Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 2256–2272
NEUROCOMPUTING
www.elsevier.com/locate/neucom

A probabilistic model of eye movements in concept formation

Jonathan D. Nelson[a,*], Garrison W. Cottrell[b]

[a]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd., La Jolla, CA 92037 1099, USA
[b]Computer Science and Engineering Department, University of California, San Diego, 9500 Gilman Dr., Dept. 0404, La Jolla, CA 92093 0404, USA

Received 5 February 2005; received in revised form 26 August 2005; accepted 1 February 2006
Available online 2 January 2007

Abstract

It has been unclear whether optimal experimental design accounts of data selection may offer insight into evidence acquisition tasks in which the learner's beliefs change greatly during the course of learning. Data from Rehder and Hoffman's [Eyetracking and selective attention in category learning, Cognitive Psychol. 51 (2005) 1–41] eye movement version of Shepard, Horland and Jenkins' classic concept learning task provide an opportunity to address these issues. We introduce a principled probabilistic concept-learning model that describes the development of subjects' beliefs on that task. We use that learning model, together with a sampling function inspired by theory of optimal experimental design, to predict subjects' eye movements on the active learning version of that task. Results show that the same rational sampling function can predict eye movements early in learning, when uncertainty is high, as well as late in learning when the learner is certain of the true category.

Keywords: Saccadic eye movement; Information theory; Bayesian reasoning; Concept formation; Optimal experimental design; Intuitive statistics

1. Introduction

Learning about the world and querying the world are both central to competent behavior in many situations. Consider the process by which a physician diagnoses a patient's illness. It is not possible to conduct every relevant medical test. The physician must conduct tests that are expected to provide the best improvement in the ability to correctly diagnose and treat the patient [54]. The new information obtained by the most recent test is used to better select the next test, in an iterative process. During this process, the physician's beliefs about the patient's illness may change drastically. Suppose a patient has a deep cough. The presumed cause of the cough may be nasal drip, caused by an allergy, and anti-allergy medication might be an appropriate treatment. However, suppose a chest X-ray were to show abnormal appearance of a lung. In this case, pneumonia or tuberculosis would become more probable explanations of the illness, and the most appropriate diagnostic tests and treatments would change drastically.

What is known about how human beliefs develop as new information is obtained, and about how people's queries relate to their beliefs at the time those queries are made? Can Bayesian statistics help explain how beliefs develop, and can theory of optimal experimental design help explain people's queries? Research on Bayesian reasoning, and on Bayesian theories of concept learning, suggests that Bayesian statistics may provide a useful language for understanding human belief revision. If beliefs can be described in a Bayesian framework, then principles of optimal experimental design [43, chapter 6, pp. 105–119] can be used to identify potentially useful queries. A variety of articles suggest that this framework may provide a useful foundation for a descriptive theory of human queries as well [27].

The present article concerns learning and selective attention in Shepard et al.'s [46] concept learning task. In this task, subjects learn to categorize objects where each object has three binary dimensions, such as size (large or small), shape (square or circle), and color (black or white). Accounts of learning in this task have largely been based

*Corresponding author. Tel.: +1 858 922 2536; fax: +1 858 587 0417.
E-mail addresses: jnelson@salk.edu (J.D. Nelson), gary@ucsd.edu (G.W. Cottrell).

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.02.026

on one or more mechanistic theories, rather than on rational principles (for a review, see [32,55]), with the notable exception of Anderson [2]. Many of these accounts of learning have as a critical component a selective attention mechanism, to describe when during learning a subject attends more to the size dimension, when to color, and so on, as a function of the type of concept being learned. However, data on selective attention per se have not been available. In order to directly measure subjects' attentional focus, Rehder and Hoffman [38,39] separated the stimulus dimensions in space as characters on a computer screen, in an eye movement version of the task. Subjects' eye movements then provided direct evidence for what stimulus dimension a subject was attending to at each point in learning.

One interesting property of Shepard et al.'s [46] task is that learning takes place over an extended period of time, involving dozens or hundreds of trials, and that subjects' beliefs can change drastically over the course of learning. In this article, we introduce a probabilistic model of learning to help explain how beliefs develop as subjects perform this task. One key property of our model of learning is our extension of earlier probabilistic learning and inference models [9,49,51] to cases of imperfect memory or noisy data. We then evaluate whether this probabilistic model of learning can offer insight into eye movements in this task using theories of optimal experimental design, rather than mechanistic (process) theories of selective attention. More specifically, we introduce a model of eye movements (queries) in which the subject's eye movements are optimal at each point in learning, but in which the specific eye movements change during learning as a function of the development of knowledge during the task. At each point in learning, our model represents a learner's knowledge as a probability distribution over the possible concepts.

The rest of this article is structured as follows. In Section 2, we discuss Shepard et al.'s [46] classic task, our Bayesian model of subjects' learning on that task, how model performance compares with human data, and novel predictions derived from our model. In Section 3, we discuss Rehder and Hoffman's [38,39] eye movement version of Shepard et al.'s task. Section 3 also introduces our eye movement model, and shows how this model can provide a rational explanation of much of Rehder and Hoffman's data. Section 4 summarizes our findings, what we take them to imply, and lays out issues for future research.

2. The task; our probabilistic concept learning model

Shepard et al.'s [46] task involves eight objects, formed by taking every combination of three binary stimulus dimensions, such as size (large or small), shape (square or circle), and color (white or black). Concepts are comprised of a subset of those eight objects.[1] There are 256 (2^8) possible concepts, although behavioral experiments have mainly considered only the 70 (8 choose 4) concepts of size 4. ([10] is a notable exception.) Shepard et al. [46] defined a taxonomy of six types of concepts; Table 1 gives illustrative examples. Type I concepts are defined by a single stimulus dimension, for example "large," or "circle". Type II concepts are defined with reference to two stimulus dimensions; an example is "large square or small circle." Type III, IV, and V concepts require attention to all three stimulus dimensions for correct classification, but above-chance performance can be achieved by attending a single dimension. One Type IV concept is "large, but also the small white circle and excluding the large black square." Type VI concepts require attention to all three stimulus dimensions; they are usually learned by memorizing the specific set of objects that is consistent with them.

Table 1
Example concepts

Type	Objects	Err.
I		8.2
II		31.2
IV		36.9
VI		70.6

Note: Err. column is average number of errors in Rehder and Hoffman's [39] study.

In an individual trial, (1) an object is drawn at random from the set of eight objects, (2) the learner judges it as consistent or inconsistent with the true, hidden concept, and (3) the learner is given feedback. The objects are usually drawn without replacement to form a block of trials, such that objects appear in a random order but with the constraint that each object appears exactly once (or twice) in a block. Shepard et al. [46] found that across several experimental manipulations, criterion performance was reached most quickly on Type I concepts. The number of trials to criterion performance was ordered as follows:

$$\text{Type I} < \text{II} < \text{III} \sim \text{IV} \sim \text{V} < \text{VI}.$$

(Types III, IV, and V were not reliably different from each other.) Nosofsky et al. [32,33] obtained similar results, with 44.0, 85.4, 121.6, 127.0, 133.8, and 189.2 trials to criterion performance, for concept Types I through VI, respectively.

[1] An equivalent alternative is to describe concepts as separating two types of objects. Here it is easier to discuss whether an object is consistent with a concept or not.

2.1. Introduction to the probabilistic learning model

In this section, we introduce a new probabilistic model of how the learner's beliefs develop over the course of many trials, in Shepard et al.'s [46] concept learning task. This model is similar in spirit to Anderson's [2] model, but makes fewer assumptions as to the cognitive processes underlying inference. We explicitly specify a generative model, which is meant to capture the learner's beliefs about the task. Learning, in our model, is optimal Bayesian inference with respect to that generative model.

We use probability theory to describe the learner's prior beliefs, and how those beliefs change in response to new information. We use standard probabilistic notation, in which random variables begin with capital letters, and specific values of random variables begin with lowercase letters. For example, C represents the learner's beliefs about the a priori probability of each of the concepts, represented in a prior probability distribution over all concepts c_i. (In our probability model, which includes 70 possible concepts, C is a vector with 70 probabilities that sum to one.) $C = c_i$ represents the event that C takes the value c_i, for instance that the true concept is "small". The learner typically wishes to know $P(C = c_i|X = x)$, the posterior probability of a particular concept given a particular observed datum x. For instance, the learner may wish to know the probability that "small" is the true concept, given that in their first trial of learning, they observe that a small black circle is consistent with the true concept. [For concise notation in the article, we will use lowercase letters to denote specific values taken by random variables, for instance c_i, rather than $C = c_i$.] The posterior probability of the concept "small" is not observed directly. Rather, $P(c_i|x)$ is calculated according to Bayes' [4] theorem, with reference to both the prior probability of the concept "small", $P(c_i)$, and the a priori likelihood of observing the small black circle as a positive example, if the true concept were "small," $P(x|c_i)$. By Bayes' theorem, the posterior probability of a particular concept is calculated as follows:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}.$$

The simulated model learner obtains the posterior probability of every concept in this way. $P(x)$, the probability of the observed datum, is a normalizing constant that is computed by summing $P(x|c_i)P(c_i)$, the numerator of the previous term, over all concepts c_i:

$$P(x) = \sum_{c_i} P(x|c_i)P(c_i).$$

2.2. Capturing the learner's intuitions with prior probabilities

If the learner had no intuition that certain concepts were more plausible than others, each concept could be given the same prior probability. However, in the case of Shepard et al.'s [46] task, some concepts, such as "large", are much more plausible than other concepts, such as random collections of objects. In a probabilistic modeling framework, we seek to capture the learner's intuitions by giving more plausible concepts higher prior probability, and less plausible concepts lower prior probability.

In the current article, our primary focus is on understanding what Rehder and Hoffman's [39] subjects believed, so as to most accurately understand subjects' eye movement behavior, rather than on providing a complete explanation for why subjects had particular beliefs. We therefore set the prior probability of each concept of a particular type to be inversely proportional to the average number of errors committed by Rehder and Hoffman's study participants ("Err." column in Table 1) when learning that type of concept, ε_i, raised to a power, λ, a free parameter in our model:

$$P(c_i) \propto (\varepsilon_i)^{-\lambda}.$$

Choice of this particular function is relatively arbitrary; it serves to provide spread in the prior distribution. If $\lambda = 0$, all concepts have equal prior probability; if λ is large, Type I concepts have most of the prior probability. All concepts of a particular type, for example the Type I concepts "small" and "circle", receive the same prior probability. Our model only assigns non-zero prior probability to the 70 concepts that include exactly 4 of the 8 objects, because most empirical research focuses on these types of concepts. Because Type III–V concepts are not reliably different from each other in the speed at which they are learned [32,46], and because Rehder and Hoffman's [39] data do not include Type III or V concepts, our model gives Type III and V concepts the same prior probability as Type IV concepts.

Note that our primary goal in setting priors in the present concept learning model is to capture the learner's beliefs as accurately as possible, so as to predict the learner's eye movements. While the errors committed by Rehder and Hoffman's [39] subjects provide a reasonable basis for understanding what the learner's prior beliefs are, and serve as a reasonable basis for a model that describes the process of learning and eye movements, those errors do not constitute an explanation of the ultimate source of the learner's beliefs about each type of concept's plausibility.

A number of a priori criteria that do not depend on empirical data could be tested as providing a basis for prior probabilities in our model. One such criterion is the number of stimulus dimensions that must be attended, on average, to correctly classify an object, if the true concept is known with certainty. Shepard et al. [46] discussed this criterion. Type I concepts require a single dimension; Type VI concepts require three dimensions, and other types of concepts require intermediate numbers of dimensions. Another criterion, which Feldman [10,11] has investigated in relation to concepts' learnability, is Boolean complexity. In future work, we intend to investigate whether some of

these a priori criteria might provide a suitable basis for the prior probability distribution.

2.3. Belief updating: the likelihood function

Although the assignment of prior probabilities is frequently thought to be the most important part of a Bayesian model, the likelihood function—which describes how new information changes beliefs—is equally critical. In the case of our concept-learning model, the likelihood function should describe the learner's beliefs about how the example data, including their labels (in the form of the feedback given after each trial), are generated.

The likelihood functions of classic Bayesian reasoning models [9], and recent Bayesian concept learning models [28,49,51] would appear to be applicable in this task.[2] However, these models have the property that if an observed datum is incompatible with a particular hypothesis, that hypothesis is eliminated from consideration. Under these models, one exposure to every possible object uniquely identifies the true concept, by eliminating all other concepts. If applied to the concept-learning task, these types of models would predict that after a single block of trials, in which each object is viewed once with feedback, the learner would know the true concept with complete certainty. This result would hold irrespective of the type of concept being learned, and irrespective of the prior probabilities of each concept, such as whether the concept had prior probability 1/10 or 1/1000. Empirical results strongly contradict this implicit claim of those models, however. In fact, no concept is learned after a single block of trials, and there are clear distinctions wherein some types of concept (such as the Type I concept "square") are learned more quickly than others (such as the Type VI concept "large black square or small white square or large white circle or small black circle"). The limitation underlying the classic likelihood functions is that a single observation completely eliminates incompatible concepts. For instance, observation of the large white circle as a positive example would immediately eliminate concepts such as "square", "small", and "black circle or white square."

To account for the empirical finding that a single example observation does not eliminate incompatible concepts, we assume that the subjects do not encode the examples perfectly in memory. To model this, we include two generators of data. One data generator is uninformative, and labels objects as consistent or inconsistent with the true concept randomly with equal probability. The other data generator is veridical, and always correctly labels example objects. We use a parameter μ, between zero and one, for the probability that the informative data generator generated a particular observation. When $\mu = 1$, the model behaves in a manner analogous to earlier probabilistic concept learning models [49,51], and learns the task in a single block of trials, irrespective of the type of concept and prior distribution. When $\mu = 0$, the likelihood function is uninformative and no learning ever takes place. Hence, μ can also be thought of as a learning rate (Fig. 1, [35]). The likelihood of a particular observation x (a combination of an object and its label as consistent or inconsistent with the true concept), given a particular concept (c_i), where g_v and g_u are the veridical and uninformative data generators, respectively, is given by:

$$P(x|c_i) = P(x|c_i, g_v)P(g_v) + P(x|c_i, g_u)P(g_u)$$

with $P(g_v) = \mu$, for the veridical generator and $P(g_u) = 1-\mu$, for the noise generator. In the case where all stimulus dimensions are observed at once, as in the classic concept-learning task, if the veridical data generator is selected in a particular trial, then there are 8 equally probable observations (4 positive and 4 negative); if the noise generator is selected, then there are 16 equally probable observations (8 positive and 8 negative). In this case $P(x|c_i)$ reduces to $(1/8)\mu + (1/16)(1-\mu)$ if x is consistent with the concept c_i, and $(1/16)(1-\mu)$ otherwise. To approximate the learner's beliefs, the likelihood function assumes that the observations are conditionally independent of each other, given the true concept. The psychological premise of this assumption is that the learner does not know that the objects are drawn without replacement in each block of trials, and believes that objects are drawn randomly throughout learning.[3] Bayes' theorem is used to update beliefs.

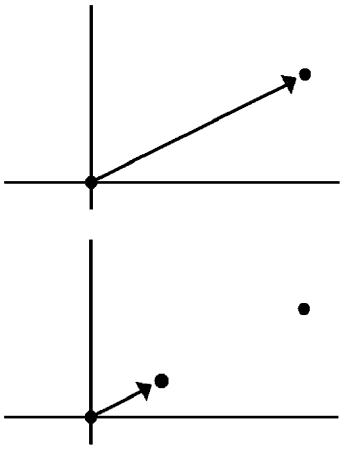

Fig. 1. Belief update if $\mu = 1$, top; if $\mu < 1$, beliefs change less but in same direction, bottom.

2.3.1. Belief updating illustrated with an example

The likelihood function is illustrated with a simplified example, where the concepts "square" and "large" each

[2]Tenenbaum [49,50] and Tenenbaum and Griffiths [51] focused on situations where the learner encounters only positive examples. However, Tenenbaum's models can update beliefs based on negative instances, provided that the likelihood function specifies how those instances are generated.

[3]We believe that this discrepancy between our generative model of this task, and the actual task, is minor. However, it would be interesting, in a future experiment with human subjects, to actually draw example objects independently, with replacement, throughout learning.

have prior probability 0.50, and the other concepts have prior probability zero. (Except for the simplified prior probability distribution, this example is identical to the full concept-learning model.) Let $\mu = 0.20$. We assume that the learner is exposed to all stimulus dimensions, as is the case in most versions of Shepard et al.'s [46] task. The likelihood of a particular observation, if all stimulus dimensions are attended, is 1/8 given a particular concept and the veridical generator, and 1/16 given a particular concept and the noise generator, as in the full-scale model.

Suppose that on a particular trial, the learner sees the large white circle as a positive example of the concept. The learner believes that with probability 0.20, the observation is veridical; and with probability 0.80, the observation is uninformative. In the equations below we use x to denote that the observed object is the large white circle, labeled as a positive example of the unknown true concept. There are 16 possible observations, corresponding to each of the 8 objects as positive and negative examples of the true concept. We use g_v to denote that the veridical generator of data is selected, and g_u to denote that the uninformative generator is selected. We use c_s to denote that the true concept is "square," and c_l to denote that the true concept is "large." If the observation is from the veridical data generator, then the posterior probability of "square" is 0:

$$P(c_s|x, g_v) = \frac{P(x|c_s, g_v)P(c_s|g_v)}{P(x|g_v)} = \frac{P(x|c_s, g_v)P(c_s)}{P(x)} = \frac{0 \times \frac{1}{2}}{\frac{1}{16}} = 0.$$

[$P(c_s|g_v) = P(c_s)$, and $P(x|g_v) = P(x)$, because $P(c_s)$ and $P(x)$ do not depend on G.] Similarly, the probability of the concept "large," given this observation, would be 1:

$$P(c_l|x, g_v) = \frac{P(x|c_l, g_v)P(c_l)}{P(x)} = \frac{\frac{1}{8} \times \frac{1}{2}}{\frac{1}{16}} = 1.$$

If the observation is uninformative, then posterior probabilities of "square" and "large" would each be 0.5, unchanged from the priors. Consider the posterior probability of "large":

$$P(c_l|x, g_u) = \frac{P(x|c_l, g_u)P(c_l)}{P(x)} = \frac{\frac{1}{16} \times \frac{1}{2}}{\frac{1}{16}} = 0.5.$$

The learner, of course, does not know whether the veridical or uninformative generator of labels was selected on this trial, and therefore must average over both possibilities. The posterior probability of the concept "square", given the observed datum x:

$$P(c_s|x) = \mu P(c_s|g_v) + (1 - \mu)P(c_s|g_u) = 0.2(0) + 0.8(0.5) = 0.4.$$

Similarly, $P(c_l|x)$ the posterior probability of "large", given the observed datum x, is 0.6.

2.3.2. Comment on μ

We introduced μ as reflecting imperfect encoding of examples in memory on the part of the learner. Other conceptualizations of μ are also viable. An ideal observer, with perfect perception and memory, would update beliefs in the same manner as our likelihood function if it believed that individual examples (such as the large white circle, if drawn on a particular trial) were correctly labeled with probability μ, and randomly labeled (positive or negative by flip of a fair coin) with probability $1-\mu$. If an ideal observer has $\mu < 1$, that observer is skeptical about the process by which examples are labeled. In fact, classic work on *conservatism* in Bayesian reasoning [9] suggests that even in situations with no memory component, μ is likely to be less than 1 for human subjects. This suggests that although we primarily describe μ as reflecting imperfect memory, it may also represent some element of skepticism on the part of subjects. Another way of describing our likelihood function would be to state that we do not know whether a particular learner remembers a particular example, but that we believe that in general, learners remember examples with probability μ. In that case, our concept leaning model might not describe the beliefs of any particular learner, but rather would describe the mean of all learners' beliefs. Computation of the model learner's posterior probabilities does not depend on which explanation of μ one favors.

Our model learner's likelihood function illustrates how imperfect encoding or memory, or the learner's belief that the data are noisy, can be modeled in a generative framework of optimal inference. This likelihood function behaves in the same manner as that of Oaksford and Chater [35], and a similar manner as conservative likelihood functions from classic research on Bayesian reasoning [9]. Most authors have viewed conservatism as an error, or at least as a suboptimal property of human inference. Our generative formulation shows that if a learner is skeptical that data are noisy, being conservative in belief updating is the optimal Bayesian behavior.

2.4. Simulation of the probability model's learning

Does our model account for results in the classic non-eye movement concept-learning task, in which all three stimulus dimensions are presented in every trial? Fig. 2, top panel, shows how the model's classification performance improves during the course of learning, for each type of concept. Qualitatively, the most important result is that the ordering observed in empirical investigations (e.g. [39,46]) is preserved, with high performance on Type I concepts achieved most quickly, followed, in order, by Type II, IV, and VI. Because the concept-learning model explicitly represents beliefs about the probability of each concept, at each point in learning, it will also prove possible to query the model learner in ways that go beyond available empirical data on human learning. Note that although the probability model that our model learner uses

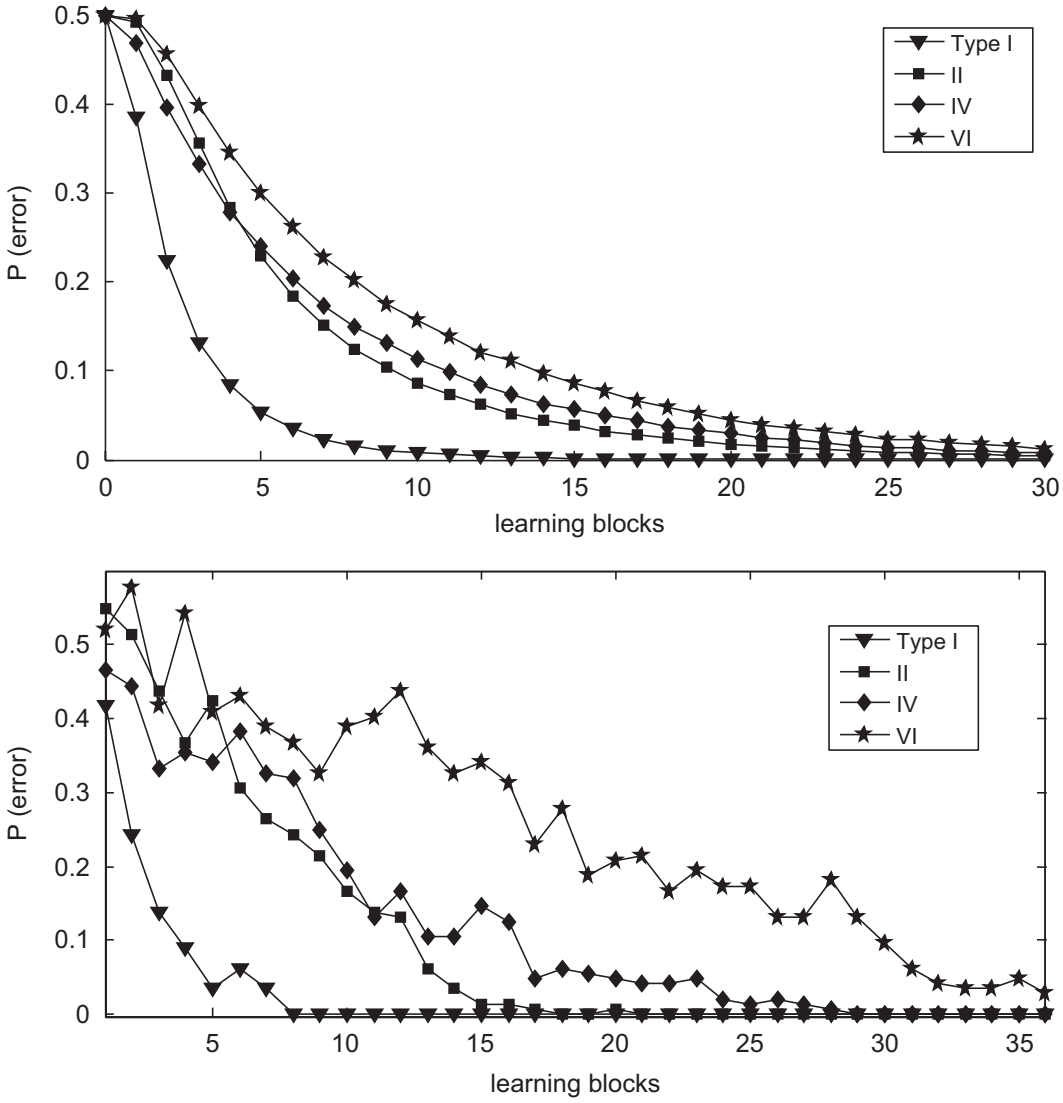

Fig. 2. Error rates through learning. Top: model learner. Bottom: data from Rehder and Hoffman [39], courtesy Bob Rehder and Aaron Hoffman.

to calculate posterior probabilities does include Type III and V concepts, for easiest comparison to Rehder and Hoffman's [39] results, we plot only Types I, II, IV, and VI in the charts in this article. We set $\lambda = 5$ and $\mu = 0.10$ in all simulations in this article, choosing those values by hand. (The qualitative pattern of results is maintained across a wide range of parameter values; in future work we intend to optimize these parameters quantitatively.)

2.4.1. The model learner's probability of error

The goal of studying concept learning is largely to know how people's *beliefs* change through experience. However, evidence for belief change has largely been in the form of *error rates* during learning. It is therefore important to assess the model learner's error rates. Here we predict the model learner's error rate (Fig. 2, top panel) using a probability matching response function, in which a response is picked proportional to the estimated probability that is correct, rather than the optimal winner-take-all response function. This appears to be necessary to keep the model learner's error rates from decreasing too rapidly, compared with the human data. Use of probability matching response functions is thought to be common in human subjects (but see [44]). Note that the choice of response function here does not affect the model's learning or eye movements.

Let x represent an unlabeled observation, such as the large white square, before the learner is given feedback during a particular trial.[4] Let A be a binary random variable that indicates whether or not the correct response is "yes" in a particular trial, e.g. whether the random object in the present trial is a positive ($A = 1$) or negative ($A = 0$) example of the unknown true concept. $P(\alpha)$ is shorthand

[4]In our model of the classic task, we assume that each observation x is of all three stimulus dimensions of an object. More generally, x could be an observation that results from querying all, or any subset of, stimulus dimensions.

for $P(A=1)$. The subscript t is shorthand to condition on the complete history of learning up to the present trial. For instance, $P_t(C)$ specifies the probability distribution that gives the learner's beliefs about the probability of each possible concept, after t blocks of learning; $P_t(c)$ is the probability of a *particular* concept c. The learner's belief that a particular unlabeled observation x is consistent with the true concept is

$$P_t(\alpha|x) = \sum_c P(\alpha|x,c)P_t(c).$$

In a given trial, the learner responds "yes" with probability $P_t(\alpha|x)$. The learner's probability of error for a particular observation x is the probability of a false positive or miss (Table 2)

$$P_t(\text{error}|x) = 2P_t(\alpha|x)[1 - P_t(\alpha|x)].$$

The model learner's overall probability of error is computed by averaging over all possible observations x

$$P_t(\text{error}) = \sum_x P(x)P_t(\text{error}|x).$$

Fig. 2, top panel, gives the model learner's error rates during learning, for Type I, II, IV, and VI concepts; the bottom panel gives error rates from Rehder and Hoffman's [39] data. The model learner shows fastest reduction in error rates for Type I concepts, consistent with Rehder and Hoffman's data. The model predicts that Type IV concepts will initially have lower error rates than Type II concepts. This phenomenon may result from the high similarity between Type IV and I concepts, such that above-chance performance on Type IV concepts can be obtained by responding in accordance with appropriate Type I concepts. It is hard to tell whether this prediction is consistent with the data, which are fairly noisy during early learning. The model predicts that Type II and IV concepts will have relatively similar error rates throughout learning, and that Type II concepts will be mastered first. These predictions appear consistent with the data. Finally, the model learner shows the slowest reduction in error for Type VI concepts, consistent with most empirical work. It does appear that the model learner underestimated the relative difficulty of Type VI concepts, compared with Type IV concepts. (However, Nosofsky and Gluck [31], did find similar error rates for Type IV and VI concepts, even late in learning.)

Table 2
Probability matching illustrated

	True state	
Response	$A=1$ $P_t(\alpha\|x)$	$A=0$ $1-P_t(\alpha\|x)$
"Yes" $P_t(\alpha\|x)$	Hit $[P_t(\alpha\|x)]^2$	False positive $P_t(\alpha\|x)$ $[1-P_t(\alpha\|x)]$
"No" $1-P_t(\alpha\|x)$	Miss $[1-P_t(\alpha\|x)]P_t(\alpha\|x)$	Correct rejection $[1-P_t(\alpha\|x)]^2$

Note: In each cell, the top line gives the event of interest; the bottom line gives its probability of occurrence. The learner believes the probability that the observation x is a positive instance of the true concept is $P_t(\alpha|x)$.

2.4.2. The model learner's beliefs about the true concept

An advantage of testing our model learner is that beliefs per se, as measured by the probability distribution $P_t(C)$, can be queried at each moment in learning, in addition to probability of correct response. Fig. 3 gives the posterior probability of the true concept, over the simulated learning blocks for four different concept classes. Before any learning has taken place, because the prior distribution is relatively peaked ($\lambda = 5$), almost all of the model learner's belief is concentrated in the six Type I concepts. (Entropy of prior beliefs in our model is 2.67 bits, barely larger than the 2.59 bits of a model with six equiprobable Type I concepts.) During the course of learning, the probability of the actual concept increases, irrespective of its type, following the ordering of Type I > II > IV > VI.

2.4.3. The model learner's uncertainty, throughout learning

While it is difficult to chart the model's beliefs about the probability of each of the 70 concepts at each point in learning, it is desirable to understand the model learner's

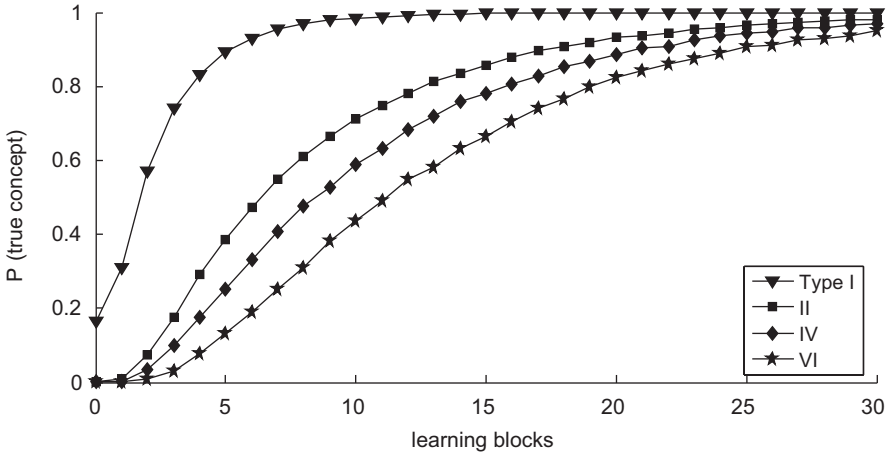

Fig. 3. The model learner's posterior probability of the correct concept.

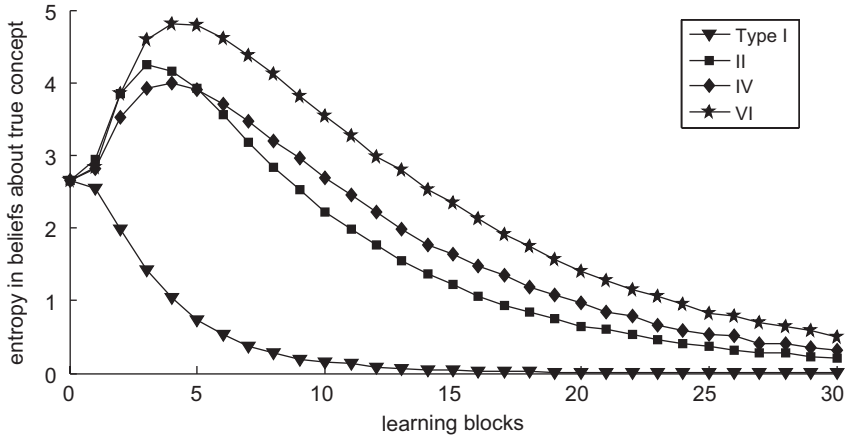

Fig. 4. Model learner's uncertainty (Shannon entropy) about the true concept, where the true concept is of the specified type.

subjective level of uncertainty over all the concepts, at each point. Shannon entropy [7,45] is well suited to quantifying this overall level of uncertainty. Fig. 4 shows the model learner's uncertainty about the true concept, according to the type of concept being learned (i.e. the curve for Type IV represents the uncertainty in the model learner's beliefs over all concepts, if the true concept is of Type IV).

$$H_t(C) = \sum_{c_i} P_t(c_i) \log_2 \frac{1}{P_t(c_i)}.$$

An interesting, unanticipated result is that for non-Type I concepts, the model learner's subjective uncertainty (Shannon entropy) actually increases in early learning. This increase in uncertainty is due to the large shift in probability mass (the model learner's beliefs) from the Type I concepts (which begin with combined probability 99.4%) to the concepts of other types. If the true concept is not Type I, then by five blocks of learning, non-Type I concepts (there are 64 of these, versus 6 concepts of Type I) will obtain a majority of the probability mass. Eventually, uncertainty decreases as the model learner narrows beliefs to the correct concept.

It would be interesting, and potentially feasible, to test our model's claim that uncertainty actually increases during early learning of non-Type I concepts. These tests should employ a modified task measuring knowledge of the true concept per se, rather than correct classification performance. With human subjects, these measures could potentially include subjective report, reaction time, or galvanic skin response. Another possibility would be to use fMRI to measure brain metabolic activity that correlates with entropy (or perhaps variance). Preuschoff et al. [37] found brain activity that correlated with uncertainty of monetary reward probability. An fMRI study could address whether uncertainty in a categorization task also correlates with brain activity. Bob Rehder (pers. commun., 2005) has noted his impression that subjects initially focus only on Type I concepts, and broaden their search if it later becomes clear that Type I concepts cannot account for the data.

3. Eye movements in concept formation

A primary motivation to develop a probabilistic model of learning on Shepard et al.'s [46] task was to evaluate whether theories of optimal experimental design (or optimal data selection) [3,27,34,43] could explain Rehder and Hoffman's [38,39] eye movement results. In this section, we describe Rehder and Hoffman's experiment and their main results. We also describe our eye movement model (formally, a utility function that quantifies possible eye movements' usefulness), and how its results may offer insight into Rehder and Hoffman's data. Note that although our concept-learning model includes all types of concepts that were considered by Shepard et al., we do not discuss model eye movement results for concept Types III and V, because Rehder and Hoffman did not obtain eye movement data from human subjects on concepts of those types.

3.1. Rehder and Hoffman's [38,39] task and eye movement data

Numerous models of category learning describe how selective attention to different stimulus dimensions evolves, yet little direct evidence is available. Rehder and Hoffman [39] devised a novel experimental design in which eye movements provided direct evidence for how selective attention was deployed at each moment in time. This was accomplished by representing the value of each stimulus dimension with abstract characters (such as $ or ¢ at the bottom right of Fig. 5) at a particular vertex of a large triangle on a computer screen, and recording eye movements throughout learning. In other words, rather than seeing a large black circle in the middle of the screen, subjects might be presented with an "x" in the bottom left, an "!" in the top middle, and a "$" in the bottom right.

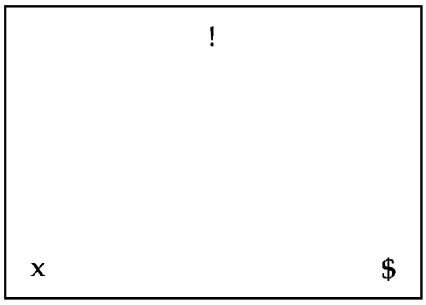

Fig. 5. Example stimulus from Rehder and Hoffman's [39] experiment. Reprinted from Fig. 2 in Rehder and Hoffman [39]. © 2005 Elsevier. Reprinted with permission.

A small black circle might be represented by having an "o" in the bottom left, and the other characters as before.

Among 72 subjects, a variety of patterns were observed. For the vast majority of participants, the following findings held:

(1) Early in learning, all three stimulus dimensions were viewed (fixated).
(2) Improvement in categorization performance occurred before subjects switched to viewing just the dimensions necessary for correct classification ("efficient" eye movements). For instance, the modal pattern for Type I learners was for error to drop from 50% (chance) to almost 0% after about 12 trials (1.5 blocks), but for the number of dimensions fixated to remain high until about 16 trials (2.0 blocks). ("Type I learner" means a subject who was assigned to learn a Type I concept.) For Type II learners, errors decreased more gradually, beginning approximately on trial 40 (5 blocks), and decreasing gradually through approximately trial 72 (9 blocks).
(3) Eventually, *efficient* eye movements (to only the necessary stimulus dimensions, given certain knowledge of the true concept) were used. For Type I concepts, efficient eye movements are to a single dimension, such as the size query for the concept "small". Rehder and Hoffman's Type I subjects in general fixated only the single relevant dimension, late in learning. For Type II concepts, efficient eye movements are to the relevant two dimensions, such as the size–shape query for the concept "large square or small circle." Type IV concepts, in the general case, require all three dimensions. In some cases, however, depending on what is learned from the first dimension, Type IV concept objects can be classified with two dimensions. Type IV subjects tended to fixate all three dimensions, although the average number of dimensions fixated by Type IV subjects was slightly less than the number of dimensions fixated by Type VI subjects (Fig. 6). The finding that Type IV subjects tended to fixate all three dimensions, even late in leaning, could suggest that subjects found it more expeditious to plan a sequence of eye movements in advance (e.g. a remembered scanpath or visual routine), rather than to fixate a single dimension and then consider what dimension to fixate next. Finally, Type VI concepts always require all three stimulus dimensions; the vast majority of subjects always considered all three dimensions.

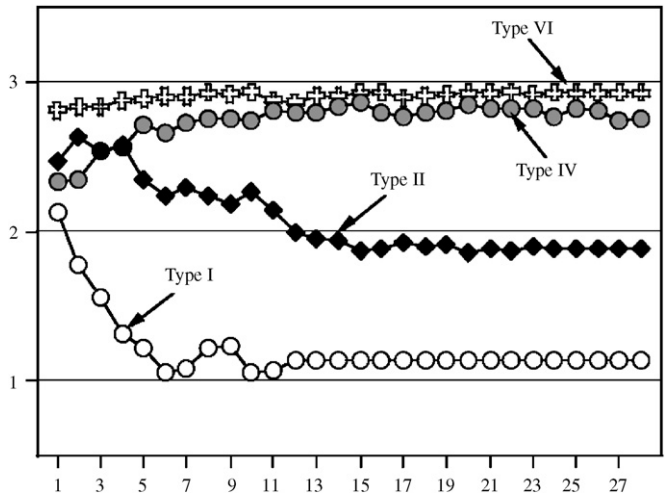

Fig. 6. Number of stimulus dimensions fixated (y-axis), during different blocks of learning (x-axis). Reprinted from Fig. 3 in Rehder and Hoffman [39]. © 2005 Elsevier. Reprinted with permission.

We will focus on these empirical findings when evaluating our eye movement model.

3.2. Quantifying eye movements' usefulness

We model the history of learning with the assumption that the learner has viewed all three stimulus dimensions, at each trial up to the present time. This matches the classic task, but is a simplification in modeling Rehder and Hoffman's [39] eye movement version, in which subjects sometimes do not view all stimulus dimensions, even early in learning. At each point in learning, our concept-learning model is used to describe the model learner's beliefs. Technically, the eye movement model is a subjective utility function for evidence acquisition, in the sense of Savage [43]. When we refer to the *usefulness* of each possible eye movement, we mean the subjective utility that the model learner expects to obtain, on average, as a result of making that eye movement. In the equations below we refer to queries, rather than eye movements per se, to reflect that a task could be designed so that information is obtained by other means, for instance by asking a question, clicking a mouse pointer on a computer screen, etc.[5]

[5]Some eye movement models, for instance the E-Z Reader models of reading [40,41] explicitly account for oculomotor processes, sub-second timing issues, saccade (eye movement) length distributions, and specific properties of visual acuity at different eccentricities (angular distances from center of gaze). Our model is relatively abstract. We disregard issues of peripheral acuity, because Rehder and Hoffman's [39] stimuli were designed to render peripheral information uninformative, except with respect to stimulus position. We disregard the visual system's preference for relatively short eye movements, because the targets—at the vertices of

(Matsuka and Corter [22], nicely show how mouse clicks can be used to study attention allocation in category learning.) At each point in learning, the learner picks between 8 possible queries: all dimensions, size–shape, size–color, color–shape, size, color, shape, or null. The task for modeling is to see whether a principled utility function, inspired by ideas from theories of optimal experimental design, can offer insight into subjects' eye movements in Rehder and Hoffman's [39] data. We propose the following function to quantify a particular query's expected usefulness:

$$u_t(Q) = pg_t(Q) + I_t(Q; C) - j(Q).$$

(Recall that the subscript t is shorthand to condition on the complete history of learning up to the present trial. C is a random variable denoting beliefs about the true concept. Q is a random variable that denotes a particular query, such as size, whose outcome is unknown, and q represents one of several possible outcomes of that query.) Each term in the usefulness function corresponds to a central idea, as introduced below:

1. $pg_t(Q)$. The learner wishes to correctly classify the randomly selected object that is presented in a particular trial (exploitation).
2. $I_t(Q; C)$. The learner wishes to learn the true concept (exploration).
3. $-j(Q)$. The learner wishes to avoid making extraneous eye movements.

The next sections explain the terms of the usefulness function in more detail, and describe the rationale behind each term.

3.2.1. Classifying a random object: $pg_t(Q)$

The learner's immediate task is correct classification on a particular trial. The eye movement model's usefulness function captures this idea (Baron, 1981, as cited in [3]) with its $pg_t(Q)$ term. This term quantifies the amount that a query is expected to improve the probability of correct response on a particular trial, the query's *probability gain*. Because objects are drawn randomly with equal probability, and 4 of the 8 objects are consistent with each concept, chance accuracy is 0.5. Formally

$$pg_t(Q) = \sum_q P_t(\text{correct}|q)P(q) - 0.5.$$

In the above equation, Q is shorthand for any particular query that the learner might make, for example Q_{size}, in which case

$$pg_t(Q_{\text{size}}) = \sum_{q\in\{\text{large,small}\}} P_t(\text{correct}|q)P(q) - 0.5.$$

The probability of correct response, given a particular outcome of a query, is

$$P(\text{correct}|q) = \max[P(\alpha|q), 1 - P(\alpha|q)],$$

where $P(\alpha|q)$ gives the probability that the correct response on the trial is "yes", given q. The $P(\text{correct}|q)$ term is derived from an optimal (not probability matching) response strategy, in which the learner always responds "yes" if $P(\alpha|q) > 0.5$, and "no" if $P(\alpha|q) \leqslant 0.5$.

Suppose that at time t, late in learning, the true concept were known with probability 1 to be "large," and that the learner were evaluating the probability gain of the size query (Q_{size}). If the observation (q) were that the size dimension takes the value "small," then $P(\alpha|q) = 0 < 0.5$, so the learner would respond "no," which results in 100% probability of being correct. If the observation q were "large," $P(\alpha|q) = 1 > 0.5$, and the learner would respond "yes," again with 100% probability of being correct. Before making an eye movement, the learner averages over each possible result of the query, in this case each with probability 0.50 of occurrence, and obtains $pg_t(Q_{\text{size}}) = 0.50$.

3.2.2. Learning the true concept: $I_t(Q; C)$

The second term in the usefulness function, $I_t(Q; C)$, quantifies a query Q's value for reducing uncertainty about the true concept C. We use the mutual information between the query Q and the concept C, given learning to date t, to measure this

$$\begin{aligned} I_t(Q; C) &= H_t(C) - H_t(C|Q) \\ &= \sum_c P_t(c)\log_2 \frac{1}{P_t(c)} \\ &\quad - \sum_q P_t(q) \sum_c P_t(c|q) \log_2 \frac{1}{P_t(c|q)}. \end{aligned}$$

This term is an exploration term to quantify the anticipated future usefulness of what is learned in the present trial. It also addresses the subject's presumed desire to identify the true concept. Technically, mutual information is the expected reduction in entropy in C from making query Q, or the initial entropy $H_t(C)$ less the conditional entropy $H_t(C|Q)$. Note that $H_t(C|Q)$ is the expected entropy given a query Q (such as size), not the actual entropy given a particular result of the query q (such as small).

Sometimes called 'expected information gain,' the use of mutual information to quantify experiments' (or queries') usefulness was proposed by Lindley [20]. Oaksford and Chater [34,35,56] used information gain in a model of Wason's [58,59] card selection task. Oaksford and Chater [34] noted that mutual information is equivalent to the directional Kullback–Liebler distance [18] from prior to posterior distribution, as well. Lee and Yu [19] suggested using mutual information to predict eye movements. Denzler and Brown [8] used information gain to control cameras in a robotic vision task. Renninger et al. [42] did so in a shape-learning task model, as did Movellan [23] in a

(*footnote continued*)
a triangle—are roughly equidistant from the current point of gaze. Timing issues would be interesting to consider in future work, such as whether longer fixations (periods of focused gaze) tend to be to more informative features.

social contingency detection model, and Steyvers et al. [48] in a causal network learning task. Nelson [27] reviews some alternate ways an exploration term could be formulated, and shows that alternate formulations give equivalent results in several psychological tasks.

3.2.3. Avoiding extraneous eye movements: $j(Q)$

Finally, we include a cost term, $j(Q)$, to represent the subjective cost of fixating one or more stimulus dimensions. This cost does not depend on the history of learning t. We set $j(Q)$ arbitrarily, to 0.04 per dimension fixated. Hence, Q_{all}, the size-shape-color query, has cost 0.12, and Q_{size} has cost 0.04. We have simulated the model with various settings of this parameter. Results do not greatly depend on the specific value chosen, so long as it is small and nonzero.

3.3. Results from the eye movement model

What eye movements does the model predict, at particular stages of learning? In each trial, there are 8 possible queries: size–shape–color, size–shape, size–color, color–shape, size, color, shape, and null. Before learning, the query to all stimulus dimensions (utility = 0.880) is more useful than any query of two (utility = 0.585) or one (utlity = 0.292) stimulus dimensions, consistent with Rehder and Hoffman's [39] empirical finding that early in learning, learners tend to fixate all stimulus dimensions. This is due to the queries' different $I_t(C; Q)$, their expected information gain with respect to the learner's beliefs about the true concept. Because no learning has taken place, this is true irrespective of the true concept. Before learning, $pg_t(Q)$ is zero for every query. (The learner's lack of knowledge of the true concept is so pronounced that learning, for instance, that the object presented in this trial is a large white circle does not help the learner do better than chance performance of 50% on this trial.) Fig. 7 illustrates how $I_t(C; Q)$, the open circles, and $pg_t(Q)$, the closed circles, change during learning of the concept "small", comparing the query to all stimulus dimensions (left panel), and the size query (right panel). Comparing the open circles on the two graphs, the $I_t(C; Q)$ term always favors the query to all stimulus dimensions, relative to the query to the size dimension alone. During learning, however, the $I_t(C; Q)$ terms tend toward zero, such that the difference between $I_t(C; Q_{\text{all}})$ and $I_t(C; Q_{\text{size}})$ is miniscule. Because of this, eventually the size query, which has the lower cost, becomes more valuable (has higher utility, or usefulness) than the query to all dimensions (Fig. 9). In other words, the model initially focuses on learning the correct concept, but as the correct Type I concept quickly becomes known, the model shifts its focus to producing a correct response.

A similar pattern holds for Type II concepts (Fig. 8). We discuss results for the Type II concept "large square or small circle." For this concept, the efficient query ($Q_{\text{efficient}}$) is to the size and shape dimensions. Type II concepts have a smaller initial difference in $I_t(C; Q_{\text{all}})$ and $I_t(C; Q_{\text{efficient}})$ than do Type I concepts. This is because a query to two stimulus dimensions is expected to provide more information than a query to a single stimulus dimension. The slower learning of Type II concepts, relative to Type I concepts, results in slower decrease of the $I_t(C; Q)$ terms and increase of the $pg_t(Q)$ terms.

For Types IV and VI concepts (not shown), the model always prefers the query to all stimulus dimensions.

3.4. Putting it all together

We now have results from both the concept-learning model and the eye movement model in hand, so we can examine their behavior together. This is the critical test to evaluate whether our model offers insight into Rehder and Hoffman's [39] data. Fig. 9, for the Type I concept, "small," shows that the error rate (dashed line) decreases

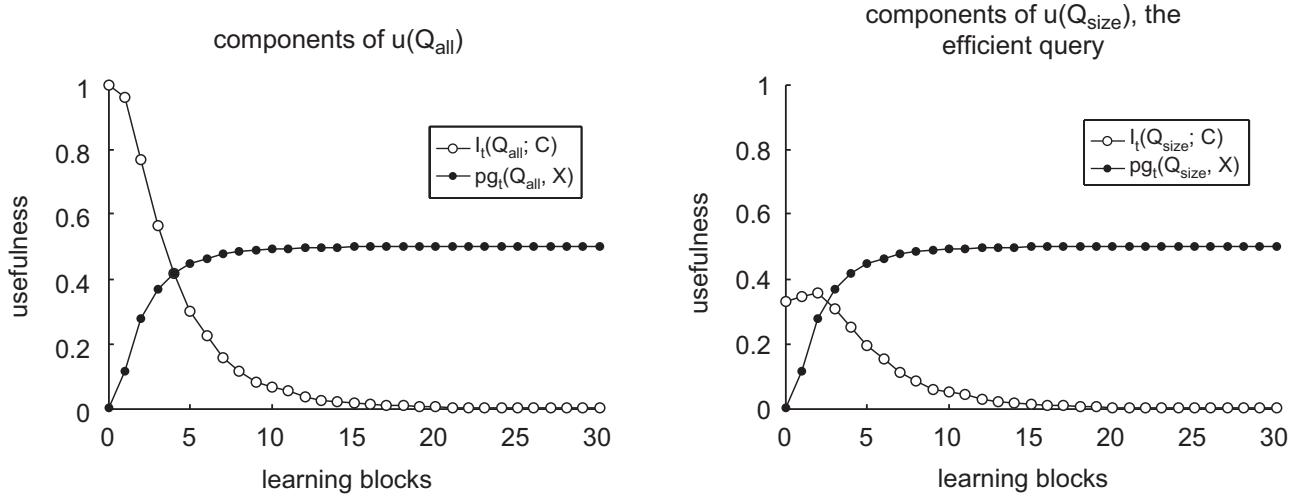

Fig. 7. Components of the usefulness movement function for the concept "small." At left, Q_{all}, the query to all stimulus dimensions; at right, Q_{size}, the efficient size query.

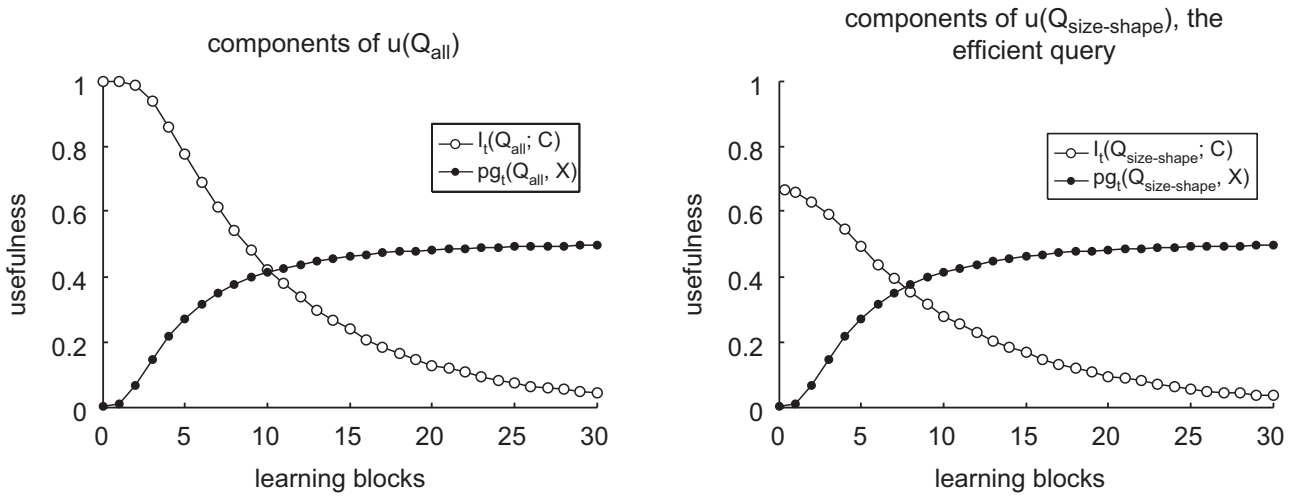

Fig. 8. Components of usefulness function for the Type II concept "large square or small circle." Left panel, the query to all stimulus dimensions; Right panel, the efficient query, to size and shape.

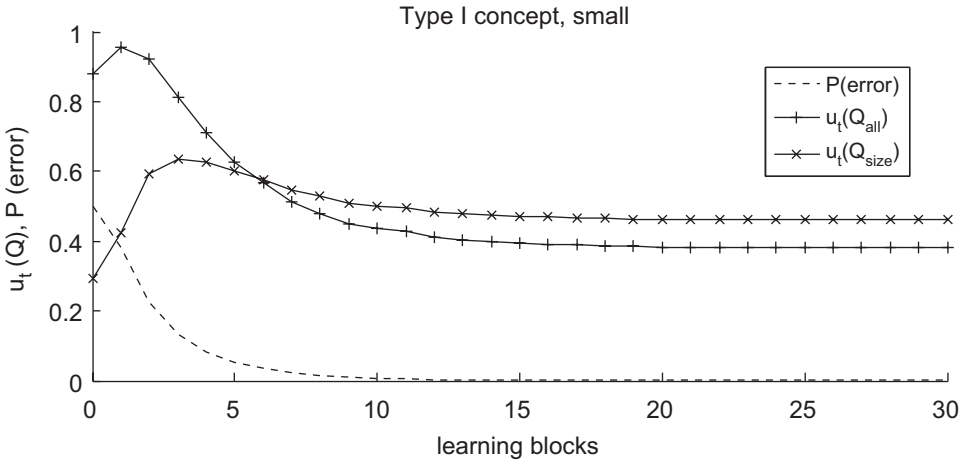

Fig. 9. Error rate and selected queries' usefulness, Type I concept, "small."

well before the efficient query (x marker) becomes more valuable than the query to all dimensions (+ marker).

Fig. 10 shows that the model behaves in a similar manner on Type II concepts. Importantly, however, because learning develops more slowly in the case of Type II concepts, the error rates decrease more slowly, and the point when the efficient query becomes more valuable occurs much later, than in the case of Type I concepts. For Type IV and VI concepts, for which Rehder and Hoffman's [39] subjects tended to fixate all stimulus dimensions throughout learning (Fig. 6), the model also rates the query to all dimensions as having the highest usefulness throughout learning.

Together, these results show that the model produces the main qualitative features that Rehder and Hoffman [39] noted in their data. In each case, the model offers a principled explanation of the result:

1. Initially, subjects consider all stimulus dimensions, because Q_{all} provides more information than any other query about the true concept, as quantified by the $I_t(Q, C)$ term. The greater usefulness with respect to the $I_t(Q, C)$ term more than overcomes the higher cost of Q_{all}, versus other queries.
2. Error rates decrease during the course of learning, even as subjects continue to fixate all stimulus dimensions, because correct classification performance results from beliefs about several possible concepts, and not just about the concept that later proves correct.
3. Subjects switch from Q_{all} to $Q_{\text{efficient}}$ after achieving high classification performance, because even though high classification performance is being achieved already, it takes additional learning to rule out plausible alternate hypotheses about the true concept.

3.5. Questions about the usefulness function

A number of points of confusion about our formulation of the usefulness function may arise. Are all the terms necessary? Could some terms be formulated differently? This section, formulated as a hypothetical question-and-answer session, discusses several points.

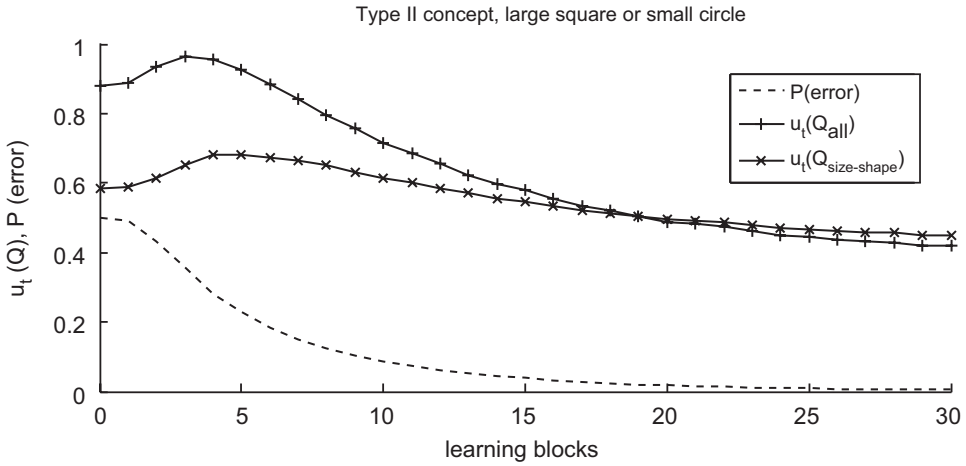

Fig. 10. Error rate and selected queries' usefulness, Type II concept, "large square or small circle."

Isn't there a correct way to quantify queries' usefulness? For the sake of argument, suppose that the task included an explicitly specified set of external rewards. For instance, suppose that the learner is paid $1 for each trial in which they guess correctly, up to a total of 300 trials. In this case, there is a unique, correct calculation of each query's usefulness, at each point in learning, and dynamic programming could in principle be used to identify what utility function applies at each point in learning, to maximize net total payoff (Javier Movellan, pers. commun., 2005). Modeling the subjective utility (the perceived value to the subjects) of money, however, would require separate specification of subjects' utility function for money, as well as of the temporal discount factor that subjects have for future reward. Furthermore, external rewards are not explicitly specified on Rehder and Hoffman's [39] task, so different people could have different ideas about what function implicitly applies. Similarly, there is no guarantee that introduction of external rewards would eliminate intrinsic goals that the subject might have, such as identifying the true concept. Accordingly, we view the current modeling objective as ascertaining whether ideas from optimal experimental design can help describe subjects' goals on Rehder and Hoffman's task, rather than as ascertaining whether subjects have a theoretically correct model of the value of information on the task.

What if the $pg_t(Q)$ term were eliminated? Intuitively, it may seem that the learner's goal is to learn the true concept, and that this goal is served by the $I_t(C, Q)$ term. Early in learning, if the $pg_t(Q)$ term were eliminated, the model concept learner would behave similarly, because the main driver for the query to all stimulus dimensions is its ability to provide information about the unknown true concept. Late in learning, however, once the true concept is known with high certainty, the model concept learner, rather than relying on the efficient query relative to the true concept, would stop making eye movements altogether (switching to the null query), because the cost of any query would more than outweigh the usefulness of its miniscule expected reduction in uncertainty about the true concept. The paradoxical result would be that as the learner comes to know the true concept with high probability, as attested by the probabilities of each concept in the model, performance on the categorization task would decrease to chance.

What if the $I_t(C, Q)$ term were eliminated? This would cause the learner to have to rely on the $pg_t(Q)$ term to identify useful eye movements. Unfortunately, in the first trial of learning it is impossible to classify the object with greater than chance (50%) accuracy, irrespective of the number of stimulus dimensions fixated. Thus, because of each query's cost of execution, every query (except the null query) would have negative usefulness, and the learner would not fixate any of the stimulus dimensions. In the second learning trial, the learner's beliefs about the concepts would be the same as the learner's initial beliefs, because no learning took place on the previous trial. Hence, the same situation would arise. The problem would iterate indefinitely, and at no point in time would the learner fixate any stimulus dimensions, or achieve greater-than-chance performance on the task.

What if the $j(Q)$ term were eliminated, by setting the cost of each query to zero? Consider what would happen when, late in learning, uncertainty is small and the true concept is known with nearly 100% certainty. Even at this point, the learner would fail to switch from fixating all stimulus dimensions to the efficient query, because Q_{all} would be marginally more valuable at reducing the miniscule remaining uncertainty about the true concept (Fig. 11). This behavior would make sense if eye movements were free. However, every eye movement (saccade) has a biomechanical cost to execute. There is also an opportunity cost in the other eye movements that must be foregone, and the inability to perceive any visual stimuli due to retinal blurring during each saccadic eye movement's execution.

Could the $I_t(Q, C)$ term be formalized differently? Several usefulness functions for quantifying the value of information have been proposed. Nelson [27] provided

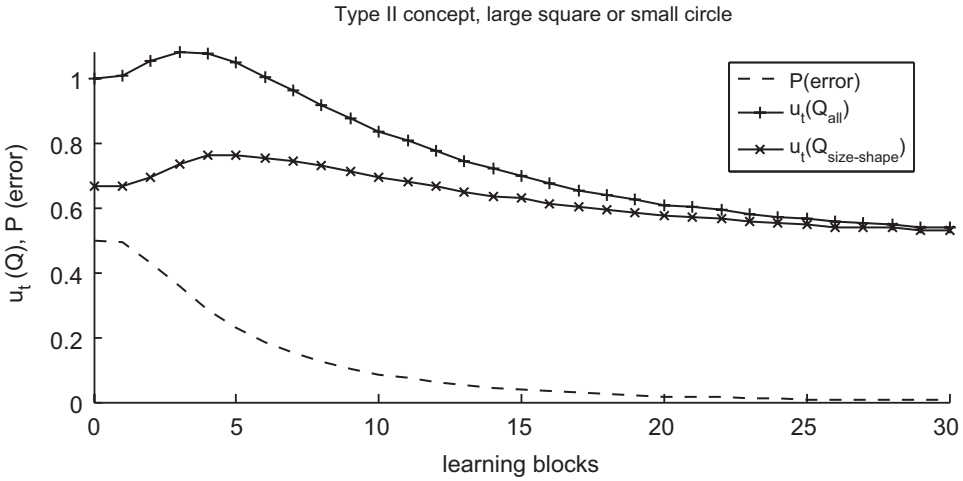

Fig. 11. If cost j is zero, Q_{all} is the most useful query, throughout learning.

experimental evidence that some usefulness functions from the literature (Bayesian diagnosticity and log diagnosticity) are not plausible models of human intuitions, and behave poorly in several situations. However, based on a review of the literature, Nelson concluded that several other usefulness functions are roughly equally reasonable as descriptive models of empirical data from human subjects. These functions include mutual information, which we use here, as well as probability gain and impact. If probability gain were to be used as the basis of an exploitation term, it would quantify the expected improvement of correctly identifying the true *concept*. (Note that this is different from the improvement in probability of correct response on the current trial, which is already quantified in the current eye movement model's $pg_t(Q)$ term.) Impact, the expected absolute change in beliefs [15,30,60] is another candidate. Nelson reported simulations suggesting that these three usefulness functions in general behave similarly, in a two-category, binary-feature scenario.[6] Other plausible usefulness functions could also be proposed. Our theoretical goal is to show that a reasonable usefulness function can offer insight onto Rehder and Hoffman's [39] eye movement data, not to argue for information gain versus other measures.

Should a term for intrinsic visual salience [12,13,53] be included? The justification for a salience term would be that certain kinds of visual stimuli (such as high color contrast areas) might be intrinsically interesting to the visual system, irrespective of one's current task. In Rehder and Hoffman's [39] task, the values of the stimulus dimensions of the objects were the only illuminated parts of the screen during each trial. A salience-based eye movement model alone could not model Rehder and Hoffman's eye movement data, namely the different patterns of eye movements observed for the different stimulus types, as the visual stimuli were the same for all types of concepts. The other terms are each necessary to model Rehder and Hoffman's eye movement data. For the sake of parsimony, we ignore visual salience in the present model, although it could potentially be important on other tasks.

What likelihood does the model learner think they will use when evaluating each possible query's usefulness? The current model makes the simplifying assumption that when evaluating possible queries' (eye movements') anticipated usefulness, $\mu = 1$. The implicit psychological claim here is that when considering what eye movements to make, the learner believes he or she will have complete confidence in the results. In future work we intend to explore the implications of using the same likelihood function as in the concept learning model, where $\mu < 1$, when calculating possible queries' usefulness.

4. Summary, discussion, and future work

Shepard et al.'s [46] task has received a great deal of study. Like much work in cognitive psychology, process models have largely inspired theories of the task. Many of these models have made claims about the learner's selective attention to different stimulus dimensions throughout learning, in the absence of selective attention data. Rehder and Hoffman's [38,39] insight was to devise an eye movement-based version of the task, so that the viewer's eye movements could provide direct evidence for how selective attention is deployed throughout learning.

Rehder and Hoffman noted that their eye movement data did not appear to fit existing models well. RULEX, a rule-based category-learning model [33] would predict eye movements to one dimension at a time, early in learning. Subjects, however, tended to fixate all stimulus dimensions early in learning. ALCOVE, a connectionist exemplar-based model [17] would predict a gradual transition

[6] Jiri "George" Najemnik (pers. commun., 2005) noted that although Najemnik and Geisler [26] used probability correct, which is equivalent to probability gain, in an ideal observer analysis of a visual search task, they obtain similar results using mutual information, and that the two functions are too similar for determination of which of them best models human data.

between eye movements to all stimulus dimensions, and eye movements to the relevant dimensions. However, subjects showed a relatively sudden switch to efficient eye movements, which occurred after error rates were already low. Rehder and Hoffman [39] hypothesized that multiple learning modules, some rule-based and some similarity-based, control learning and selective attention on their task. Rehder and Hoffman [39] suggested that perhaps the late switch to efficient eye movements happens because subjects strategically "abandon alternative learning strategies after one module has solved the learning problem," or because of "cognitive limits that influence how objects get prioritized in a visual field, or both." Our results show that recourse to such a complicated set of explanations of Rehder and Hoffman's data may not be necessary, even though some aspects seem consistent with ALCOVE, and other aspects with RULEX.

There were two main goals underlying the research discussed in this article. The first was to evaluate whether a probabilistic model that describes learning as optimal Bayesian inference could help explain Shepard et al.'s task. It appears that the Bayesian concept learning model at least qualitatively matches key properties in data from Shepard et al.'s task, such as rapid reduction in error for Type I concepts, similar error rates for Types II and IV concepts through much of learning, and the expected ordering of concept types in number of trials to criterion performance. Our second goal was to evaluate whether, given that Bayesian learning model, eye movements (selective attention) on Rehder and Hoffman's [38,39] task could be modeled as a process of optimal experimental design. We view our eye movement modeling results as a proof of concept that a usefulness (utility) function that is inspired by principled statistical ideas can reproduce important properties of human behavior. A model with an unchanging set of *goals* (the usefulness function), inspired by ideas from the theory of optimal experimental design, and with *beliefs* that develop in a principled way, modeled with Bayesian statistics, produces the key findings that Rehder and Hoffman [39] reported: early fixations to all stimulus dimensions, the gradual reduction in classification error, and the late shift to efficient eye movements.

This article reports work in progress. In the future we hope to address several issues:

1. What is the source of prior beliefs? Here we primarily focused on identifying what the learner's beliefs are, at each point in learning, and on how those beliefs change during learning. We therefore used error counts from Rehder and Hoffman's [39] data as a basis for assigning prior probabilities. A priori statistical measures, such as Boolean complexity [10], could perhaps also be used in assignment of prior probabilities. It would be remarkable if subjects' intuitions about different concepts' relative plausibility could be explained in such a principled manner.
2. What range of eye movement (utility) functions can offer insight into Rehder and Hoffman's [39] data? We chose a usefulness function that seemed plausible, based on the task and on other research using optimal experimental design concepts to understand human intuitions (reviewed by Nelson [27]). A variety of other formulations of the usefulness function could be considered. For instance, the learner's goal of identifying the correct concept, which we quantified with mutual information (information gain), $I_t(C, Q)$, might alternately be quantified with impact (absolute change in beliefs impact, [15,30,60]), or with probability gain [3]. It would also be helpful to more thoroughly consider the explore/exploit tradeoff. Exploration, in our model, is quantified by $I_t(C, Q)$, and exploitation is quantified by $pg_t(Q)$. In the present model, we simply added those terms, and did not optimize their relative weight. However, subjects may place more weight on one of these terms. In general, we sought to be conservative in not optimizing every aspect of model construction, to avoid overfitting. In the future, we hope that minimum description length [36] or other techniques may enable us to consider a wider range of models, while insuring against overfitting empirical data.
3. What more can be learned from a more fine-grained analysis of the data? Rehder and Hoffman [39] reported that their subjects could be clustered according to their eye movements and error data. Optimizing model parameters to subgroups of subjects (see [57]), or even to individual subjects, could potentially inform discussion of the subjects and the model alike. Another important analysis of data would be a backward stationarity analysis of errors on individual stimulus items, to see whether (or the extent to which) individual items show gradual decrease in error (Steve Link, pers. commun., 2005).

The models introduced in this article contribute to a growing body of research in rational analyses of cognitive tasks [1,2, pp. 409–410]. This line of research is analogous to ideal observer analyses of perceptual tasks [16,24]. It was presaged by Brunswik's [5] molar analysis and Marr's [21] computational analysis; others have recently discussed it as well [6,25]. The focus is on the nature of problems that cognition solves, on what would constitute optimal solutions to those problems, and on whether those optimal solutions might approximate human cognition. Our results, of course, do not demonstrate what architecture the brain uses to learn categories, to categorize objects, or to deploy selective attention via eye movements. However, our results demonstrate that a single model, with a small number of parameters, based on Bayesian statistics and theory of optimal experimental design, can (1) provide a parsimonious account of classic concept learning results, and (2) help explain the central findings that Rehder and Hoffman noted in their compelling eye movement data.

In life, virtually every interesting learning or inference scenario involves some active contribution of the learner in deciding where to look, what questions to ask or experiments to conduct, or what data to attend to. Optimal experimental design-inspired models of evidence acquisition, however, have seldom considered tasks where beliefs change so drastically, as information is sought out, decisions are made, and feedback is given, over dozens or hundreds of trials. The present model suggests that unified, rational accounts of learning and information acquisition may apply to a wide range of behavior. Our model contributes to a growing body of work using information- and decision-theoretic concepts to model perceptual tasks [19,23,26,42,47,52]. We hope it may also spur work to explore the relationship between active information acquisition in perceptual, such as eye movement, and other learning and inference tasks.

Acknowledgments

Bob Rehder and Aaron Hoffman kindly corresponded about their experiment and findings, and provided their data to us. Bob Rehder, Javier Movellan, Michael Lee, Tim Marks, Flavia Filimon, and two anonymous reviewers provided feedback on a draft of this manuscript, and numerous ideas on the research and its explication. Nathaniel Smith, Josh Tenenbaum, Eric Wiewiora, and Danke Xie provided helpful suggestions on this work. JDN was funded by NIMH Grant 5T32MH020002-05 to T. Sejnowski and by NSF Grant DGE 0333451 to GWC. GWC is supported by NIH Grant MH57075. This research was conducted while Jonathan Nelson was a graduate student at the Cognitive Science Dept., UCSD. An earlier version of this work was presented by Nelson et al. [29]. New developments in this research will be posted at http://www.jonathandnelson.com/.

References

[1] J.R. Anderson, The Adaptive Character of Thought, Erlbaum, Hillsdale, NJ, 1990.
[2] J.R. Anderson, The adaptive nature of human categorization, Psychol. Rev. 98 (1991) 409–429.
[3] J. Baron, Rationality and Intelligence, Cambridge University Press, Cambridge, 1985.
[4] T. Bayes, An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. Lond. 53 (1763) 370–418.
[5] E. Brunswik, The Conceptual Framework of Psychology, University of Chicago Press, Chicago, 1952.
[6] N. Chater, M. Oaksford, Ten years of the rational analysis of cognition, Trends Cognitive Sci. 3 (2) (1999) 57–65.
[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[8] J. Denzler, C.M. Brown, Information theoretic sensor data selection for active object recognition and state estimation, IEEE Trans. Pattern Anal. Machine Intell. 24 (2002) 145–157.
[9] W. Edwards, Conservatism in human information processing, in: B. Kleinmuntz (Ed.), Formal Representation of Human Judgment, Wiley, New York, 1968, pp. 17–52.
[10] J. Feldman, Minimization of Boolean complexity in human concept learning, Nature 407 (2000) 630–633.
[11] J. Feldman, The simplicity principle in human concept learning, Curr. Direct. Psychol. Sci. 6 (2003) 227–232.
[12] L. Itti, P. Baldi, A principled approach to detecting surprising events in video, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[13] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Machine Intell. 20 (11) (1998) 1254–1259.
[15] J. Klayman, Y.-W. Ha, Confirmation, disconfirmation, and information, Psychol. Rev. 94 (1987) 211–228.
[16] D.C. Knill, W. Richards (Eds.), Perception as Bayesian Inference, Cambridge University Press, Cambridge, UK, 1996.
[17] J.K. Kruschke, ALCOVE: an exemplar-based connectionist model of category learning, Psychol. Rev. 99 (1992) 22–44.
[18] S. Kullback, R.A. Liebler, Information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86.
[19] T.S. Lee, S.X. Yu, 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, MIT Press, Cambridge, MA, 2000, pp. 834–840.
[20] D.V. Lindley, On a measure of the information provided by an experiment, Ann. Math. Stat. 27 (1956) 986–1005.
[21] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W. H. Freeman, San Francisco, 1982.
[22] T. Matsuka, J. Corter, Process tracing of attention allocation in category learning, submitted.
[23] J.R. Movellan, An infomax controller for real time detection of contingency, in: Proceedings of the International Conference on Development and Learning, Osaka, Japan, July, 2005.
[24] J.R. Movellan, J.L. McClelland, The Morton–Massaro law of information integration: implications for models of perception, Psychol. Rev. 108 (1) (2001) 113–148.
[25] J.R. Movellan, J.D. Nelson, Probabilistic functionalism: a unifying paradigm for the cognitive sciences, Behav. Brain Sci. 24 (2001) 690–692.
[26] J. Najemnik, W.S. Geisler, Optimal eye movement strategies in visual search, Nature 434 (17 March 2005) 387–391.
[27] J.D. Nelson, Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain, Psychol. Rev. 112 (4) (2005) 979–999.
[28] J.D. Nelson, J.B. Tenenbaum, J.R. Movellan, Active inference in concept learning, in: J.D. Moore, K. Stenning (Eds.), Proceedings of the 23rd Conference of the Cognitive Science Society, Erlbaum, Mahwah, NJ, 2001, pp. 692–697.
[29] J.D. Nelson, G.W. Cottrell, J.R. Movellan, Explaining eye movements during learning as an active sampling process, in: International Conference on Development and Learning, October 2004.
[30] R.S. Nickerson, Hempel's paradox and Wason's selection task: logical and psychological puzzles of confirmation, Think. Reason. 2 (1996) 1–32.
[31] R.M. Nosofsky, M.A. Gluck, Adaptive networks, exemplars, and classification rule learning, in: Paper presented at the 30th Annual Meeting of the Psychonomic Society, Atlanta, 1989 [cited in Anderson, 1991].
[32] R.M. Nosofsky, M. Gluck, T.J. Palmeri, S.C. McKinley, P. Glauthier, Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961), Memory Cognit. 22 (1994) 352–369.
[33] R.M. Nosofsky, T.J. Palmeri, S.C. McKinley, Rule-plus-exception model of classification learning, Psychol. Rev. 101 (1994) 53–79.
[34] M. Oaksford, N. Chater, Rational explanation of the selection task, Psychol. Rev. 103 (1996) 381–391.
[35] M. Oaksford, N. Chater, A revised rational analysis of the selection task: exceptions and sequential sampling, in: M. Oaksford, N. Chater

(Eds.), Rational Models of Cognition, Oxford University Press, Oxford, 1998, pp. 372–393.
[36] M.A. Pitt, I.J. Myung, S. Zhang, Toward a method of selecting among computational models of cognition, Psychol. Rev. 109 (2002) 472–491.
[37] K. Preuschoff, P. Bossaerts, S.R. Quartz, Neural differentiation of expected reward and risk in human subcortical structures. Neuron. 51, (2006), 381–390. DOI:10.1016/j.neuron.2006.06.024.
[38] B. Rehder, A.B. Hoffman, Eyetracking and selective attention in category learning, in: R. Alternman, D. Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society, Cognitive Science Society, Boston, MA, 2003.
[39] B. Rehder, A.B. Hoffman, Eyetracking and selective attention in category learning, Cognitive Psychol. 51 (2005) 1–41.
[40] E.D. Reichle, A. Pollatsek, D.L. Fisher, K. Rayner, Toward a model of eye movement control in reading, Psychol. Rev. 105 (1998) 125–157.
[41] E.D. Reichle, K. Rayner, A. Pollatsek, The E–Z reader model of eye-movement control in reading: comparisons to other models, Behav. Brain Sci. 26 (2003) 445–526.
[42] L.W. Renninger, J. Coughlan, P. Verghese, J. Malik, An information maximization model of eye movements, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 1121–1128.
[43] L.J. Savage, The Foundations of Statistics, Wiley, New York, 1954.
[44] D.R. Shanks, R.J. Tunney, J.D. McCarthy, A re-examination of probability matching and rational choice, J. Behav. Decision Mak. 15 (2002) 233–250.
[45] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[46] R.N. Shepard, C.I. Hovland, H.M. Jenkins, Learning and memorization of classifications, Psychol. Monogr.: Gen. Appl. 75 (13) (1961) 1–42.
[47] N. Sprague, D. Ballard, Eye movements for reward maximization, in: S. Thrun, L.K. Saul, B. Scholkpf (Eds.), Advances in Neural Information Processing Systems, vol. 16, MIT Press, Cambridge, MA, 2003.
[48] M. Steyvers, J.B. Tenenbaum, E.-J. Wagenmakers, B. Blum, Inferring causal networks from observations and interventions, Cognitive Sci. 27 (2003) 453–489.
[49] J.B. Tenenbaum, A Bayesian framework for concept learning, Ph.D. Thesis, MIT, 1999.
[50] J.B. Tenenbaum, 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, MIT Press, Cambridge, MA, 2000, pp. 59–65.
[51] J.B. Tenenbaum, T.L. Griffiths, Generalization, similarity, and Bayesian inference, Behav. Brain Sci. 24 (4) (2001) 629–640.
[52] J. Trommershäuser, L.T. Maloney, M.S. Landy, Statistical decision theory and the selection of rapid, goal-directed movements, J. Opt. Soc. Am. A 20 (7) (2003) 1419–1433.
[53] K. Yamada, G.W. Cottrell, A model of scan paths applied to face recognition, in: Proceedings of the Seventeenth Annual Cognitive Science Conference, Pittsburgh, PA, Erlbaum, Mahwah, NJ, 1995, pp. 55–60.
[54] F. Yokota, K.M. Thompson, Value of information literature analysis: a review of applications in health risk management, Med. Decision Mak. 24 (2004) 287–298.
[55] L. Zhang, G.W. Cottrell, A computational model which learns to selectively attend in category learning, in: Paper presented at the International Conference on Development and Learning, Osaka, Japan, July 2005.
[56] M. Oaksford, N. Chater, A rational analysis of the selection task as optimal data selection, Psychological Review 101 (1994) 608–631.
[57] T.J. Palmeri, R.M. Nosofsky, Recognition memory for exceptions to the category rule, Journal of Experiment Psychology: Learning, Memory, and Cognition 21 (1995) 548–568.
[58] P.C. Wason, Reasoning, in: B.M. Foss (Ed.), New horizons in psychology Harmondsworth, England, Penguin, 1966, pp. 135–151.
[59] P.C. Wason, Reasoning about a rule, Quarterly Journal of Experimental Psychology 20 (1968) 273–281.
[60] G.L. Wells, R.C.L. Lindsay, On estimating the diagnosticity of eyewitness nonidentifications, Psychological Bulletin 88 (1980) 776–784.

Jonathan D. Nelson is interested in the statistical principles that underlie human cognition, and the adaptiveness of human cognition in natural environments. One focus is using Bayesian statistics, decision theory, and theories of optimal experimental design to understand human evidence acquisition, both in cognitive and perceptual (especially eye movement) tasks. Another interest is using fMRI to study the usefulness of information in the brain. Dr. Nelson's research has included behavioral investigation of human eye movement; investigation of the pragmatic communicative bases of risky-choice framing effects; and behavioral and modeling investigations of active information acquisition on a number concept-learning task. A recent project compared different utility functions for evidence acquisition, considering their relative potential to model human intuitions, together with their theoretical bases. Jonathan Nelson received his Ph.D. in August 2005, from the Cognitive Science Department, University of California, San Diego. He has studied with Javier Movellan, Marty Sereno, Gary Cottrell, and Terry Sejnowski. He is currently a postdoctoral scholar at the Salk Institute for Biological Studies.

Garrison W. Cottrell is a Professor of Computer Science and Engineering at UC San Diego. Professor Cottrell's main interest is Cognitive Science, in particular, building working models of cognitive processes and using them to explain psychological or neurological processes. In recent years, he has focused upon face processing, including face recognition, face identification, and facial expression recognition. He has also worked in the areas of modeling psycholinguistic processes, such as language acquisition, reading, and word sense disambiguation. He received his Ph.D. in 1985 from the University of Rochester under James F. Allen (thesis title: A connectionist approach to word sense disambiguation). He then did a postdoc with David E. Rumelhart at the Institute of Cognitive Science, UCSD, until 1987, when he joined the CSE Department.