Chapter 7

Towards a rational theory of human information acquisition

to appear in M. Oaksford and N. Chater (Eds.),
The probabilistic mind: prospects for rational models of cognition

Jonathan D. Nelson

jnelson@salk.edu

http://www.jonathandnelson.com/

Computational Neurobiology Lab, Salk Institute for Neural Computation and Cognitive Science Department, University of California, San Diego, USA

Introduction

Many situations require judiciously choosing particular pieces of information to acquire before otherwise acting. Consider the case of a scientist studying gravity. Dropping an apple from a bridge, to see whether it will fall, is a possible experiment. But it does not appear to be a useful experiment, perhaps because all plausible theories predict the apple will fall. In a Bayesian optimal experimental design (OED) framework, simply testing predictions of various theories does not constitute good science. Rather, possible experiments are judged useful to the extent that plausible competing theories make contradictory predictions about the results.

Beginning roughly 1980, psychologists began to use OED theories of the value of information as normative (theoretically optimal) or descriptive models of human queries and intuitions about the value of information (Table 7.10). Examples of scenarios studied include people's intuitions about what medical tests are most useful (Baron *et al.*, 1988), about what features would be most helpful to categorize objects (Skov & Sherman, 1986; Slowiaczek *et al.*, 1992), and about what cards are most useful on Wason's selection task (Klayman & Ha, 1987; Oaksford & Chater, 1994); human eye movements for learning an object's shape (Walker-Renninger *et al.*, 2005), for finding an object hidden in noise (Najemnik & Geisler, 2005), or for categorization (Nelson & Cottrell, 2007); and monkeys' eye movements to query the location of a hidden target (Nakamura, 2006).

In life, of course, information gathering is frequently also coupled with other goals (Baron *et al.* 1988; Box & Hill, 1967; Chater *et al.* 1998; Chater & Oaksford, 1999; Lindley, 1956). For instance, a foraging animal may wish to learn the geographical

distribution of a food source, while simultaneously maximizing immediate energy intake. Similarly, a physician diagnosing an illness would want to balance the expected informational value of diagnostic tests with their monetary cost and possible harm to the patient. (An expensive test that deposits large quantities of radiation on the patient, such as a CT scan, should be avoided if an inexpensive blood test provides equivalent information.) It would be inappropriate to build mathematical models of these situations based solely on information gathering; the other goals must also be incorporated. There is evidence that in some situations people can optimize fairly arbitrary utility functions (Trommershäuser *et al.*, 2003) and that in other situations, people have difficulty doing so (Baron & Hershey, 1988). In this chapter, however, our focus is on situations in which information gathering is the only goal. A unified theory of information gathering might be possible in these situations, depending on the consistency of intuitions across people and across tasks. Whether or not a unified theory is plausible will be discussed further on.

The issue of how to assess potential experiments' usefulness is complicated by the fact that an experiment's outcome is not known until the experiment is conducted. But in some cases, people have strong intuitions about which experiments (or questions, or queries) are most useful. The planet Vuma scenario (Fig. 7.1), introduced by Skov and Sherman (1986), illustrates this. The experimental subject imagines that they are on the planet Vuma, which is inhabited by two species of creatures, gloms and fizos. Because of the blinding light and blowing sand, the creatures are effectively invisible. The task is to categorize a randomly chosen creature as a glom or a fizo, by asking either whether it wears a hula hoop or whether it drinks iced tea. As in scientific inference, the hypotheses of interest (species of creature) cannot be queried directly, but probabilistic predictions of the hypotheses (presence or absence of features) can be tested. Suppose that P(*glom*) = P(*fizo*) = 0.50, P(*hula* | *glom*) = 0.90, P(*hula* | *fizo*) = 0.50, P(*drink* | *glom*) = 0.28, and P(*drink* | *fizo*) = 0.32. Which feature would be the most useful to query, based on the probabilities involved?

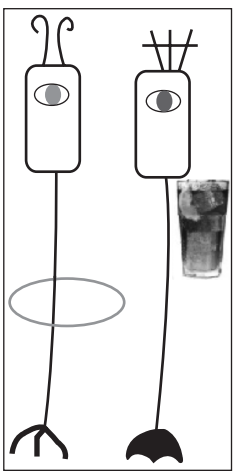

Fig. 7.1. Planet Vuma.

Most everyone thinks the hula feature is more useful. Why? Note that its usefulness depends on whether the feature is present or absent. Absence of the hula feature provides strong evidence for fizo: using Bayes' (1763) theorem, we can calculate that P(*fizo* | ~*hula*) = 0.88. Presence of the hula feature provides moderate evidence for glom: P(*glom* | *hula*) = 0.64. To assess the usefulness of the hula question, however, we need precise methods to:

(1) quantify the usefulness of each possible *answer* (presence or absence of the feature), and

(2) weight each possible answer's usefulness, to quantify the usefulness of the *question*.

Several mathematical ideas for doing this have been proposed, in a Bayesian optimal experimental design (OED) framework. (Optimal data selection, a term used by Oaksford & Chater, 1994, 2003, and optimal experimental design are equivalent terms.) Note that environmental probabilities, and posterior probabilities of fizo or glom, given presence or absence of a feature, do not depend on what feature is queried. Suppose, for sake of illustration, that we were to use probability gain (Baron 1981, as cited in Baron, 1985, pp. 130–167) to calculate the usefulness of obtained answers. Probability gain defines an answer's usefulness as the extent to which the answer improves the probability of correct guess (or reduces probability of error). If probability gain is used to calculate the utility (usefulness) of each answer (query result), absence of the hula hoop is worth 0.33 utility units, and presence of the hula hoop is worth 0.14 utility units. But how do we calculate the usefulness of the hula *question*? We could average the usefulness of the two answers, e.g. (0.33 + 0.14)/2 = 0.24. However, the hula hoop is present 70% of the time, so it would not make sense to weight presence and absence equally. Savage (1954, chap. 6) suggested weighting the usefulness (utility) of each answer according to its probability of occurrence. If we use probability gain to measure utility, then the expected utility of the hula question:

$$\begin{aligned} \mathrm{eu}(Hula) &= \mathrm{P}(hula)\ \mathrm{u}(hula) + \mathrm{P}(\sim hula)\ \mathrm{u}(\sim hula) \\ &= 0.70\ \mathrm{u}(hula) + 0.30\ \mathrm{u}(\sim hula) \\ &= 0.20 \text{ utility units.} \end{aligned}$$

Note that here and throughout the chapter, we use capitals for random variables, such as the hula question (*Hula*), and lowercase for specific values taken by random variables, such as the answers that a hula hoop is worn (*hula*) or is not worn (~*hula*).

We can think of calculating the expected utility of the question as a balancing game (Fig. 7.2), in which we represent particular answers' utilities with weights on the right side of a fulcrum, and find what weight will balance the fulcrum on the left side. Balancing the fulcrum, as described below, is mathematically equivalent to calculating the expected usefulness of a question's possible answers. The more frequent an answer, the farther from the center of the fulcrum. Suppose we let 0.01 utility unit = 1 gram, and 1% probability of occurrence = 1 cm. We would then weight the hula absent answer at 33 grams, and put it 30 cm right of the fulcrum; and the hula present answer at 14 grams, and put it 70 cm right of the fulcrum. To find out what the hula

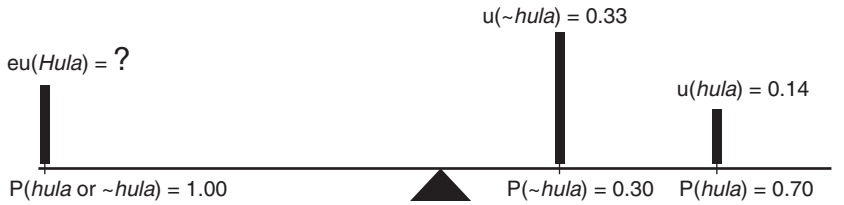

Fig. 7.2. Fulcrum diagram to illustrate expected a question's expected utility (at left) as a function of the utility and probability of its possible answers (at right).

question is worth, we check how much weight needs to be placed 100 cm to the left of the fulcrum to balance it: in this case, 20 grams. The *expected utility* of the hula hoop question is therefore 0.20 utility units, if probability gain is used to measure utility. (We could similarly calculate that the drink question is worth only 0.02 utility units.)

We would like to know whether human information acquisition is optimal (rational, in the sense of Anderson, 1990). But, as mentioned above, in this chapter we are dealing with situations where no extrinsic utility (cost) function is given. Several different optimal experimental design (OED) theories of the value of information have been proposed, in most cases by statisticians considering how to calculate the usefulness of possible experiments. It would therefore be unreasonable to arbitrarily pick a particular measure of the value of information, against which to test people's intuitions. However, it would be remarkable if people's intuitions approximate any of these theories. But the data will push us to conclude something even stronger. We will find (1) that among the OED models, some are better motivated than others as normative models of the value of information, and (2) that among the OED models, the better-motivated models also provide better description of human intuition and behavior!

Optimal experimental design theories of the usefulness of information

Each theory that we will consider follows Savage's (1954, chap. 6) suggestion to define a question Q's usefulness as the expected usefulness, given current knowledge, of the possible answers q_j:

$$\mathrm{eu}(Q)=\sum_{q_j}\mathrm{P}(q_j)\mathrm{u}(q_j)$$

The six OED theories of the value (utility or usefulness) of information that we will consider are Bayesian diagnosticity, log diagnosticity, information gain (uncertainty reduction), Kullback-Liebler (KL) distance, impact (absolute difference), and probability gain (error reduction). Table 7.1 gives how each theory calculates the usefulness of a particular answer q. Individual c_i are particular categories, on a categorization task. (Mathematically, we will treat possible categories c_i as hypotheses to test.) The inquirer's (or learner's) goal in each case is to identify which category (or hypothesis)

Table 7.1. OED theories for the value of information on categorization tasks

OED theory	usefulness of obtained answer q, u(q)
probability gain	$\max_{c_i} P(c_i \mid q) - \max_{c_i} P(c_i)$
information gain	$\sum_{c_i} P(c_i)\log\frac{1}{P(c_i)} - \sum_{c_i} P(c_i \mid q)\log\frac{1}{P(c_i \mid q)}$
KL distance	$\sum_{c_i} P(c_i \mid q)\log\frac{P(c_i \mid q)}{P(c_i)}$
impact	$0.5\sum_{c_i} abs(P(c_i \mid q) - P(c_i))$
Bayesian diagnosticity	$\max\left(\frac{P(q \mid c_1)}{P(q \mid c_2)}, \frac{P(q \mid c_2)}{P(q \mid c_{1.})}\right)$
log diagnosticity	$\log\max\left(\frac{P(q \mid c_1)}{P(q \mid c_2)}, \frac{P(q \mid c_2)}{P(q \mid c_{1.})}\right)$

is correct. Nelson (2005, pp. 996–997) gives example calculations of each theory. Here are some very brief highlights of each OED theory:

— Probability gain (Baron, 1981, as cited in Baron, 1985) quantifies the value of an answer as the extent to which that answer improves probability of correct guess of the true category. Maximizing probability gain corresponds to maximizing probability correct, as well as to minimizing probability of error. In this chapter we assume an optimal response strategy (rather than probability matching or other suboptimal strategies) when calculating probability gain.

— Information gain (Lindley, 1956; Box & Hill, 1967; Fedorov, 1972) is based on the idea that an answer that reduces uncertainty (Shannon, 1948, entropy) about the true category is useful. Oaksford and Chater (1994, 2003) used expected information gain (mutual information) in their probabilistic models of Wason's selection task as optimal data selection.

— Kullback-Liebler (KL) distance defines an answer's usefulness as the distance from prior to posterior beliefs about the categories (Kullback & Liebler, 1951; Cover & Thomas, 1991). The expected KL distance of a query and the expected information gain of a query are identical (Oaksford & Chater, 1996), although KL distance and information gain sometimes give different ratings of the usefulness of individual answers.

— Impact measures the usefulness of information as the absolute change from prior to posterior beliefs. Impact appears to have been independently proposed on three

separate occasions (by Wells & Lindsay, 1980; Klayman & Ha, 1987, pp. 219–220; Nickerson, 1996). It was initially defined for situations with two categories; Nelson (2005) generalized it to situations with two or more discrete categories. It has been slightly reformulated here, in a way functionally equivalent to Nelson (2005), but so that the presence or absence of very-low-probability categories does not change its scale. Impact and probability gain are equivalent if prior probabilities of the categories are equal.

— Bayesian diagnosticity (Good, 1950, 1975, 1983) and log diagnosticity are based on the likelihood ratio of the answer obtained. (Note that despite their names, these measures are in no way more Bayesian than any others; nor does being Bayesian compel the use of these or other measures.) The idea is that answers are useful if they are more probable given one category than given the other category. However, both of these diagnosticity measures have serious flaws (Nelson, 2005). One flaw is that Bayesian diagnosticity and log diagnosticity are unbounded, and consider a question to be infinitely useful if there is any probability that its answer will rule-out one hypothesis, even if the probability of that answer is very low.[1] Another is that in some cases they are independent of priors. Another limitation is that they are only defined for cases with exactly two categories, c_1 and c_2, whereas the other theories are defined for two or more categories c_i. In some cases, both diagnosticity measures make strange claims; in others, only one of the measures does. For instance, Nelson describes how log diagnosticity has bizarre nonmonotonicities in some situations in which diagnosticity is monotonic.

What does each OED theory of the usefulness of information say about our example scenario? As we would hope, these six OED theories agree that the hula question is far more useful than the drink question (Table 7.2).

In the remainder of this chapter we will consider a number of issues important for a rational theory of human information acquisition. How do various OED theories differ from each other? Which are best motivated? Which best describe human intuitions? Do people follow suboptimal heuristic strategies when picking queries? Do monkeys, people, and other animals have similar intuitions about the value of information? Finally, can OED principles be used to design the most informative experiments for addressing these issues?

Nonnegativity, additivity, and sequential questions

Before making statements about whether human information acquisition is rational, it would be helpful to know what theoretical OED models best capture the value of information. This is difficult in the situations we are considering, because there

[1] Suppose there were a gender discrimination task, in which you could choose one feature to look at, to tell whether a person is male or female. Suppose further that one man in the world had a distinctive tattoo on his left arm. The diagnosticity measures consider looking for that tattoo to be an infinitely useful query.

Table 7.2. Usefulness of particular answers, and expected usefulness of questions, in example scenario.

OED theory	eu(***Hula***)	u(***hula***)	u(***~hula***)	eu(***Drink***)	u(***drink***)	u(***~drink***)
probability gain	0.2000	0.1429	0.3333	0.0200	0.0333	0.0143
information gain (bits)	0.1468	0.0597	0.3500	0.0014	0.0032	0.0006
KL distance (bits)	0.1468	0.0597	0.3500	0.0014	0.0032	0.0006
impact	0.2000	0.1429	0.3333	0.0200	0.0333	0.0143
Bayesian diagnosticity	2.7600	1.8000	5.0000	1.0840	1.1429	1.0588
$\log_{10}$ diagnosticity	0.3884	0.2553	0.6990	0.0348	0.0580	0.0248

Note. In this example, P(*glom*) = P(*fizo*) = 0.50, P(*hula* | *glom*) = 0.90, P(*hula* | *fizo*) = 0.50, P(*drinks* | *glom*) = 0.28, and P(*drinks* | *fizo*) = 0.32. Therefore, P(*hula*) = 0.70, and P(~*hula*) = 0.30. The expected utility of the hula hoop question is given by eu(*Hula*). The utility of the answers that the hula hoop is (or is not) worn are given by u(*hula*) and u(~*hula*), respectively.

are no (at least no obvious) unique externally-imposed utility functions. Despite some statements in the literature (Good, 1975, pp. 52–53, said Bayesian diagnosticity 'was central to my first book and occurred also in at least 32 other publications…. What I say thirty-three times is true'), we will refrain from calling any particular utility optimal by definition. Two people could both be perfectly calibrated to the same environment, completely Bayesian in how they update their beliefs as new information is obtained, and yet choose to use different utility functions to evaluate potential queries' usefulness.

One issue in the literature is whether it is better to have a measure of information that is always positive if beliefs change (nonnegativity), such as impact or KL distance; or a measure of information in which two questions' sum utility, if the questions are answered sequentially, is the same as the utility if both questions are answered at once (additivity), such as probability gain or information gain. As a very rough analogy, we can think of current beliefs as a location on a journey to the desired destination, where a single category will have probability one, and the other categories will have probability zero. In this analogy, the various OED models of the utility of information are different metrics of distance traveled (the nonnegative measures), or of the distance remaining toward the destination (the additive measures). Distance traveled may or may not correspond to getting closer to the destination (as lost travelers can attest). It is hopefully intuitive that one could not have a measure that is guaranteed to be positive if beliefs change (e.g. an odometer), and also guaranteed to be additive (because driving in circles is possible, and beliefs sometimes do fluctuate in ways that do not bring the goal closer).

To make this more explicit, suppose a physician determines that a particular patient has Disease A with 76% probability, and Disease B, C, or D each with 8% probability.

Suppose further that a diagnostic test, Test 1, provides evidence against Disease A, such that after seeing the test results the physician determines that each disease has 25% probability. Information gain and probability gain treat these results of Test 1 as having negative utility, whereas the other measures value the change in beliefs. Now suppose that Test 2 provides such strong evidence for Disease A that beliefs once again become 76% Disease A, and 8% for each of the other diseases. Any measure of the value of information that positively values changes in beliefs would regard results of both Test 1 and Test 2 (obtained after Test 1) as useful. KL distance, impact, Bayesian diagnosticity, and log diagnosticity are examples of such measures. By contrast, measures of the value of information that are additive (probability gain and information gain) view results of Test 1 as having negative utility, but results of Test 2 (obtained after Test 1) as having equal and opposite positive utility.

Evans and Over (1996) stated that the possible negativity of information gain was counterintuitive and theoretically problematic. (This criticism would apply equally to probability gain.) Evans and Over were also concerned that the probability of a category might fluctuate, sometimes above 50% and sometimes below 50%, indefinitely, as new information came in. The following section uses an example Planet Vuma scenario which was designed so that it is possible for beliefs about the probability of a category go from 75% to 25%, and back to 75%. The scenario illustrates (1) that indefinite fluctuation of beliefs of this sort is implausible, and (2) that nonnegativity itself can cause counterintuitive results. For instance, Bayesian diagnosticity values both Test 1 and Test 2 in the above medical diagnosis example, if the test results are obtained sequentially. However, if those test results are obtained simultaneously, such that beliefs about the probability of the diseases do not change, then Bayesian diagnosticity does not value them at all, despite the fact that the same information is obtained, and resulting beliefs are identical.

In the discussion below, *F1* refers to the question, whose answer is unknown, about whether or not feature 1 is present, and *f1* and *~f1* are the specific answers that feature 1 is present or not. eu(*F1*) is the expected utility (usefulness) of querying whether or not feature 1 is present; u(*f1*) and u(*~f1*) are the utility of learning that feature 1 is or is not present, respectively. Suppose that P(*glom*) = 0.75, P(*fizo*) = 0.25; P(*f1* | *glom*) = 0.11, P(*f1* | *fizo*) = 0.99; P(*f2* | *glom*) = 0.99, and P(*f2* | *fizo*) = 0.11. The features are conditionally independent, given the species (there are no symmetries).
Consider two learners, who encounter the same creature, as follows:

— the *sequential learner* first asks about feature 1, learning whether or not it is present, and then asks about feature 2, and learns whether it is present or not;

— the *all-at-once learner* asks about features 1 and 2 in a single question, and learns in one fell swoop whether each feature is present or not.

Both learners have perfect knowledge of environmental statistics and are optimal Bayesians in updating their beliefs. This means that irrespective of what creature they encounter, both learners come to identical posterior beliefs (Table 7.3). Yet if the learners use certain OED models to compute the utility of information, they may disagree about the information's usefulness! Two examples will illustrate.

Table 7.3. Possible feature values and posterior beliefs in example scenario.

Feature values	*f1, f2*	*f1, ~f2*	*~f1, f2*	*~f1, ~f2*
Probability of these feature values	0.1089	0.2211	0.6611	0.0089
P(*glom* | these feature values)	0.7500	0.0037	0.9996	0.7500
P(*fizo* | these feature values)	0.2500	0.9963	0.0004	0.2500

Note. P(*glom*)=0.75, P(*fizo*) = 0.25, P(*f1* | *glom*) = 0.11, P(*f1* | *fizo*) = 0.99, P(*f2* | *glom*) = 0.99, and P(*f2* | *fizo*) = 0.11. Features are class-conditionally independent. Probability of these combinations of feature values, and posterior probabilities of the categories, are the same for both the sequential and all-at-once learners.

Both features present case. The vast majority of the time, exactly one feature is present, such that the learner becomes almost certain of the true category once the presence or absence of each feature is known (Table 7.3). (This argues against the assertion that beliefs could fluctuate indefinitely.) About 11% of the time, however, both features are present. For the sequential learner, this causes beliefs about the probability of glom to change from 75% to 25%, after feature 1 is observed, and then back to 75% after feature 2 is observed. For the all-at-once learner, beliefs do not change at all, because $P(glom \mid f1, f2) = 0.75$, and $P(fizo \mid f1, f2) = 0.25$. Note, again, that both learners had the same prior beliefs, obtained the same information, and have the same posterior beliefs. How does each learner value this information (Table 7.4)? The all-at-once learner considers this information to be useless, irrespective of which OED utility they use to measure the value of information. Sequential learners who use information gain or probability gain to measure the value of information also consider the information to be useless. However, sequential learners who use KL distance, impact,

Table 7.4. How the sequential and all-at-once learners value queries if both features are present, in example scenario.

OED model	**Sequential learner: u(*f1*) + u(*f2* | *f1*)**	**All-at-once learner: u(*f1, f2*)**
probability gain	0.0000	0.0000
information gain	0.0000	0.0000
KL distance	1.5850	0.0000
Impact	1.0000	0.0000
Bayesian diagnosticity	18.0000	1.0000
$\log_{10}$ diagnosticity	1.9085	0.0000

Note. The two learners have the same posterior beliefs: P(*glom* | *f1*, *f2*) = 0.75, and P(*fizo* | *f1*, *f2*) = 0.25; these beliefs are equal to the prior beliefs. However, if KL distance, impact, Bayesian diagnosticity, or log diagnosticity are used to measure queries' usefulness, the learners disagree about whether the obtained information is useful.

Table 7.5. Sequential learner. Expected usefulness of *F1* question, eu(*F1*). Usefulness of each possible answer to *F1* question, right two columns.

OED model	eu(*F1*)	u(*f1*)	u(∼*f1*)
probability gain	0.1650	0.0000	0.2463
information gain	0.5198	0.0000	0.7758
KL distance	0.5198	0.7925	0.3855
impact	0.3300	0.5000	0.2463
Bayesian diagnosticity	62.6000	9.0000	89.0000
$\log_{10}$ diagnosticity	1.6210	0.9542	1.9494

Note. P(*f1*) = 0.33, P(∼*f1*) = 0.67, eu(*F1*) = 0.33· u(*f1*) + 0.67· u(∼*f1*).

Bayesian diagnosticity or log diagnosticity regard each obtained answer as informative, despite the fact that beliefs (in the end) did not change at all!

Expected utility, averaging over all possible answers. Would the two learners agree, even on average, about the usefulness of the obtained information in our example scenario? The sequential learner first experiences the usefulness of the first question (*F1*, about whether feature 1 is present), and then experiences the usefulness of the second question (*F2*), given what they already learned about the first feature. For the sequential learner, then, we need to calculate the expected usefulness of the first question, eu(*F1*) (Table 7.5), plus the expected usefulness of the second question given the first question, eu(*F2* | *F1*) (Table 7.6). For the all-at-once learner, who learns the presence or absence of both features simultaneously, we only need to compute eu(*F1*, *F2*) (Table 7.7). Do the two learners, who obtain the same information and have the same posterior beliefs, equally value that information (Table 7.8)? If the learners use KL distance, probability gain, or information gain, then on average, they equally value the

Table 7.6. Sequential learner. Expected usefulness of *F2* question, given *F1* question, eu(*F1* | *F2*). Usefulness of each possible answer to *F2* question, given each possible answer to *F1* question, in four right columns.

OED model	eu(*F1* | *F2*)	u(*f2* | *f1*)	u(∼*f2* | *f1*)	u(*f2* | ∼*f1*)	u(∼*f2* | ∼*f1*)
probability gain	0.0544	0.0000	0.2463	0.0033	–0.2463
information gain	0.1846	0.0000	0.7758	0.0302	–0.7758
KL distance	0.1846	0.7925	0.3855	0.0035	1.2093
impact	0.1133	0.5000	0.2463	0.0033	0.2463
Bayesian diagnosticity	27.4000	9.0000	89.0000	9.0000	89.0000
$\log_{10}$ diagnosticity	1.1831	0.9542	1.9494	0.9542	1.9494

Note. See Table 7.3 for the probability of each combination of feature values.

Table 7.7. All-at-once learner. Expected usefulness of *F2* and *F1* questions, answered simultaneously, eu(*F1*, *F2*). Usefulness of each possible set of answers to *F1* and *F2* questions, in four right columns.

OED model	eu(*F1*, *F2*)	u(*f1*, *f2*)	u(*f1*, ~*f2*)	u(~*f1*, *f2*)	u(~*f1*, ~*f2*)
probability gain	0.2194	0.0000	0.2463	0.2496	0.0000
information gain	0.7044	0.0000	0.7758	0.8060	0.0000
KL distance	0.7044	0.0000	1.9586	0.4104	0.0000
impact	0.3300	0.0000	0.7463	0.2496	0.0000
Bayesian diagnosticity	706.7600	1.0000	801.0000	801.0000	1.0000
$\log_{10}$ diagnosticity	2.5616	0.0000	2.9036	2.9036	0.0000

Note. See Table 7.3 for the probability of each combination of feature values.

obtained information. If the learners use impact, Bayesian diagnosticity, or log diagnosticity, they do not equally value that information, even on average. Suppose the learners were to use Bayesian diagnosticity. For the sequential learner, the total expected diagnosticity is 90 (eu($F1$) = 62.6; eu($F2 \mid F1$) = 27.4). Yet the all-at-once learner values the same information as having expected diagnosticity of 706.76! Which property is more important:

1. additivity, where eu($f1$) + eu($f2 \mid f1$) = eu($f1$, $f2$); or
2. nonnegativity in cases where beliefs change?

It seems that with respect to these properties, we can have additivity (probability gain, information gain) or nonnegativity (KL distance, impact, Bayesian diagnosticity, log diagnosticity), but not both. The above example shows how lacking additivity can lead to strange contradictions between sequential and all-at-once learners. Is nonnegativity also a critical property? Actually, neuroeconomic theories (e.g. Schultz, 1998) suggest

Table 7.8. Expected usefulness, calculated sequentially (sum of two questions) vs. all at once.

OED model	Sequential learner: eu(*F1*) + eu(*F2* | *F1*)	All-at-once learner: eu(*F1*, *F2*)
probability gain	0.2194	0.2194
information gain	0.7044	0.7044
KL distance	0.7044	0.7044
impact	0.4433	0.3300
Bayesian diagnosticity	90.0000	706.7600
$\log_{10}$ diagnosticity	2.8041	2.5616

Note: learners who use impact, Bayesian diagnosticity, or log diagnosticity will on average experience different sum utility if querying the features sequentially, vs. simultaneously, despite having the same prior and posterior beliefs.

that having negative utility is critical for learning. Savage's (1954, chap. 6) account of information acquisition also allows for zero or negative utility. It is ultimately an empirical question whether people experience some information as having zero or negative utility. Physiological measures such as recording from individual neurons (Nakamura, 2006), or EEG, MEG, fMRI, and galvanic skin response, as well as behavioral experiments, could potentially address this.

Would it ever make sense to use the diagnosticity measures?

Nelson (2005) argued that Bayesian diagnosticity and log diagnosticity are poor theoretical models of the utility of information, and are not needed to explain empirical data. Perhaps because those measures highly (even infinitely) value high certainty, several people have subsequently inquired whether the diagnosticity measures *should* be used in situations that require high certainty. The short answer is no. If a learner wishes to maximize a particular goal, they should directly compute the utility of candidate queries with respect to achieving that goal (Savage, 1954), without use of any OED theories of the value of information. However, we can still ask whether the diagnosticity measures might approximate the learner's goals in situations that require high certainty. This section illustrates that even in situations where high certainty is required, relying on the diagnosticity measures can be counterproductive (Table 7.9). In each situation here:

- querying feature 2 leads to higher probability of achieving the learner's goal;
- KL distance, information gain, impact, and probability gain prefer feature 2;
- Bayesian diagnosticity and log diagnosticity either prefer feature 1 or are indifferent between the features.

Scenario 1. Suppose that the learner must become 99% sure of the true category for the information to be useful, and therefore wishes to maximize the probability that their posterior beliefs will be at least 99% in favor of one category or the other. Would they then want to use Bayesian diagnosticity to choose what question to ask?

Table 7.9. Evidence-acquisition scenarios with more specific goals than information acquisition alone

Scenario	P(*glom*)	P(*f1* | *glom*), P(*f1* | *fizo*)	P(*f2* | *glom*), P(*f2* | *fizo*)	Goal considered
1	0.50	0.000, 0.100	0.010, 0.990	achieve 99% probability of either hypothesis
2	0.70	0.000, 0.200	0.000, 0.800	falsify working hypothesis
3	0.70	0.600, 0.001	0.001, 0.800	almost eliminate working hypothesis
4	0.70	0.200, 0.000	0.800, 0.000	eliminate alternate hypothesis

Note: Where P(*glom*) = 0.70, glom is the working hypothesis, and fizo is the alternate hypothesis.

Feature 2 offers 100% chance of being useful; feature 1 offers 5% chance of being useful. Yet the Bayesian diagnosticity (and log diagnosticity) of feature 1 is infinitely greater than that of feature 2.

Scenario 2. Suppose the learner needs to falsify their working hypothesis (glom) with 100% confidence for the information to be useful. This goal implements Popper's (1959) suggestion that scientists should do the experiment with the best hope of falsifying their current best hypothesis. Feature 2 falsifies the working hypotheses 24% of the time, whereas feature 1 falsifies the working hypothesis only 6% of the time. Yet Bayesian diagnosticity is indifferent between the features, as both features have infinite diagnosticity.

Scenario 3. Suppose the learner wishes to almost eliminate their working hypothesis (glom), by reducing its probability to no more than 1%. If feature 2 is present, this goal is achieved: $P(glom \mid f2) = 0.0029$; feature 2 is present about 24% of the time. Feature 1 *never* achieves this criterion, yet the diagnosticity measures prefer it.

Scenario 4. Suppose the learner wishes to have the highest probability of eliminating their *alternate* hypothesis (or, equivalently, achieving 100% confidence in their working hypothesis). If either feature is present, the alternate hypothesis is ruled-out. Feature 2 is present four times as often as feature 1 (56% versus 14% of the time). Yet the diagnosticity measures are indifferent between the features.

Research to date. Which OED model best explains human intuitions?

Several articles (Table 7.10), involving a diverse set of tasks, have examined whether human (or monkey) information acquisition follows OED principles. Some recent work has even explored whether eye movements could also be modeled in an OED framework, where each eye movement is modeled as a query of the visual scene that returns high-resolution information from the center of gaze and lower-resolution information from the periphery.

Which OED model best approximates human intuitions about the value of information, and choices of questions to ask? Most articles use only a single OED theory of information. Thus, it is entirely possible that different theories would disagree about which queries are most informative, and about whether human information acquisition is rational. To address this possibility, Nelson (2005) re-analyzed the tasks in several articles (Skov & Sherman, 1986; Baron *et al.*, 1988; Slowiaczek *et al.*, 1992; Oaksford & Chater, 2003; McKenzie & Mikkelsen, 2007) to identify the predictions of each of six OED models of the value of information, on each task. There was high agreement between models on which questions were most (and least) useful.[2]

[2] An important caveat is that small changes in the hypotheses and prior probabilities in a model can have strong implications on the apparent usefulness of subjects' queries (Nelson, 2005; Nelson *et al.*, 2001). How can researchers know what hypotheses and prior probabilities most accurately model subjects' beliefs? The complexity of various hypotheses (Feldman 2000, 2003, 2006; Nelson & Cottrell, 2007), meta-hypotheses about the types of hypothesis spaces that are feasible (Kemp *et al.*, 2006), generalization ratings (Tenenbaum, 1999, 2000; Tenenbaum & Griffiths, 2001), and error data can provide guidance. Use of natural sampling

Table 7.10. Value of information tasks that have been modeled in an OED framework

Task	References
Medical diagnosis	Good & Card (1971); Card & Good (1974); Baron, Beattie, & Hershey (1988)
Personality characteristics	Trope & Bassok (1982, 1983); Bassok & Trope (1983–1984)
Planet Vuma	Skov & Sherman (1986); Slowiaczek *et al*. (1992); Garcia-Marques *et al*. (2001); Nelson (2005); McKenzie (2006)
Selection task, Reduced array selection task	Oaksford & Chater (1994, 1996, 1998, 2003); Laming (1996); Over & Jessop (1998); Oaksford *et al*. (1997, 1999); Klauer (1999); Hattori (2002); Oaksford & Wakefield (2003)
Covariation assessment	McKenzie & Mikkelsen (2006)
Hypothesis testing	Klayman & Ha (1987); Klayman (1987)
2-4-6 task, active number concept task	Baron (1985); Ginzburg & Sejnowski (1996); Nelson & Movellan (2001); Nelson *et al*. (2001)
Alien mind reading (inferring causal structure)	Steyvers *et al*. (2003)
Urns and poker chips	Baron (1985)
Eyewitness identification	Wells & Lindsay (1980); Wells & Olson (2002)
Social contingency detection	Movellan (2005)
Eye movement tasks:	
Free viewing	Lee & Yu (2000); Itti & Baldi (2006); Bruce & Tsotsos (2006)
Reading	Legge *et al*. (1997, 2002); Legge *et al*. (2002)
Visual search	Najemnik & Geisler (2005); Zhang & Cottrell (submitted)
Shape learning	Renninger *et al*. (2005, 2007)
Contrast entropy reduction	Raj *et al*. (2005)
Target detection	Nakamura (2006)
Concept learning	Nelson & Cottrell (2007)

Note. Most of the articles here concerned experimental data with human subjects; some were purely theoretical. Pseudodiagnosticity articles (e.g. Doherty *et al*., 1996) address similar issues. Some eye movement models are 'bottom-up,' driven by image properties in a task-blind manner, in an OED framework.

to convey environmental statistics to subjects (Knowlton *et al*., 1994; Oaksford & Wakefield, 2003) may help ensure that subjects assimilate the hypothesized probabilities better than via reading words and numbers alone. People seem well calibrated to statistics of visual scenes (Knill & Richards, 1996; Kersten *et al*., 2004; Yuille *et al*., 2004; Knill, 2006). Given this, eye movement tasks, in which each eye movement is modeled as a query of a visual scene, can also be helpful. Irrespective of how subjects learn prior probabilities, it is helpful to show (1) that subjects' beliefs are accurately described by a particular probability model, or (2) that the ordering of the usefulness of various queries does not depend on which plausible model best describes subjects' beliefs.

This result supported the feasibility of a rational (Anderson, 1990) theory of information acquisition, suggesting that irrespective of which OED model is the best motivated theoretically, human behavior closely agrees with it. However, this result left unclear which theoretical model best matches human intuitions.

Is it possible to differentiate which OED models best describe human intuitions? Nelson (2005) used computer optimization to find a limiting scenario in which the various OED models strongly disagree about which features are most useful, and then tested that scenario (Table 7.11) with human subjects. Subjects were asked to rank order the features from most to least useful, if they could just ask about one feature to classify a creature as a glom or fizo. Most OED models predict that the harmonica feature will be ranked as most useful, and the hula feature as least useful. Bayesian diagnosticity and log diagnosticity, however, hold that the hula feature is infinitely useful, because it sometimes (with probability 1 in 200) leads to certainty of the true species.

What did the 148 subjects think? Responses were strongly positively correlated with most OED models (Spearman rank correlations of 0.69–0.78), yet negatively correlated with Bayesian diagnosticity (correlation = −0.41) and with log diagnosticity (correlation = −0.22). A majority of the subjects gave a rank order that exactly matched information gain, KL distance, impact, or probability gain. No subjects gave a rank order that matched either Bayesian diagnosticity or log diagnosticity. This result suggests that the diagnosticity measures do not provide the most accurate approximation of human intuitions about the value of information.

Results from Nelson's (2005) experiment do not distinguish among probability gain, impact, information gain, and KL distance, because the predictions of those models were highly (or perfectly) correlated. Nelson did describe situations in which the remaining theories make moderately contradictory predictions. However, these situations include unequal prior probabilities of the categories (glom and fizo), which could not be effectively conveyed to subjects with a words-and-numbers-based experiment. If the desired environmental probabilities could be effectively conveyed to subjects, it could prove feasible to discriminate which of the remaining OED models best describes human intuitions.

Table 7.11. Features used in Nelson's (2005) behavioral experiment

	Drink	**Harmonica**	**Gurgle**	**Hula**
P(*feature* | *glom*)	0.0001	0.01	0.30	0.99
P(*feature* | *fizo*)	0.30	0.99	0.70	1.00

Note: Prior P(*glom*) = P(*fizo*) = 0.50. From Table 10 in Nelson (2005), copyright American Psychological Association; adapted with permission.

Suboptimal heuristics vs. OED models

The research discussed in the chapter suggests that OED theories provide a good approximation to human intuition about the usefulness of information. But do OED theories provide the best descriptive account? Perhaps people rely on heuristic strategies that only partly approximate optimal models. How could we tell? Here we briefly consider three heuristic strategies that have been reported in the literature. (Many others, including fast and frugal heuristics, e.g. Chase *et al.*, 1998, could be considered as well.)

Feature difference heuristic. Skov and Sherman (1986), and Slowiaczek *et al.* (1992) noted that many subjects query features with the highest absolute difference in feature probabilities, e.g. the feature with maximal abs($P(f \mid h_1) - P(f \mid h_2)$). Those authors called this strategy heuristic, implying that it is suboptimal. However, Nelson (2005), who also observed this heuristic empirically, proved that the feature difference heuristic exactly implements the impact OED model. This is true in all situations where there are two categories of objects and binary features, irrespective of the prior probabilities of the categories and the specific feature probabilities. Because the feature difference heuristic exactly implements an optimal model, its use does not support claims of suboptimality in human information acquisition.

Information bias. Baron *et al.* (1988) reported that even when a feature (a medical test, in their experiment) had zero probability gain, subjects judged it as useful if it would change beliefs. Baron *et al.* called this phenomenon information bias. This phenomenon may be a bias if probability gain, which Baron *et al.* used as their normative benchmark, is uniquely applicable as a normative model. However, information gain, KL distance, and impact also value queries that lead to changed beliefs but not improvement in probability of correct guess. In other words, whether or not information bias is in fact a bias depends on what normative model one applies to the situation.

Hypothesis confirmation. Skov and Sherman (1986), in a planet Vuma scenario, defined hypothesis-confirming queries as testing features for which $P(f \mid h_1) > P(f \mid h_2)$, where h_1 is the focal hypothesis. (As Klayman, 1995, noted, if Bayes' theorem is used to update beliefs, no question can introduce bias, although some questions can be more informative than others. Skov and Sherman's definition of hypothesis confirmation, which we adopt here, implies that a feature's presence favors h_1, and its absence favors h_2.) The questions available to subjects in Skov and Sherman's experiment included a variety of high, medium, and low usefulness features. Most subjects were in conditions where glom or fizo was marked as the working hypothesis, such that features could also be classified as confirmatory or disconfirmatory. Subjects were very sensitive to usefulness: about 74% of choices of questions were to high usefulness features, 21% to medium usefulness features, and 4% to low usefulness features (chance would be 33% each). Were subjects also influenced by hypothesis confirmation? Despite Skov and Sherman's statement (p. 93, echoed later in reviews by Klayman, 1995, Nickerson, 1998) that subjects showed a 'strong and consistent tendency to ask hypothesis-confirming questions,' the evidence was comparatively weak. About 60% of choices were to hypothesis-confirming features (chance would be 50%). This result is consistent with 80% of subjects being indifferent to whether a feature is hypothesis-confirming or not, and just 20% of subjects preferring hypothesis-confirming features.

But because there was no tradeoff between usefulness and hypothesis confirmation, the results do not show whether *any* subject would give up *any* information to test a hypothesis-confirming feature.

Optimal experimental design as a way to implement strong inference

By and large, research to date suggests that human information acquisition is in good accord with rational (Bayesian optimal experimental design) principles, on a wide variety of tasks. In situations where some OED theories are better motivated than others, the better-motivated theories tend to better describe human intuitions and queries. Yet the tasks studied to date barely scratch the surface of the theoretical issues that are important in information acquisition. For instance, planning several steps ahead is more efficient than just planning the next query, but little research has studied sequential queries. Similarly, whether human information acquisition is sensitive to class-conditional feature dependencies (such as bilateral symmetries in vertebrates) is not known. Some research suggests that monkey eye movements (Nakamura, 2006) may follow OED principles. It would be helpful to also study other cases of perceptual information acquisition, such as bat echolocation and rat whisking. Can OED principles be used to design the most informative experiments for addressing these issues?

Platt (1964), inspired by a tradition of scientific thought, suggested that science progresses most quickly when scientists enumerate alternate hypotheses to explain a phenomenon, devise experiments in which those hypotheses make strongly contradictory predictions, and then conduct those experiments. A limit in some research is that the stimuli were not designed to maximally differentiate between competing theoretical models. In some cases the theoretical account reduces to 'theory X claims query 1 is more useful than query 2; subjects agree; therefore subjects follow theory X.' The problem is that if every theory holds that query 1 is more useful than query 2, it is neither surprising nor interesting that people agree. A challenge in future work is to find and test cases in which candidate theories most strongly disagree about which queries people will conduct. If people have strong intuitions and consistent behavior in those cases, the results will of course contribute to theory of human information acquisition. But equally importantly, those scenarios may enhance our own developing concepts (Baron 2002, 2004) of what utilities are best suited to quantify the value of information in scientific inference and in other tasks.

Author note

I thank Nick Chater, Ulrike Hahn, and Mike Oaksford both for ideas towards the research discussed here, and for helpful feedback on a draft manuscript; Flavia Filimon, Craig McKenzie, Javier Movellan, Jon Baron, Terry Sejnowski, and Gary Cottrell for research ideas; Craig Fox for suggesting certain scenarios in the 'Would it ever make sense to use the diagnosticity measures' section; Tim Marks for suggesting the 'blinding light and blowing sand' interpretation of planet Vuma; and NIH 5T32MH020002 for funding support. Any corrections will be posted at http://www.jonathandnelson.com/.

References

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

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

Baron, J. (2002). *Value trade-offs and the nature of utility: Bias, inconsistency, protected values, and other problems*. Paper for conference on behavioral economics. American Institute for Economic Research, Great Barrington, MA, July, 2002.

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

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.

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

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

Bruce, N., & Tsotsos, J. K. (2006). Saliency Based on Information Maximization. In Y Weiss, B Sch&oumlaut;lkopf, & J Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 155–162). Cambridge, MA: MIT 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, pp. 60.1–60.23). Oxford, England: Blackwell.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998) Visions of rationality. *Trends in Cognitive Science*, **2**(6), 206–214.

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, England: Oxford University Press.

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

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

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**, 644–654.

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

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, **407**, 630–633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, **6**, 227–232.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, **50**, 339–368.

Garcia-Marques, L., Sherman, S. J., & Palma-Oliveira. J. M. (2001). Hypothesis testing and the perception of diagnosticity. *Journal of Experimental Social Psychology*, **37**, 183–200.

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

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

Good, I. J. (1983). *Good thinking*. Minneapolis: 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.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In Y. Weiss, B. Scholköpf, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 547–554). Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2006). Learning overhypotheses. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 417–422).

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, **55**, 271–304. doi:10.1146/annurev.psych.55.090902.142005

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

Klayman, J. (1987). An information theory analysis of the value of information in hypothesis testing. Retrieved May 23, 2005, from http://www.chicagocdr.org/cdrpubs/

Klayman, J. (1995). Varieties of confirmation bias. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from a cognitive perspective*. New York: Academic Press.

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

Knill, D. C. (2006). Learning Bayesian priors for depth perception [Abstract]. *Journal of Vision*, **6**(6), 412, 412a, http://journalofvision.org/6/6/412/, doi:10.1167/6.6.412.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

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

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**, 364–373.

Lee, T. S., & Yu, S. X. (2000). An information-theoretic framework for understanding saccadic eye movements. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 834–840). Cambridge, MA: MIT Press.

Legge, G. E.; Hooven, T. A.; Klitz, T. S.; Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, **42**, 2219–2234.

Legge, G. E.; Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal observer model of reading. *Psychological Review*, **104**, 524–553.

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

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition*, **23**(3), 577–588.

McKenzie, C. R. M., & Mikkelsen, L. A. (2006). A Bayesian view of covariation assessment. *Cognitive Psychology*, **54**(1), 33–61. doi:10.1016/j.cogpsych.2006.04.004

Movellan, J. R. (2005). An infomax controller for real time detection of contingency. In *Proceedings of the International Conference on Development and Learning*, Osaka, Japan, July, 2005.

Najemnik, J., & Geisler, W. S. (2005, March 17). Optimal eye movement strategies in visual search. *Nature*, **434**, 387–391.

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

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

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

Nelson, J. D., & Movellan, J. R. (2001). Active inference in concept learning. *Advances in Neural Information Processing Systems*, **13**, 45–51.

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

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

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

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

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

Oaksford, M., & Chater, N. (1998). A revised rational analysis of the selection task: Exceptions and sequential sampling. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 372–393). Oxford, England: Oxford University Press.

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

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition*, **31**(1), 143–154.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning*, **5**, 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**, 441–458.

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

Platt, J. R. (1964). Strong inference. *Science*, **146**(3642), 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. C. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America, A: Optics, Image Science, and Vision*, **22**(10), 2039–2049.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1121–1128). Cambrdige, MA: MIT Press.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, **7**(3), 6, 1–17, http://journalofvision.org/7/3/6/, doi:10.1167/7.3.6.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, **80**, 1–27.

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**, 93–121.

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

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

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Ph.D. Thesis, MIT.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, **24**(4), 629–640.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, **16**(3–4), 255–275.

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.

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

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, **8**, 155–167.

Yuille, A. L., Fang, F., Schrater, P., & Kersten, D. (2004). Human and ideal observers for detecting image curves. In S. Thrun; L. Saul & B. Schoelkopf (Eds.), *Advances in neural information processing systems* (Vol. 16). Cambridge, MA: MIT Press.

Zhang, L., & Cottrell, G. W. (submitted). Probabilistic search 1.0: a new theory of visual search.