law of Total Probability.} \end{aligned}$$

(All knowledge to date becomes prior knowledge when calculating the usefulness of subsequent questions; for instance, if features have class-conditional dependencies, obtained answers may change feature probabilities in addition to P(glom) and P(fizo), as the General Discussion considers.) Is the Drink or Hula question more useful in this example? Each sampling norm gives its own means of calculating an individual answer's usefulness. However, all norms discussed in this paper define a question's usefulness as the *expected usefulness* of the individual answers:

$$usefulness(Hula) = P(hulaWorn) * usefulness(hulaWorn) + P(hulaNotWorn) * usefulness(hulaNotWorn)$$

Equations for the hulaWorn answer and the Hula question follow; specific values for the sample questions are given in Table A1.

Diagnosticity

$$diagnosticity(hulaWorn) = \max\left(\frac{P(hulaWorn \mid glom)}{P(hulaWorn \mid fizo)}, \frac{P(hulaWorn \mid fizo)}{P(hulaWorn \mid glom)}\right), \text{ and}$$

$$diagnosticity(Hula) =$$
$$P(hulaWorn) * diagnosticity(hulaWorn) \quad + \quad P(hulaNotWorn) * diagnosticity(hulaNotWorn).$$

Log₁₀ diagnosticity

$$log_{10}\ diagnosticity(hulaWorn) = log_{10}\ \max\left(\frac{P(hulaWorn \mid glom)}{P(hulaWorn \mid fizo)}, \frac{P(hulaWorn \mid fizo)}{P(hulaWorn \mid glom)}\right), \text{ and}$$

$$log_{10}\ diagnosticity(Hula) = \quad P(hulaWorn) * log_{10}\ diagnosticity(hulaWorn)$$
$$+ \quad P(hulaNotWorn) * log_{10}\ diagnosticity(hulaNotWorn).$$

Information gain

The information gain of the answer hulaWorn,

information gain(hulaWorn) = H(Species) – H(Species | hulaWorn), where

$$H(Species) = P(glom) * \log_2 \frac{1}{P(glom)} \quad + \quad P(fizo) * \log_2 \frac{1}{P(fizo)}\text{ , and}$$

$$H(Species \mid hulaWorn) = \quad P(glom \mid hulaWorn) * log_2 \frac{1}{P(glom \mid hulaWorn)}$$
$$+ \quad P(fizo \mid hulaWorn) * log_2 \frac{1}{P(fizo \mid hulaWorn)}.$$

The information gain of the Hula question, the mutual information between Hula and Species,

$I(Hula, Species) = H(Species) - H(Species \mid Hula)$, where the conditional entropy,

H(Species | Hula)= P(hulaWorn)*H(Species | hulaWorn)
+ P(hulaNotWorn)*H(Species | hulaNotWorn).

Kullback Liebler (KL) distance

$$
\begin{aligned}
KL\ distance(hulaWorn) = \quad & P(glom \mid hulaWorn)\ \log_2 \frac{P(glom \mid hulaWorn)}{P(glom)} \\
& + \\
& P(fizo \mid hulaWorn)\ \log_2 \frac{P(fizo \mid hulaWorn)}{P(fizo)}, \text{and}
\end{aligned}
$$

$$
\begin{aligned}
KL\ distance(Hula) = \quad & P(hulaWorn) * KL\ distance(hulaWorn) \\
& + P(hulaNotWorn) * KL\ distance(hulaNotWorn).
\end{aligned}
$$

Probability gain

probabilityGain(hulaWorn) =
max(P(fizo|hulaWorn), P(glom|hulaWorn)) - max(P(glom), P(fizo)).

$probabilityGain(Hula) = P(CorrectGuess \mid Hula) - P(CorrectGuess)$, where

$P(CorrectGuess) = \max(P(glom), P(fizo))$, and

$$
\begin{aligned}
P(CorrectGuess \mid Hula) = \quad & P(hulaWorn) * \max(P(glom \mid hulaWorn), P(fizo \mid hulaWorn)) \\
+ \quad & P(hulaNotWorn) * \max(P(glom \mid hulaNotWorn), P(fizo \mid hulaNotWorn))
\end{aligned}
$$

Impact

From the definition in the text,

impact(hulaWorn) = 0.5 * abs(P(glom | hulaWorn) - P(glom))
+ 0.5 * abs(P(fizo | hulaWorn) - P(fizo)).

Because P(fizo | hulaWorn)=1-P(glom | hulaWorn), and P(fizo)=1-P(glom),

impact(hulaWorn) = abs(P(glom | hulaWorn) - P(glom)).

If a probability model has two categories, the impact of a question is a constant multiple of the difference in feature probabilities (see the footnote in the introduction of impact, in the text):

impact(Hula) = 2*P(glom)*P(fizo)*abs(P(hulaWorn | glom) – P(hulaWorn | fizo)).

[insert Table A1 about here]

Appendix B: limiting cases of disagreement between norms

This appendix gives several limiting cases of disputes between each pair of norms, as obtained in Simulation 2. Each optimization shows results for several prior probabilities, to illustrate the systematic relationship of feature probabilities to the prior P(glom). Each table (B1-B6) gives results for an optimization comparing one pair of norms. Tables are organized such that one norm prefers the "Hula" question, and the other norm prefers the "Drink" question, as noted.

Precise definition of DStr and Preference Strength. Disagreement Strength (*DStr*) is defined to be high when one norm strongly prefers one particular question, and another norm strongly prefers the other question. DStr, therefore, requires a measure of each individual norm's strength of preference between Drink and Hula. The absolute value of the difference in usefulness of Hula and Drink, as measured by each norm, was computed in each of 100,000 random trials, in which P(glom), P(wearsHula | glom), etc., were all random probabilities. For probability gain, cases where both features had probability gain zero, and were therefore tied, were excluded; no other ties occurred. Endpoints to the distributions underlying each norm's strength of preference were added where they exist. In a novel trial, a particular norm's degree of preference is quantified as a percentile of the set of 100,000 previously observed differences in the usefulness of Hula and Drink. Linear interpolation was used so that DStr would be continuously valued. DStr is the geometric mean (the square root of the product) of two norms' strengths of (contrary) preferences.

[insert tables B1-B6 about here]

Appendix C: sample stimulus from Vuma experiment

Imagine you are visiting the planet Vuma. There are 1 million creatures on Vuma; 50 % of them are Gloms, and 50 % are Fizos. All creatures are invisible to the human eye, so you cannot learn about them by looking at them, but only by asking them questions.

Suppose you have just met a creature from Vuma. Your job is to tell which of the two kinds of creatures it is. The table below gives information about the percent of Gloms, and the percent of Fizos, with certain characteristics. (The characteristics are listed in a random order.)

	Characteristic			
	drinks tea	wears a hula hoop	plays harmonica	gurgles a lot
proportion of Gloms	30 %	99 %	1 %	70 %
proportion of Fizos	0.01 %	100 %	99 %	30 %

Imagine that to help you find out the identity of the creature, you could ask it *one* yes or no question, about one of its characteristics. For instance, if "swims fast" were a characteristic, you could ask "Do you swim fast?" The creature answers truthfully.

Considering the information given, which of the possible questions would be most useful to help you learn whether the creature is a Glom or Fizo? Please rank the questions below, putting a "1" in the box beside the most useful question, a "2" in the box beside the next-most-useful question, and so on. If two questions are equally useful, give them the same rank.

Question	*Rank*
Do you drink tea?	
Do you wear a hula hoop?	
Do you play harmonica?	
Do you gurgle a lot?	

Footnotes

[1] *Pseudodiagnosticity* papers (Doherty, Mynatt, Tweney, & Schiavo, 1979; Doherty, Schiavo, Tweney, & Mynatt, 1981; Beyth-Marom & Fischoff, 1983; Doherty, Chadwick, Garavan, Barr, & Mynatt, 1996; Maggi, Butera, Legrenzi, & Mugny, 1998) also use diagnosticity as a sampling norm. In a typical experiment subjects are given d_1 (datum$_1$) and d_2 and are trying to determine whether h_1 or h_2 is the correct hypothesis. They may select two of four possible pieces of information: $P(d_1|h_1)$, $P(d_1|h_2)$, $P(d_2|h_1)$, or $P(d_2|h_2)$. The normatively correct behavior, which makes it possible to calculate posterior probabilities, is to select $P(d_1|h_1)$ and $P(d_1|h_2)$, or $P(d_2|h_1)$ and $P(d_2|h_2)$. Subjects frequently make other selections, which is nonnormative irrespective of the sampling norm used. Unfortunately, these empirical data do not discriminate the relative plausibility of different sampling norms as descriptive models.

[2] Proof of equivalence of heuristic feature difference maximization strategy and impact sampling norm, where there are two categories c_1 and c_0, and a question Q with possible answers q_1 and q_2. If $P(q_1 |c_1)>P(q_1 |c_0)$, then impact(Q)

$= P(q_1)*\text{impact}(q_1) + P(q_2)*\text{impact}(q_2)$

$= P(q_1)*\text{abs}[\ P(c_1|q_1) - P(c_1)\ ] + P(q_2)*\text{abs}[\ P(c_1|q_2) - P(c_1)\ ]$, because there are two c_i,

$= P(q_1)*[\ P(c_1|q_1) - P(c_1)\ ] + P(q_2)*[\ P(c_1) - P(c_1|q_2)\ ]$, because $P(q_1|c_1)>P(q_1|c_0)$,

$= P(q_1)*P(c_1|q_1) - P(c_1)*P(q_1) + P(c_1)*P(q_2) - P(q_2)*P(c_1|q_2)$

$= P(c_1)*P(q_1|c_1) - P(c_1)*P(q_1) + P(c_1)*P(q_2) - P(c_1)*P(q_2|c_1)$

$= P(c_1)*[P(q_1|c_1) - P(q_1) + P(q_2) - P(q_2|c_1)\ ]$

$= P(c_1)*[P(q_1|c_1) - P(q_1) + (1-P(q_1)) - (1-P(q_1|c_1))\ ]$,
because $P(q_1) = 1-P(q_2)$; $P(q_2|c_1) = 1-P(q_1|c_1)$,

$= 2*P(c_1)*[\ P(q_1|c_1) - P(q_1)\ ]$

$= 2*P(c_1)*[P(q_1|c_1) - P(c_1)*P(q_1|c_1) - P(c_0)*P(q_1|c_0)]$, by Law of Total Probability,

$= 2*\mathrm{P}(c_1)*[\ (1 - \mathrm{P}(c_1))*\mathrm{P}(q_1|c_1) - \mathrm{P}(c_0)*\mathrm{P}(q_1|c_0)\ ]$

$= 2*\mathrm{P}(c_1)*[\ \mathrm{P}(c_0)*\mathrm{P}(q_1|c_1) - \mathrm{P}(c_0)*\mathrm{P}(q_1|c_0)\ ]$, because $\mathrm{P}(c_0) = 1 - \mathrm{P}(c_1)$

$= 2*\mathrm{P}(c_1)*\mathrm{P}(c_0)*[\ \mathrm{P}(q_1|c_1) - \mathrm{P}(q_1|c_0)\ ]$

By a similar derivation, if $\mathrm{P}(q_1|c_1)>\mathrm{P}(q_1|c_0)$, then

$\mathrm{impact}(Q) = 2*\mathrm{P}(c_1)*\mathrm{P}(c_0)*[\ \mathrm{P}(q_1|c_0) - \mathrm{P}(q_1|c_1)\ ]$.

In both cases,

$\mathrm{impact}(Q) = 2*\mathrm{P}(c_1)*\mathrm{P}(c_0)*\mathrm{abs}[\ \mathrm{P}(q_1|c_0) - \mathrm{P}(q_1|c_1)\ ]$.

[3] Nick Chater (personal communication, 2005) raised this final issue.

[4] If reducing the number of possible diseases improves treatment, irrespective of the true disease, probability gain might not be uniquely justified. Shanks, Tunney, and McCarthy (2002), on a two arm bandit task that did not involve active sampling, found that some subjects did seek to maximize average winnings. One could imagine a modified task that includes asking questions, where probability gain is uniquely compatible with the goal of maximizing average winnings.

[5] In experiment 5 joint probabilities of each disease and each test, for instance that 32% of patients have disease 1 and a positive result from test A, were given. A set of conditional probabilities uniquely implies a set of joint probabilities, and vice versa.

[6] Some research on the abstract selection task (Oaksford & Chater, 1998, 2003; Hattori, 2002) has reported information gain values scaled to sum to 1. The present paper reports non-normalized values; normalized values give the same ordering.

[7] The present analysis fixed P(dependence hypothesis)=0.5 and P(error)=0.1, like Oaksford and Chater (2003, henceforth "OC2003"). But whereas OC2003 optimized P(A) and P(2) to fit data from each of several dozen experiments in the literature, the present analysis fixed

P(A)=0.22 and P(2)=0.27, to match the mean best fit parameters reported by OC2003. For each experiment, OC2003 used a logistic function to transform the information gain of each card into a probability of selection, as proposed by Hattori (2002). A further step in analysis of the selection task would be to replicate OC2003's optimizations, finding the best fit parameters of P(A) and P(2) for each sampling norm, for each of the several dozen experiments OC2003 fit. Those results, for instance goodness of fit in each experiment, and number of experiments in which a model is rejected, could potentially speak to the relative descriptive plausibility of each sampling norm, even without overt disagreements on the ordering of the cards' usefulness.

[8] In one of Inhelder and Piaget's (1955/1958) studies subjects experimented with string length, weight on the string, etc., to learn what variables control the movement of a pendulum. Scientific reasoning research (Wason, 1960; Klahr, 2000; Kuhn, Amsel, & O'Loughlin, 1988, pp. 161-183; Kuhn, 1989, 2002; Zimmerman, 2000) frequently uses similarly rich tasks. Another related area is contingency, the degree to which one variable predicts another variable's occurrence later in time (Jenkins & Ward, 1965; Allan, 1993; Anderson & Sheu, 1995).

[9] McKenzie and Mikkelsen (in press) did not specify prior probabilities of h_1 and $h_{0.}$ The diagnosticity and log diagnosticity of an observation (answer q_j) is independent of priors; e.g. $\log_2$ diagnosticity(A)=2.46, irrespective of whether P(h_1)=P(h_2)=0.5, or P(h_1)=0.9999 and P(h_2)=0.0001. As the other norms are sensitive to priors, equal priors of h_1 and h_0 were used.

[10] Methods and results are briefly described here; complete details on the design of the simulations and results are included in the supplementary material posted online, and at http://www.jonathandnelson.com.

[11] Strict formalists may state that diagnosticity(Drink) is undefined. But clearly, $\lim_{P(drink|fizo)\to 0^{+}} diagnosticity(Drink) = \infty$. This difficulty stems from the definition of diagnosticity

and log diagnosticity. Features, like P(hasDNA | human), can have probability 0 or 1.

[12] A pilot experiment with unequal prior probabilities of glom and fizo was conducted. Several subjects indicated that they had forgotten that priors were not equal when rating the questions. Because of this possible confound, and because of the default assumption that priors are equal (Fox & Rottenstreich, 2003), equal prior probabilities were used. Impact and probability gain are identical in this two category, equal prior probability case.

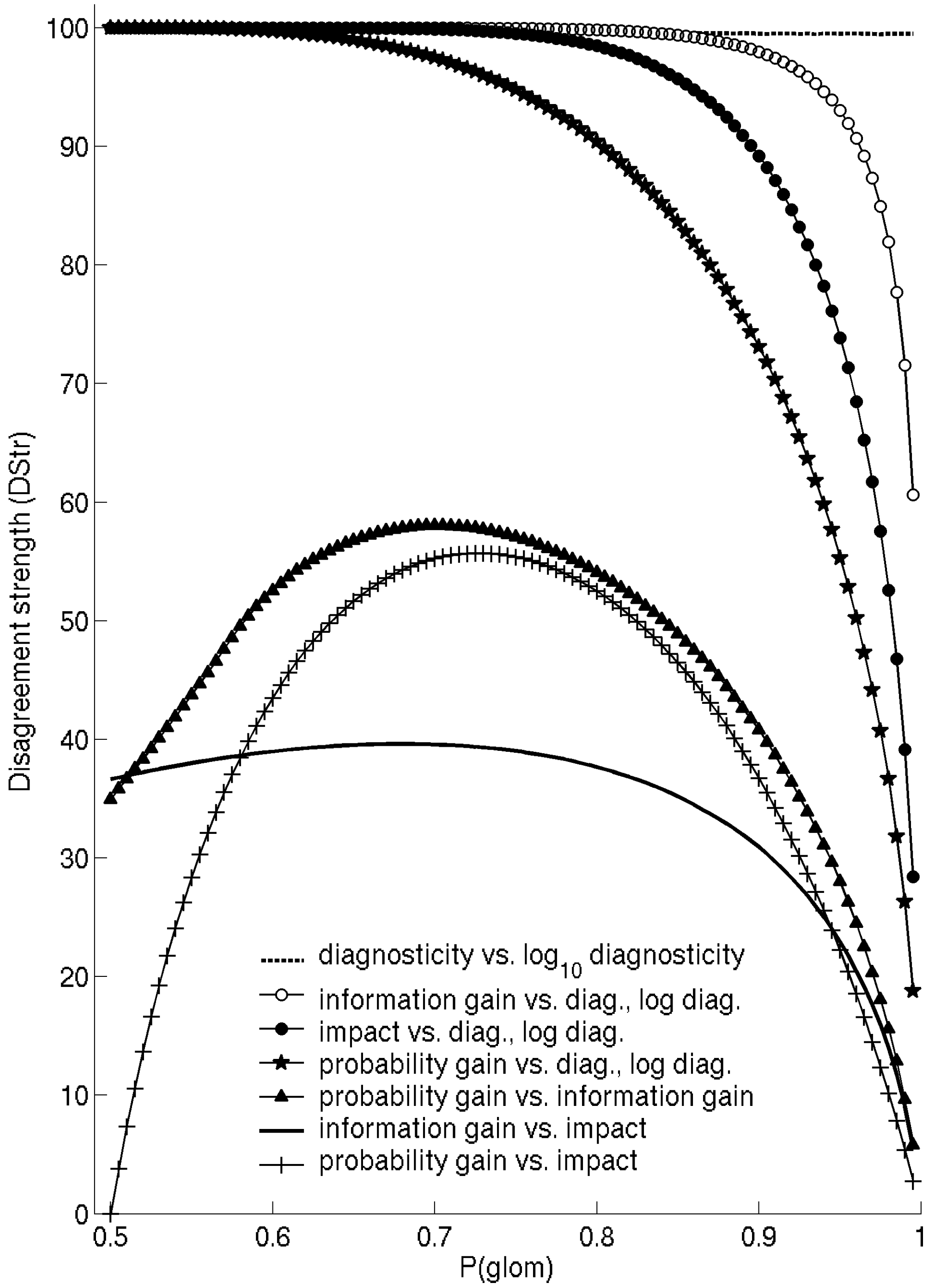
[13] A further note is that reinforcement learning researchers (Kearns & Singh, 2002, p. 211) have found that learning an unknown probability model within a finite time is not possible if rewards can be infinite.

Figure caption

Figure 1

Simulation 2 results. Strength of limiting cases of disagreement between pairs of sampling norms, for different prior probabilities

[Note: please use the included REV.Nelson105.fig1.eps file for figure 1. This tiff file is embedded here as a reference only.]



Tables and captions

Table 1

Sampling norms and probabilistic information-gathering tasks

Sampling Norm	Task	# Hyps.	Reference	Paper Type
Diagnosticity or log diagnosticity	Covariation assessment	2	McKenzie & Mikkelsen (in press)	Review, experimemt
	Medical diagnosis	2	Good & Card (1971) Card & Good (1974)	Theory
	Hypothesis testing	2	Klayman & Ha (1987)	Theory
	Introvert vs. extravert	2	Trope & Bassok (1982, 1983) Bassok & Trope (1983-1984)	Experiment
	Planet Vuma	2	Skov & Sherman (1986); Slowiaczek, Klayman, Sherman, & Skov (1992)	Experiment
	Selection task, Causal conditionals	2	Evans & Over (1996); Over & Jessop (1998)	Theory
Information gain, KL distance	Selection task	2	Oaksford & Chater (1994, 2003)	Review, theory
		2	Oaksford, Chater & Grainger (1999)	Experiment
		2, 3	Hattori (2002)	Theory
	Reduced array selection task	2	Oaksford, Chater, Grainger, & Larkin (1997); Oaksford & Chater (1998)	Experiment
	2-4-6 task	2	Ginzburg & Sejnowski (1996)	Experiment
	Number concept task	millions	Nelson, Tenenbaum, & Movellan (2001)	Experiment
	Alien mind-reading	2, 18	Steyvers, Tenenbaum, Wagenmakers, & Blum (2003)	Experiment
	Hypothesis testing	2	Klayman (1987)	Theory
	Selection task, Causal conditionals	2	Over & Jessop (1998)	Theory
Probability gain	2-4-6 task	3	Baron (1985)	Theory
	Urns and poker chips	5	Baron (1985)	Theory
	Medical diagnosis	3	Baron, Beattie, & Hershey (1988)	Experiment
Impact (absolute change)	Hypothesis testing	2	Klayman & Ha (1987)	Theory
	Selection task, Hempel's paradox	2	Nickerson (1996)	Theory

Table 2

Properties of several sampling norms

Property	Diag.	Log d.	Info. gain	KL dist.	Prob. gain	Impact
Usefulness(q_j) finite			Yes	Yes	Yes	Yes
Usefulness(q_j) ≥ 0	Yes	Yes		Yes		Yes
Usefulness(q_j&q_k)= Usefulness(q_j)+Usefulness(q_k)			Yes		Yes	

Note. Usefulness(q_j) denotes the usefulness (utility) of the answer q_j. Blank cells indicate absence of the corresponding property.

Table 3

Features used by Skov and Sherman (1986). Slowiaczek et al. (1992) used features C-F and M-P

Usefulness	Feature	% of gloms, fizos with each feature	Info. gain-KL dist.	Prob. gain, impact	Diagn.	Log_{10} diag.
Low	A, B	48, 52; 52, 48	0.001	0.020	1.083	0.035
	C, D, E, F	28, 32; 32, 28; 68, 72; 72, 68	0.001	0.020	1.084	0.035
Medium	G, H, I, J	15, 45; 45, 15; 55, 85; 85, 55	0.080	0.150	1.982	0.276
	K, L	34, 66; 66, 34	0.075	0.160	1.941	0.288
High	M, N, O, P	10, 50; 50, 10; 50, 90; 90, 50	0.147	0.200	2.760	0.388
	Q, R	26, 74; 74, 26	0.173	0.240	2.846	0.454

Note. Multiple features, for example the features A and B, appear on the same line if all sampling norms agree that those features are equally useful. Semicolons separate features. For example: 48% of gloms and 52% of fizos have feature A; 52% of gloms and 48% of fizos have feature B.

Table 4

Features used by Slowiaczek et al. (1992, experiment 3b), and each feature's usefulness

	% of gloms, fizos with each feature	Info. gain-KL dist.	Prob. gain, impact	Diagn.	Log_{10} diag.
extreme features	90, 55; 45, 10	0.11766	0.175	2.4239	0.3347
non-extreme features	65, 30; 30, 65	0.09052	0.175	2.0792	0.31754

Note. Semicolons separate features. For example, 95% of gloms and 55% of fizos have one of the extreme features; 45% of gloms and 10% of fizos have the other extreme feature. Both Planet Vuma and medical diagnosis scenarios were used.

Table 5

Re-analysis of experiment 4 in Baron et al. (1988)

Test:	1	2	3	4	5	6	7	8	9	10
P(positive | disease):										
Disease A	0.75	0	0.75	0.50	0.50	0.50	0.50	0.25	1.00	0.50
Disease B	0.75	1.00	1.00	1.00	0	0	1.00	0	0	0.50
Disease C	0.75	0	1.00	0	0	1.00	1.00	0	0	0.50
Subjects' ratings	21	61	40	34	26	26	48	25	75	9
Usefulness:										
Probability gain	0	0.200	0	0	0	0	0	0	0.240	0
Information gain	0	0.795	0.115	0.350	0.264	0.350	0.264	0.115	0.943	0
Impact	0	0.243	0.077	0.141	0.154	0.141	0.154	0.077	0.307	0

Note. Prior probabilities of diseases A, B, and C, were 0.64, 0.24, and 0.12, respectively. Subjects' ratings, and the conditional probabilities of a positive test given each disease, are adapted from Table 4 in "Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty," by J. Baron, J. Beattie, and J. C. Hershey, 1988, *Organizational Behavior and Human Decision Processes, 42(1)*, p. 102. Copyright Elsevier; adapted with permission.

Table 6

Re-analysis of experiments 5 and 6 in Baron et al. (1988)

Test:	1	2	3	4	5	6	7	8	9	10	11
P(positive | disease):											
Disease A	0.50	1.00	0.81	0	1.00	0	0	1.00	0	1.00	1.00
Disease B	0.50	1.00	0	1.00	0.50	1.00	0.50	0	0	0	1.00
Disease C	0.50	0	0	0	0	1.00	0	1.00	1.00	0	1.00
Subjects' ratings:											
Expt. 5	-	42	56	64	41	69	44	65	42	64	-
Expt. 6	-	64	62	75	52	75	41	69	56	79	-
Usefulness:											
Probability gain	0	0.120	0.118	0.240	0.120	0.240	0.120	0.240	0.120	0.240	0
Information gain	0	0.529	0.550	0.795	0.555	0.942	0.289	0.795	0.529	0.943	0
Impact	0	0.141	0.245	0.243	0.205	0.307	0.122	0.243	0.141	0.307	0

Note. Prior probabilities of diseases A, B, and C, were 0.64, 0.24, and 0.12, respectively. Baron et al. stated that most subjects rated tests 1 and 11 zero, but did not report subjects' ratings of those tests. Subjects' ratings, and the conditional probabilities of a positive test given each disease, are adapted from Table 5 in "Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty," by J. Baron, J. Beattie, and J. C. Hershey, 1988, *Organizational Behavior and Human Decision Processes, 42(1)*, p. 106. Copyright Elsevier; adapted with permission.

Table 7

Correlation between three sampling norms and subjects' average ratings in Baron et al.'s (1988) study

	Experiment		
	4	5	6
Probability gain	0.84	0.91	0.83
Information gain	0.89	0.86	0.94
Impact	0.89	0.88	0.81

Note. Correlations exclude tests 1 and 11, in experiments 5 and 6, for which Baron et al. (1988) did not report subjects' ratings. Correlations use Pearson's *r*. A similar pattern, with slightly lower correlation values, results if Spearman's rank correlation is used.

Table 8

Cards' usefulness on the selection task

	A (*P*)	2 (*Q*)	3 (*not-Q*)	K (*not-P*)
Information gain-KL distance	0.324	0.200	0.066	0.040
Probability gain-impact	0.315	0.257	0.095	0.089
Diagnosticity	4.980	3.120	2.001	1.548
Log_{10} diagnosticity	0.664	0.493	0.191	0.162

Note. Each card's usefulness was calculated with respect to Oaksford and Chater's (2003) belief model, with P(dependence hypothesis)=0.5, P(error)=0.1, P(A)=0.22, and P(2)=0.27.

Table 9

Usefulness of each cell, relative to McKenzie and Mikkelsen's (in press) belief model

Observation	Info. gain, KL dist.	Prob. Gain, Impact	Diagnosticity	Log_{10} diag.	Log_{2} diag.
A	0.38	0.35	5.50	0.74	2.46
B	0.08	0.17	2.00	0.30	1.00
C	0.08	0.17	2.00	0.30	1.00
D	0.0005	0.01	1.06	0.02	0.08

Note. McKenzie and Mikkelsen's model describes answers' (q_j), rather than questions' (Q), usefulness, illustrating the variety of tasks in which explicit sampling norms may be calculated. Particular answers' information gain and KL distance are not necessarily the same, although they are in this case.

Table 10

Feature probabilities, and usefulness of each question, as calculated by each norm

	Drink	Gurgle	Harmonica	Hula
Proportion of Gloms	30%	70%	1%	99%
Proportion of Fizos	0.01%	30%	99%	100%
Diagnosticity	451.36	2.33	99.00	infinite
Log_{10} diagnosticity	0.65	0.37	2.00	infinite
Information gain-KL distance	0.17	0.12	0.92	0.01
Probability gain-impact	0.15	0.20	0.49	0.01

Table 11

Experiment results

%	(#)	Pattern	Description	Pattern correlation with:			
of SS				Info. gain-KL dist.	Prob. gain-impact	Diag.	Log diag.
32	(48)	Harmonica > Drink > Gurgle > Hula	Info. gain-KL dist.	1.00	0.80	-0.40	-0.20
27	(40)	Harmonica > Gurgle > Drink > Hula	Prob. gain-impact	0.80	1.00	-0.80	-0.40
18	(26)	Drink > Harmonica > Gurgle > Hula	Close to info. gain	0.80	0.40	-0.20	-0.40
5	(8)	Harmonica > Drink = Gurgle > Hula	Avg. info., prob. gain	0.95	0.95	-0.63	-0.32
18	(26)	Other patterns, 1-3 responses each	(Various patterns)	0.26	0.20	0.04	0.21
0	(0)	Hula > Drink > Harmonica > Gurgle	Diagnosticity	-0.40	-0.80	1.00	0.80
0	(0)	Hula > Harmonica > Drink > Gurgle	Log diagnosticity	-0.20	-0.40	0.80	1.00
100	(148)	Aggregate results	(Various patterns)	0.78	0.69	-0.41	-0.22

Note. Spearman's rank correlation coefficient, with correction for ties, was used. Mean correlations are given for rows that aggregate multiple patterns of response rankings.

Table 12

Natural environment feature distribution and usefulness values

	Skirt (or dress)		Glasses			Beard		Earrings		Hair	
Feature values and distribution:	no	yes	none	sun-	yes	yes	no	yes	no	short	long
% of males	100	0	67	6	27	16	84	2	98	93	7
% of females	98	2	83	3	14	0	100	47	53	7	93
Usefulness of each feature:											
Information gain-KL distance	0.010		0.025			0.084		0.235		0.634	
Probability gain	0.010		0.065			0.062		0.220		0.420	
Impact	0.010		0.080			0.080		0.225		0.430	
Diagnosticity	infinite		1.412			infinite		7.056		13.296	
Log_{10} diagnosticity	infinite		0.093			infinite		0.532		1.123	

Note. About 51% of individuals were male. When hair completely obscured the ears, the ears were classified as not having earrings.

Table A1

Example questions' and answers' usefulness, as calculated by each sampling norm

	Diagn.	Log_{10} diag.	Info. gain	KL dist.	Prob. gain	Impact
Hula	9.000	0.954	0.456	0.456	0.200	0.336
hulaWorn (p=0.34)	9.000	0.954	0.148	0.752	0.094	0.494
hulaNotWorn (p=0.66)	9.000	0.954	0.615	0.303	0.255	0.255
Drink	1.496	0.173	0.026	0.026	0.000	0.084
drinksTea (p=0.36)	1.667	0.222	-0.099	0.044	-0.117	0.117
doesn'tDrink (p=0.64)	1.400	0.146	0.096	0.016	0.066	0.066

Note. P(glom)=0.70, P(wearsHula | glom)=0.1, P(wearsHula | fizo)=0.9, P(drinksTea | glom)=0.3, and P(drinksTea | fizo)=0.5. Individual answers can have negative information gain or probability gain, as drinksTea illustrates. Questions' usefulness is nonnegative irrespective of which sampling norm is used. The Drink question has probability gain zero because it does not improve probability of correct guess. Information gain and KL distance are equivalent when evaluating questions, but not individual answers, as shown.

Table B1

Log diagnosticity (prefers Hula) vs. diagnosticity (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm's preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50%	99.90	0.10	53.66	0.01	99.66	-99.70	99.62	99.16	87.06	88.50
60	99.90	0.10	0.01	62.76	99.68	-99.78	99.58	97.45	69.45	79.14
70	99.87	0.13	0.01	62.03	99.68	-99.65	99.70	95.87	58.12	74.63
80	0.17	99.82	0.01	76.77	99.61	-99.80	99.41	82.60	27.90	45.30
90	99.85	0.21	29.73	0.01	99.51	-99.56	99.47	93.94	52.86	65.41
99.5	99.86	0.16	68.27	99.99	99.50	-99.64	99.33	33.17	2.62	5.58

Table B2

Information gain (prefers Hula) vs. diagnosticity (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm’s preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50%	99.99	0.01	100.00	99.99	100.00	-100.00	-100.00	100.00	100.00	100.00
60	99.99	0.01	0.01	0.00	100.00	-100.00	-100.00	100.00	99.43	100.00
70	0.01	99.99	0.01	0.00	99.99	-100.00	-100.00	99.98	95.02	99.77
80	99.99	0.01	99.99	100.00	99.83	-100.00	-100.00	99.67	81.74	96.98
90	0.01	99.99	0.01	0.00	97.98	-100.00	-100.00	96.01	53.42	79.56
99.5	99.99	0.01	99.99	100.00	60.64	-100.00	-100.00	36.77	3.53	8.08

Note. Results for impact vs. diagnosticity, information gain vs. log diagnosticity, and impact vs. log diagnosticity are not presented separately, because each of those optimizations produced feature probabilities that were virtually identical to the results for information gain vs. diagnosticity. DStr values for information gain vs. log diagnosticity were essentially identical to those for information gain vs. diagnosticity. DStr values for impact vs. diagnosticity, and impact vs. log diagnosticity, follow a qualitatively similar pattern; those values can be approximated from the strengths of each norm’s preference in Table B2.

Table B3

Probability gain (prefers Hula) vs. diagnosticity (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm's preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50%	0.01	99.99	0.00	0.01	100.00	-100.00	-100.00	100.00	100.00	100.00
60	0.01	99.99	86.87	100.00	99.72	-100.00	-100.00	99.99	99.43	99.75
70	0.01	99.99	9.14	0.00	97.48	-100.00	-100.00	99.95	95.02	99.19
80	0.01	99.99	49.48	100.00	90.41	-100.00	-100.00	97.86	81.74	74.60
90	0.01	99.99	87.77	100.00	73.09	-100.00	-100.00	95.42	53.42	74.51
99.5	0.01	99.99	16.43	0.00	18.78	-100.00	-100.00	36.15	3.53	6.81

Note. Results for probability gain vs. log diagnosticity are not given separately because they were indistinguishable from results for probability gain vs. diagnosticity.

Table B4

Information gain (prefers Hula) vs. probability gain (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm's preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50%	0.00	31.44	28.91	71.09	34.94	100.00	100.00	37.42	-32.61	-35.75
60	66.67	100.00	87.51	56.42	52.55	100.00	100.00	51.87	-53.24	8.71
70	57.14	0.00	95.25	55.87	58.04	100.00	100.00	63.38	-53.15	45.89
80	75.00	0.00	1.58	46.12	54.07	100.00	100.00	64.94	-45.02	55.54
90	11.11	100.00	99.69	51.39	40.80	100.00	100.00	58.39	-28.50	45.21
99.5	0.50	100.00	0.00	50.21	5.81	0.00	0.00	18.27	-1.82	4.07

Table B5

Information gain (prefers Hula) vs. impact (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm’s preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50%	67.72	100.00	28.80	71.26	36.64	100.00	100.00	39.16	-31.13	-34.28
60	100.00	65.60	71.53	25.70	38.98	100.00	100.00	41.76	7.81	-36.39
70	100.00	62.47	27.70	77.91	39.57	100.00	100.00	44.08	41.96	-35.52
80	100.00	58.88	73.13	17.29	37.66	100.00	100.00	44.03	46.14	-32.21
90	100.00	53.76	75.23	11.06	30.92	100.00	100.00	40.42	28.70	-23.65
99.5	0.00	54.84	15.95	98.86	5.59	100.00	100.00	13.30	1.98	-2.35

Table B6

Impact (prefers Hula) vs. probability gain (prefers Drink)

P(glom)	wearsHula:		drinksTea:		DStr	Strength of each norm's preference:				
	% gloms	% fizos	% gloms	% fizos		Diagn.	Log diag.	Info. gain	Prob. gain	Impact
50.5%	1.98	0.00	0.07	1.19	3.78	100.00	100.00	8.81	-3.97	3.60
60	33.33	0.00	0.00	17.27	43.49	0.00	0.00	49.13	-40.20	47.03
70	57.14	0.00	100.00	69.66	55.25	0.00	0.00	59.21	-49.93	61.12
80	25.00	100.00	0.00	40.66	52.48	0.00	0.00	59.99	-45.77	60.18
90	88.89	0.00	0.00	47.59	36.71	0.00	0.00	55.25	-29.44	45.78
99.5	0.50	100.00	0.00	45.82	2.75	0.00	0.00	19.70	-1.68	4.44

Note. Impact and probability gain are identical when there are two equiprobable hypotheses. There is therefore no case of disagreement between those norms in the Vuma scenario when P(glom)=0.50. Cases of slight disagreement are observed when P(glom)=0.505, included here.