Running head: EXEMPLAR MODELS AND BAYESIAN INFERENCE

# Exemplar models as a mechanism for performing Bayesian inference

Lei Shi Thomas L. Griffiths

Helen Wills Neuroscience Institute and Department of Psychology

University of California, Berkeley

Naomi H. Feldman

Department of Cognitive and Linguistic Sciences

Brown University

Adam N. Sanborn

Gatsby Computational Neuroscience Unit

University College London

**Word count:** 12,135

**Address for correspondence:**
Tom Griffiths
University of California, Berkeley
Department of Psychology
3210 Tolman Hall # 1650
Berkeley CA 94720-1650
**E-mail:** tom_griffiths@berkeley.edu **Phone:** (510) 642 7134 **Fax:** (510) 642 5293

**Abstract**

Probabilistic models have recently received much attention as accounts of human cognition. However, most research using probabilistic models has focused on formulating the abstract problems behind cognitive tasks and their optimal solutions, rather than considering mechanisms that could implement these solutions. Exemplar models are a successful class of psychological process models that use an inventory of stored examples to solve problems such as identification, categorization, and function learning. We show that exemplar models can be used to perform a sophisticated form of Monte Carlo approximation known as importance sampling, and thus provide a way to perform approximate Bayesian inference. Simulations of Bayesian inference in speech perception, generalization along a single dimension, making predictions about everyday events, concept learning, and reconstruction from memory show that exemplar models can often account for human performance with only a few exemplars, for both simple and relatively complex prior distributions. These results suggest that exemplar models provide a possible mechanism for implementing at least some forms of Bayesian inference.

## Exemplar models as a mechanism for performing Bayesian inference

Much of cognition and perception involves inference under uncertainty, using limited data from the world to evaluate underdetermined hypotheses. Probabilistic models provide a way to characterize rational solutions to these problems, with probability distributions encoding the beliefs of agents and Bayesian inference updating those distributions as data become available. As a consequence, probabilistic models are becoming increasingly widespread in both cognitive science and neuroscience, providing explanations of behavior in domains as diverse as motor control (Körding & Wolpert, 2004), reasoning (Oaksford & Chater, 1994), memory (Anderson & Milson, 1989), and perception (Yuille & Kersten, 2006). However, these explanations are typically presented at Marr's (1982) computational level, focusing on the abstract problem being solved and the logic of that solution. Unlike many other formal approaches to cognition, probabilistic models are usually not intended to provide an account of the mechanisms underlying behavior – how people actually produce responses consistent with optimal statistical inference.

Understanding the mechanisms that could support Bayesian inference is particularly important since probabilistic computations can be extremely challenging. Representing and updating distributions over large numbers of hypotheses is computationally expensive, a fact that is often viewed as a limitation of rational models (e.g., Kahneman & Tversky, 1972; Gigerenzer & Todd, 1999). The question of how people could perform Bayesian inference can be answered at at least two levels (as suggested by Marr, 1982). One kind of answer focuses on the neural level, exploring ways in which systems of neurons could perform probabilistic computations. The language of such answers is that of neurons, tuning curves, firing rates, and so forth, and several recent papers have explored ways in which systems of neurons could perform probabilistic computations (e.g., Ma, Beck, Latham, & Pouget, 2006; Zemel, Dayan, & Pouget, 1997). A second kind of answer is at the level of psychological processes – showing that the Bayesian inference can be performed using mechanisms that are no more complex than those used in

psychological process models. The language of such answers is representations, similarity, activation, and so forth, and some preliminary work has been done in this direction (Kruschke, 2006; Sanborn, Griffiths, & Navarro, 2006).

Our focus in this paper is on a familiar class of psychological process models known as exemplar models. These models assume that people store many instances ("exemplars") of events in memory, and evaluate new events by activating stored exemplars that are similar to those events (Medin & Schaffer, 1978; Nosofsky, 1986). It is well known that exemplar models of categorization can be analyzed in terms of nonparametric density estimation, and implement a Bayesian solution to this problem (Ashby & Alfonso-Reese, 1995). Here we show that exemplar models can be used to solve problems of Bayesian inference more generally, providing a way to approximate expectations of functions over posterior distributions. Our key result is that exemplar models can be interpreted as a sophisticated form of Monte Carlo approximation known as *importance sampling*. This result illustrates how at least some cases of Bayesian inference can be performed using a simple mechanism that is a common part of psychological process models.

Our analysis of Bayesian inference using exemplar models is also an instance of a more general strategy for exploring possible psychological mechanisms for implementing rational models. Importance sampling is one of a variety of methods used for approximating probabilistic computations in computer science and statistics. These methods are used because they provide efficient approximate solutions to problems that might be intractable to solve exactly. If we extend the principle of optimality underlying rational models of cognition to incorporate constraints on processing, we might expect to see similarities between the approximation schemes used by computer scientists and statisticians and the mechanisms by which probabilistic computations are implemented in the human mind. In some cases, as for importance sampling and examplar models, the resulting "rational process models" provide a way to connect the abstract level of analysis used in many probabilistic models of cognition with existing ideas about psychological processes.

Establishing a stronger connection between rational models of cognition and psychological

mechanisms has been a goal of cognitive scientists at least since Simon (1957) introduced the notion of "bounded rationality." Several different strategies for taking into account the effects of information-processing constraints have been considered, including incorporating those constraints into the optimization process involved in rational analysis (e.g., Anderson, 1990), handicapping rational models to produce behavior closer to that of human participants (e.g., Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), and rejecting the principle of optimization altogether in favor of finding simple but effective heuristics (e.g., Gigerenzer & Todd, 1999). The idea of developing rational process models shares characteristics with all of these strategies, with its focus being on finding psychologically plausible processes that can be justified as approximations to rational statistical inference. Such processes will ideally generalize beyond the solutions to specific optimization problems, or schemes for handicapping specific models, and provide a new way to look at the mechanistic or heuristic accounts that psychologists have developed in order to explain aspects of human behavior.

The plan of the paper is as follows. We first introduce the mathematical formulation of exemplar models and Bayesian inference. We then discuss how exact Bayesian inference can be approximated, focusing on Monte Carlo methods. A Monte Carlo method known as importance sampling is discussed in detail and its connection to exemplar models is established. The remainder of the paper explores the capacity of exemplar models to perform Bayesian inference in various tasks. These include a range of cognitive tasks from perception, generalization, prediction and concept learning. We also use simulations of performance on these tasks to investigate the effects of different kinds of capacity limitations and ongoing storage of exemplars in memory.

## Exemplar models

Human knowledge is formed by observing examples. When we learned the concept "dog," we were not taught to remember the physiological and anatomical characteristics of dogs, but instead, saw examples of various dogs. Based on the large inventory of examples of dogs we have

seen, we are able to reason about the properties of dogs, and make decisions about whether new objects we encounter are likely to be dogs. Exemplar models provide a simple explanation for how we do this, suggesting that we do not form abstract generalizations from experience, but rather store examples in memory and use those stored examples as the basis for future judgments (e.g., Medin & Schaffer, 1978; Nosofsky, 1986).

An exemplar model consists of stored exemplars $X^* = \{x_1^*, x_2^*, \cdots, x_n^*\}$, and a similarity function $s(x, x^*)$, measuring how closely a new observation $x$ is related to $x^*$.[1] On observing $x$, all exemplars are activated in proportion to $s(x, x^*)$. The use of the exemplars depends on the task (Nosofsky, 1986). In an identification task, where the goal is to identify the $x^*$ of which $x$ is an instance, the probability of selecting $x_i^*$ is

$$p_r(x_i^*|x) = \frac{s(x, x_i^*)}{\sum_{j=1}^{n} s(x, x_j^*)}, \tag{1}$$

where $p_r(\cdot)$ denotes the response distribution resulting from the exemplar model, and we assume that participants use the Luce-Shepard rule (Luce, 1959; Shepard, 1962) in selecting a response, with no biases towards particular exemplars. In a categorization task, where each exemplar $x_j^*$ is associated with a category $c_j$, the probability that the new object $x$ is assigned to category $c$ is given by

$$p_r(c|x) = \frac{\sum_{j|c_j=c} s(x, x_j^*)}{\sum_{j=1}^{n} s(x, x_j^*)}, \tag{2}$$

where again we assume a Luce-Shepard rule without biases towards particular categories.

While exemplar models have been most prominent in the literature on categorization, the same basic principles have been used to define models of function learning (DeLosh, Busemeyer, & McDaniel, 1997), probabilistic reasoning (Juslin & Persson, 2002), and social judgment (Smith & Zarate, 1992). These models pursue a similar approach to models of categorization, but associate each exemplar with a quantity other than a category label. For example, in function learning each exemplar is associated with the value of a continuous variable rather than a discrete

category index. The procedure for generating responses remains the same as that used in Equations 1 and 2: the associated information is averaged over exemplars, weighted by their similarity to the stimulus. Thus, the predicted value of some associated information $f$ for a new stimulus $x$ is

$$\hat{f} = \frac{\sum_{j=1}^{n} f_j s(x, x_j^*)}{\sum_{j=1}^{n} s(x, x_j^*)}, \tag{3}$$

where $f_j$ denotes the information associated with the $j$th exemplar. The identification and categorization models can be viewed as special cases, corresponding to different ways of specifying $f_j$. Taking $f_j = 1$ for $j = i$ and 0 otherwise yields Equation 1, while taking $f_j = 1$ if $c_j = c$ and 0 otherwise yields Equation 2. Equation 3 thus provides the general formulation of an exemplar model that we will analyze.

## Bayesian inference

Many cognitive problems can be formulated as evaluating a set of hypotheses about processes that could have produced observed data. For example, perceiving speech sounds requires considering what sounds might be consistent with an auditory stimulus (Feldman, Griffiths, & Morgan, 2009), generalizing a property from one object to another involves considering the set of objects likely to possess that property (Shepard, 1987), predicting the duration of an ongoing event necessitates reasoning from its current duration to a hypothetical future endpoint (Griffiths & Tenenbaum, 2007), and learning a concept from examples means evaluating a space of possible concepts (Tenenbaum & Griffiths, 2001). Even reconstructing information from memory can be analyzed as an inference about the nature of that information from the data provided by a noisy memory trace (Huttenlocher, Hedges, & Vevea, 2000).

Bayesian inference provides a solution to problems of this kind. Letting $h$ denote a hypothesis and $d$ the data, assume a learner encodes his or her degrees of belief regarding the hypotheses before seeing $d$ using a probability distribution, $p(h)$, known as the *prior* distribution.

Then, the degrees of belief after seeing $d$ are given by the *posterior* distribution, $p(h|d)$, obtained from Bayes' rule

$$p(h|d) = \frac{p(d|h)p(h)}{\int_{\mathcal{H}} p(d|h)p(h)\,dh}, \tag{4}$$

where $\mathcal{H}$ is the set of hypotheses under consideration (the *hypothesis space*), and $p(d|h)$ is a distribution indicating the probability of seeing $d$ if $h$ were true, known as the *likelihood*.

While our analysis applies to Bayesian inference in the general case, we will introduce it using a specific example that is consistent with several of the psychological tasks we consider later in the paper. We will return to the general case after working through this specific example. Assume we observe a stimulus $x$, which we believe to be corrupted by noise and potentially missing associated information, such as a category label. Let $x^*$ denote the uncorrupted stimulus, and $z$ denote the missing data. Often, our goal is simply to reconstruct $x$, finding the $x^*$ to which it corresponds. In this case, $z$ can be empty. Otherwise, we seek to infer both $x^*$ and the value of $z$ which corresponds to $x$. We can perform both tasks using Bayesian inference.

The application of Bayes' rule is easier to illustrate in the case where $z$ is empty, where we simply wish to infer the true stimulus $x^*$ from noisy $x$. We use the probability distribution $p(x|x^*)$ to characterize the noise process, indicating the probability with which the stimulus $x^*$ is corrupted to $x$, and the probability distribution $p(x^*)$ to encode our a priori beliefs about the probability of seeing a given stimulus. We can then use Bayes' rule to compute the posterior distribution over the value of the uncorrupted stimulus, $x^*$, which might have generated the observation $x$, obtaining

$$p(x^*|x) = \frac{p(x|x^*)p(x^*)}{\int p(x|x^*)p(x^*)\,dx^*}, \tag{5}$$

where $p(x|x^*)$ is the likelihood and $p(x^*)$ is the prior.

This analysis is straightforward to generalize to the case where $z$ contains missing data, such as the label of the category from which $x$ was generated. In this case, we need to define our prior as a distribution over both $x^*$ and $z$, $p(x^*,z)$. We can then use Bayes' rule to compute the posterior

distribution over the uncorrupted stimulus, $x^*$, and missing data, $z$, which might have generated the observation $x$, obtaining

$$p(x^*, z|x) = \frac{p(x|x^*)p(x^*, z)}{\int\int p(x|x^*)p(x^*, z)\, dx^*\, dz}, \tag{6}$$

where we also assume that the probability of observing $x$ is independent of $z$ given $x^*$, so $p(x|x^*, z) = p(x|x^*)$.

## Evaluating expectations by Monte Carlo

Posterior distributions on hypotheses given data can be used to answer a variety of questions. To return to the example above, a posterior distribution on $x^*$ and $z$ can be used to evaluate the properties of $x^*$ and $z$ given $x$. A standard way to do this is to use the expectation of a function over the posterior distribution. For any function $f(x^*, z)$, the posterior *expectation* of that function given $x$ is

$$E\left[f(x^*, z)|x\right] = \int\int f(x^*, z)p(x^*, z|x)\, dx^*\, dz, \tag{7}$$

being the average of $f(x^*, z)$ over the posterior distribution. Since $f(x^*, z)$ can pick out any property of $x^*$ and $z$ that might be of interest, many problems of reasoning under uncertainty can be expressed in terms of expectations. For example, we could compute the posterior mean of $x^*$ by taking $f(x^*, z) = x^*$, or calculate the posterior probability that $z$ takes a particular value by taking $f(x^*, z)$ to be 1 when $z$ has that value, and 0 otherwise.

Evaluating expectations over the posterior distribution can be challenging: it requires computing a posterior distribution, which is a hard problem in itself, because the integrals in Equation 7 can range over many values for $x^*$ and $z$. Consequently, Monte Carlo methods are often used to approximate expectations. Monte Carlo methods approximate the expectation of a function with respect to a probability distribution with the average of that function at points drawn from the distribution. Assume we want to evaluate the expectation of a function $g(y)$ over the distribution $p(y)$, $E_p\left[g(y)\right]$ (where we use $y$ as a generic random variable, instead of $x^*$ and $z$). Let $\mu$ denote the

value of this expectation. The law of large numbers justifies

$$\mu = E_p[g(y)] = \int g(y)p(y)\,dy \approx \frac{1}{m}\sum_{j=1}^{m} g(y_j), \tag{8}$$

where the $y_j$ are all drawn from the distribution $p(y)$.

This simple Monte Carlo method requires that we are able to generate samples from the distribution $p(y)$. However, this is often not the case: it is quite common to encounter problems where $p(y)$ is known at all points $y$ but hard to sample from. If a *surrogate distribution* $q(y)$ is close to $p(y)$ but easy to sample from, a form of Monte Carlo called *importance sampling* can be applied (see Neal, 1993 for a detailed introduction, and Robert & Casella, 1999 for a mathematical treatment). Manipulating the expression for the expectation of $g$ gives

$$\int g(y)p(y)\,dy = \frac{\int g(y)p(y)\;dy}{\int p(y)\,dy} = \frac{\int g(y)\frac{p(y)}{q(y)}q(y)\,dy}{\int \frac{p(y)}{q(y)}q(y)\,dy}. \tag{9}$$

The numerator and denominator of this expression are each expectations with respect to $q(y)$. Applying simple Monte Carlo (with the same set of samples from $q(y)$) to both,

$$\mu = E_p[g(y)] \approx \frac{\sum_{j=1}^{m} g(y_j)\frac{p(y_j)}{q(y_j)}}{\sum_{j=1}^{m}\frac{p(y_j)}{q(y_j)}}, \tag{10}$$

where each $y_j$ is drawn from $q(y)$. The ratios $\frac{p(y_j)}{q(y_j)}$ are "importance weights" on the samples $y_j$, correcting for having sampled from $q(y)$ rather than $p(y)$. Intuitively, these weights capture how important each sampled value should be to calculating the expectation, and give importance sampling its name. If the $y_j$ are sampled directly from $p(y)$, they are given equal weight, each having an importance weight of 1. However, when the $y_j$ are sampled from surrogate distribution $q(y)$, they bear nonuniform importance weights due to the difference between $p(y)$ and $q(y)$. Samples with higher probability under $p(y)$ than $q(y)$ occur less often than they would if we were sampling from $p(y)$, but receive greater weight, counter-balancing the lower sampling frequency, with the opposite applying to samples with higher probability under $q(y)$ than $p(y)$.

Importance sampling is a useful method for approximating expectations when simple Monte Carlo cannot be applied because generating samples from the target distribution is difficult. However, using an importance sampler can make sense even in cases where simple Monte Carlo can also be applied. First, it allows a single set of samples to be used to evaluate expectations with respect to a range of distributions, through the use of different weights for each distribution. Second, the estimate of $\mu$ produced by the importance sampler can have lower variance than the estimate produced by simple Monte Carlo, if the surrogate distribution is chosen to place high probability on values of $y$ where both $p(y)$ and the contribution of $g(y)$ to the expectation are large. 2

Both simple Monte Carlo and importance sampling can be applied to the problem of evaluating the expectation of a function $f(x^*,z)$ over a posterior distribution on $x^*$ and $z$ with which we began this section. Simple Monte Carlo would draw values of $x^*$ and $z$ from the posterior distribution $p(x^*,z|x)$ directly. Importance sampling would generate from surrogate distribution, $q(x^*,z)$, and then reweight those samples. One simple choice of $q(x^*,z)$ is the prior, $p(x^*,z)$. If we sample from the prior, the weight assigned to each sample is the ratio of the posterior to the prior

$$\frac{p(x^*,z|x)}{p(x^*,z)} = \frac{p(x|x^*)}{\int\int p(x|x^*)p(x^*,z)\,dx^*\,dz}, \tag{11}$$

where we use the assumption that $p(x|x^*,z) = p(x|x^*)$. Substituting these weights into Equation 10 and cancelling constants, we obtain

$$E\left[f(x^*,z)|x\right] \approx \frac{\sum_{j=1}^{m} f(x_j^*,z_j)p(x|x_j^*)}{\sum_{j=1}^{m} p(x|x_j^*)}, \tag{12}$$

where we assume that $x_j^*$ and $z_j$ are drawn from $p(x^*,z)$. Because the weights on the samples are based on the likelihood, this approach is sometimes known as *likelihood weighting*.

Figure 1 provides a visual illustration of the approximation of Bayesian inference using importance sampling. Here, the goal is to recover the true value of a noisy observation $x$, which is

done by computing the posterior expectation $E[x^*|x]$. This can be done applying Equation 12 with $f(x^*, z) = x^*$. First, exemplars $x^*_j$ are drawn from prior distribution $p(x^*)$, producing the collection of sampled values shown in Figure 1 (a). Then, these exemplars are given weights proportional to the likelihood $p(x|x^*)$. In this case, the likelihood is a Gaussian distribution with its mean at $x^*$, and the same standard deviation for each value of $x^*$. Since the Gaussian is symmetric in its arguments (in this case, $x$ and $x^*$), the function used to assign weights to each $x^*_j$ is also Gaussian, with its mean at $x$, as illustrated in Figure 1 (b). Finally, $E[x^*|x]$ is estimated by the weighted sum $\sum_j x^*_j p(x|x^*_j)$ normalized by $\sum_j p(x|x^*_j)$. The posterior expectation moves the estimate of $x^*$ closer to the nearest mode of the prior distribution, as shown in Figure 1 (c), appropriately combining prior knowledge with the noisy observation. This computation is straightforward despite the complicated shape of the prior distribution.

The success of this importance sampling scheme for approximating posterior expectations depends on how much probability mass the prior and posterior distribution share. This can be understood by considering how the variance of the importance weights depends on the relationship between the surrogate and target distributions. The variance of the importance weights determines the stability of the estimate produced by importance sampling: If only a few samples have high weights, then the estimate of the expectation will be based only on those samples. Figure 2 provides some intuitions for this phenomenon. If the prior largely overlaps with the posterior, as in Figure 2 (a), the importance weights have little variance and the estimate produced by the sampler is fairly stable. If the prior does not overlap with the posterior, as in Figure 2 (b), few samples from the prior fall in the region with higher posterior probability, and these samples are given all the weight. The estimate is then solely dependent on these samples and is highly unstable. In intermediate cases, such as that shown in Figure 2 (c) where the prior is a multi-modal distribution and the posterior is one of the modes, stable results are obtained if enough samples are drawn from each of the modes. In cases where there is not a close match between prior and posterior, a reasonably large number of samples needs to be drawn from the prior to ensure a good

approximation.

## Exemplar models as importance samplers

Inspection of Equations 3 and 12 yields our main result: Exemplar models can be viewed as implementing a form of importance sampling. More formally, assume $X^*$ is a set of $m$ exemplars $x^*$ and associated information $z$ drawn from the probability distribution $p(x^*, z)$, and $f_j = f(x_j^*, z_j)$ for some function $f(x^*, z)$. Then the output of Equation 3 for an exemplar model with exemplars $X^*$ and similarity function $s(x, x^*)$ is an importance sampling approximation to the expectation of $f(x^*, z)$ over the posterior distribution on $x^*$ and $z$, as given in Equation 6, if two conditions are fulfilled: the $x_j^*$ and $z_j$ making up $X^*$ are sampled from the prior $p(x^*, z)$ and the similarity function $s(x, x^*)$ is proportional to the likelihood $p(x|x^*)$. Returning to Figure 1, the $x_i^*$ are now exemplars, and the importance weights reflect the amount of activation of those exemplars based on similarity to the observed data $x$.

The two conditions identified in the previous paragraph are crucial in establishing the connection between exemplar models and importance sampling. They are also reasonably natural assumptions, if we assume that exemplars are stored in memory as the result of experience, and that similarity functions are flexible and can vary from task to task. For most perceptual tasks of the kind we have been considering here, the prior $p(x^*, z)$ represents the distribution over the states of the environment that an agent lives in. Thus, sampling $x_j^*$ and $z_j$ from the prior is equivalent to storing randomly generated events in memory. The second condition states that the similarity between $x$ and $x^*$ corresponds to the likelihood function, subject to a ratio constant. This is straightforward when the stimulus $x$ exists in the same space as $x^*$, as when $x$ is a noisy observation of $x^*$. In this case, similarity functions are typically assumed to be monotonically decreasing functions in space, such as exponentials or Gaussians, which map naturally to likelihood functions (Nosofsky, 1986, 1990; Ashby & Alfonso-Reese, 1995).

This connection between exemplar models and importance sampling provides an alternative

rational justification for exemplar models of categorization, as well as a more general motivation for these models. The justification for exemplar models in terms of nonparametric density estimation (Ashby & Alfonso-Reese, 1995) provides a clear account of their relevance to categorization, but does not explain why they are appropriate in other contexts, such as identification (Equation 1) or the general response rule given in Equation 3. In contrast, we can use importance sampling to provide a single explanation for many uses of exemplar models, such as categorization, identification and function learning, viewing each as the result of approximating an expectation of a particular function $f(x^*, z)$ over the posterior distribution $p(x^*, z|x)$. For categorization, $z$ is the category label and the quantity of interest is $p(z = c|x)$, the posterior probability that $x$ belongs to category $c$. Hence, $f(x^*, z) = 1$ for all $z = c$ and 0 otherwise. For identification, the question is whether the observed $x$ corresponds to a specific $x^*$, so $f(x^*, z) = 1$ for that $x^*$ and 0 otherwise, regardless of $z$. For function learning, $z$ contains the value of the continuous variable associated with $x^*$, and $f(x^*, z) = z$. Similar analyses apply in other cases, with exemplar models providing a rational method for answering questions expressed as an expectation of a function of $x^*$ and $z$.

## A general scheme for approximating Bayesian inference

The equivalence between exemplar models and importance sampling established in the previous section focuses on the specific problem of interpreting a noisy stimulus. However, the idea that importance sampling constitutes a psychologically plausible mechanism for approximating Bayesian inference generalizes beyond this specific problem. In the general case an agent seeks to evaluate a hypothesis $h$ in light of data $d$, and does so by computing the posterior distribution $p(h|d)$ as specified by Equation 4. An expectation of a function $f(h)$ over the posterior distribution can be approximated by sampling hypotheses from the prior, $p(h)$, and weighting the

samples by the likelihood, $p(d|h)$. Formally, we have

$$E[f(h)|d] = \int f(h)p(h|d)\,dh \approx \frac{\sum_{j=1}^{m} f(h_j)p(d|h_j)}{\sum_{j=1}^{m} p(d|h_j)}, \tag{13}$$

where $h_j$ is drawn from the prior $p(h)$.

Approximating Bayesian inference by importance sampling in this general case can also be interpreted as a kind of exemplar model, but here the stored "exemplars" correspond to hypotheses rather than stimuli. As in a standard exemplar model, these hypotheses can be stored in memory as the consequence of previous learning events. Each hypothesis needs to be weighted by its likelihood, which no longer has a natural interpretation in terms of similarity, but represents the degree to which a hypothesis is "activated" as a result of observing the data. Thus, all that is required for an agent to be able to approximate Bayesian inference in this way is to store hypotheses in memory as they are encountered, and to activate those hypotheses in such a way that the hypotheses that best account for the data receive the most activation.

The theoretical properties of importance sampling suggest that exemplar models of the kind considered in this and the preceding section may provide a way to approximate Bayesian inference in at least some cases. Specifically, we expect that importance sampling with a relatively small number of samples drawn from the prior should produce an accurate approximation to Bayesian inference in cases where prior and posterior share a reasonable amount of probability mass. This can occur in cases where the data are relatively uninformative, either as a result of small samples or high levels of noise. Despite this constraint, we anticipate that there will be a variety of applications in which exemplar models provide a good enough approximation to Bayesian inference to account for existing behavioral data.

In the remainder of the paper we present a series of simulations evaluating exemplar models as a scheme for approximating Bayesian inference in five tasks. These tasks are selected to illustrate the breadth of this approach, and to allow us to explore the effect of number of exemplars

on performance, as well as the consequences of other variants on the basic importance sampling scheme intended to reflect possible psychological or biological constraints. In general, we use the notation from the original papers in describing these simulations. However, in each case we formulate the underlying problem to be solved by Bayesian inference, and relate it back to either the specific or general problems of Bayesian inference we have considered in establishing the connection to exemplar models, identifying the correspondence between the relevant variables.

### Simulation 1: The perceptual magnet effect

Categorical perception of speech sounds was first demonstrated by Liberman, Harris, Hoffman, and Griffith (1957), who showed that listeners' discrimination of stop consonants was little better than would be predicted on the basis of categorization performance, with sharp discrimination peaks at category boundaries. Evidence has also been found in vowels for a *perceptual magnet effect*, a language-specific shrinking of perceptual space specifically near category prototypes, presumably due to a perceptual bias toward category centers (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). However, perception of vowels differs from that of stop consonants in that it is continuous rather than strictly categorical, with listeners showing high levels of within-category discrimination (Fry, Abramson, Eimas, & Liberman, 1962). Because of the high level of within-category discriminability in vowels, the perceptual magnet effect has been difficult to capture through traditional labeling accounts of categorical perception.

Feldman et al. (2009) argued that the perceptual magnet effect arises because listeners are trying to recover the phonetic detail (e.g., formant values) of a speaker's target production from a noisy speech signal. Under this account, listeners perform a Bayesian denoising process, recovering the intended formant values of the noisy speech sounds they hear. Speech sounds are assumed to belong to phonetic categories in the native language, and listeners can use their knowledge of these categories to guide their inferences of the speaker's target production. Because this account assumes that listeners are trying to recover phonetic detail, it predicts a baseline level

of within-category discrimination while still allowing categories to influence listeners' perception.

The Bayesian model introduced by Feldman et al. (2009) assumes that speakers, in producing a speech sound, sample a phonetic value for their target production $T$ from a Gaussian phonetic category $c$ with category mean $\mu_c$ and category variance $\sigma_c^2$. Listeners hear a speech sound $S$, which has been perturbed by articulatory, acoustic, and perceptual noise. This noisy speech sound $S$ is normally distributed around the target production $T$ with noise variance $\sigma_S^2$. The prior on target productions is therefore a mixture of Gaussians representing the phonetic categories of the language,

$$p(T) = \sum_c p(T|c)p(c) = \sum_c N(T|\mu_c, \sigma_c^2)p(c), \tag{14}$$

where $N(T|\mu_c, \sigma_c^2)$ is the probability density at $T$ given a Gaussian distribution with mean $\mu_c$ and variance $\sigma_c^2$. The likelihood function represents the noise process that corrupts a target production $T$ into a speech sound $S$, and is given by the Gaussian function representing speech signal noise,

$$p(S|T) = N(S|T, \sigma_S^2). \tag{15}$$

Listeners hear the speech sound $S$ and use Bayes' rule to compute the posterior mean (ie. the expectation $E[T|S]$) and optimally recover the phonetic detail of a speaker's target production, marginalizing over all possible category labels.

The problem of inferring $T$ from $S$ is directly analogous to the problem of inferring a true stimulus $x^*$ from a noisy stimulus $x$ that we considered when introducing importance sampling. To complete the analogy, the category $c$ corresponds to the missing information $z$, and the expectation $E[T|S]$ corresponds to $E[x^*|x]$. This expectation can thus be approximated by an importance sampler of the form given in Equation 12, with $f(x^*, z) = x^*$. By the equivalence between importance sampling and exemplar models, this means that we can approximate the Bayesian solution to the problem of inferring $T$ from $S$ using an exemplar model.

An exemplar model derived through importance sampling provides a psychologically

plausible implementation of the model introduced by Feldman et al. (2009), allowing listeners to optimally recover speakers' target productions using unlabeled exemplars. This implementation has two specific advantages over the original Bayesian formulation. First, there is evidence that infants as young as six months show a language-specific perceptual magnet effect even though they are still forming phonetic categories (Kuhl et al., 1992), and importance sampling allows them to perform this computation without any explicit category knowledge. Category labels are not required, and the distribution of exemplars need not follow any parametric distribution. Second, importance sampling directly parallels the neural network model of the perceptual magnet effect proposed by Guenther and Gjaja (1996), allowing the Bayesian model and the neural network model to be interpreted as convergent descriptions of the same perceptual process.

To calculate the expected target production $T$ using importance sampling, listeners need to store their percepts of previously encountered speech sounds, giving them a sample from $p(T)$, the prior on target productions (Equation 14).[3] Upon hearing a new speech sound, they weight each stored exemplar by its likelihood $p(S|T)$ (Equation 15) and take the weighted average of these exemplars to approximate the posterior mean as

$$E[T|S] \approx \frac{\sum_{j=1}^{m} T_j p(S|T_j)}{\sum_{j=1}^{m} p(S|T_j)}, \tag{16}$$

where $T_j$ denotes the formant value of a stored target production.

We compared the performance of this exemplar model to multidimensional scaling data from Iverson and Kuhl (1995) on adult English speakers' discrimination of 13 equally-spaced stimuli in the /i/ and /e/ categories. The discrimination data were obtained through an AX task in which subjects heard pairs of stimuli and pressed a button to indicate whether the stimuli were identical. Responses and reaction times were used in a multidimensional scaling analysis to create a one-dimensional map of perceptual space, shown in Figure 3. The data show a non-linear mapping between acoustic space and perceptual space, with portions that are more nearly

horizontal corresponding to areas in which perceptual space is shrunk relative to acoustic space. Sounds near phonetic category centers are closer together in perceptual space than sounds near category boundaries, despite being separated by equal psychophysical distances. We simulated the performance of exemplar models with ten and fifty exemplars drawn from the prior, examining both the performance of individual simulated participants and the results of aggregating across participants. The results of this simulation, shown together with the multidimensional scaling data in Figure 3, suggest that a relatively small number of exemplars suffices to capture human performance in this perceptual task. Model performance using ten exemplars already demonstrates the desired effect, and with fifty exemplars, the model gives a precise approximation that closely mirrors the combined performance of the 18 subjects in Iverson and Kuhl's multidimensional scaling experiment.

In addition to giving a simple psychological mechanism for approximating Bayesian inference in this task, importance sampling provides a link between the Bayesian model and a previous account of the perceptual magnet effect. The exemplar model considered in this section is isomorphic to a neural mechanism proposed by Guenther and Gjaja (1996) to create a bias toward category centers. In Guenther and Gjaja's neural map, the firing preferences of a population of neurons come to mirror the distribution of speech sounds in the input. Upon hearing a speech sound, listeners recover a percept of that speech sound by taking a weighted average of firing preferences in the neural map. The weights, or neural activations, are determined by the similarity between a neuron's firing preference and the speech sound heard. This perceptual mechanism implements an importance sampler: Firing preferences of individual neurons constitute samples from the prior, and the activation function plays the role of the likelihood. The activation function in the neural map differs from the Gaussian function assumed in the Bayesian model, but both implement the idea that exemplars with similar acoustic values should be weighted most highly. The correspondence between these two models suggests that Monte Carlo methods such as importance sampling may provide connections not just to psychological processes, but to the

neural mechanisms that might support probabilistic computations. We return to this possibility in the General Discussion.

### Simulation 2: The universal law of generalization

In a celebrated paper, Shepard (1987) showed that generalization gradients decrease exponentially with psychological distance across many experimental situations. He then gave a probabilistic explanation for this phenomenon that was later formulated in a Bayesian framework (Myung & Shepard, 1996; Tenenbaum & Griffiths, 2001). Here, we use the notation originally introduced by Shepard. Assume that we observe a stimulus **0** that has a certain property (or "consequence"). What is the probability that a test stimulus **x** has the same property? Shepard analyzed this problem by assuming that **0** and **x** were points in a psychological space, and the set of stimuli sharing a property defined a consequential region in the space. We know that the original stimulus **0** belongs to this region, and we want to evaluate whether the test stimulus **x** does. We thus want to compute the probability that the **x** falls into an unknown consequential region containing **0**.

The first question we can answer is which consequential regions **0** could have come from. This is a problem of Bayesian inference, where consequential regions are hypotheses and observing that **0** belongs to the region constitutes data. In the case of one-dimensional generalization, we might take consequential regions to be intervals along that dimension, parameterized by their center $c$ and size $s$. We then want to compute the posterior distribution on intervals $(c,s)$ given the information that $\mathbf{0} \in (c,s)$. This can be done by defining a prior $p(c,s)$ and likelihood $p(\mathbf{0}|c,s)$. Shepard (1987) assumed that all locations of consequential regions are equally probable, so the distribution of $c$ is uniform and the prior distribution $p(c,s)$ is specified purely in terms of a distributon on sizes, $p(s)$. The likelihood is obtained by assuming that **0** is sampled uniformly at random from the interval given by $(c,s)$, resulting in $p(\mathbf{0}|c,s) = 1/m(s)$ for all intervals $(c,s)$ containing **0**, where $m(s)$ is a measure of the volume of a region of size $s$ (in one dimension, the length of the interval), and $p(\mathbf{0}|c,s) = 0$ for all other intervals. Prior and likelihood

can then be combined as in Equation 4 to yield a posterior distribution over consequential regions.

With a posterior distribution over consequential regions in hand, the probability that $\mathbf{x}$ belongs to one of the consequential regions containing $\mathbf{0}$ is obtained by summing the posterior probabilities of the regions containing $\mathbf{x}$. This can be expressed as the integral

$$p(\mathbf{x}|\mathbf{0}) = \int_{s,c} \mathbf{1}(\mathbf{x} \in (c,s)) p(c,s|\mathbf{0})\, ds\, dc, \tag{17}$$

where $\mathbf{1}(\mathbf{x} \in (c,s))$ is an indicator function that equals 1 if $\mathbf{x}$ is in the region parameterized by $(c,s)$ and 0 otherwise. This integral can also be viewed as an expectation of the indicator function $\mathbf{1}(\mathbf{x} \in (c,s))$ over the posterior distribution $p(c,s|\mathbf{0})$.

By viewing Equation 17 as an expectation, it becomes clear that it can be approximated by importance sampling, and thus by an exemplar model. Identifying a consequential region does not match the form of the simple stimulus de-noising problem that we used in demonstrating equivalence between importance sampling and exemplar models, requiring us to use the more general idea that Bayesian inference can be approximated by storing hypotheses sampled from the prior and activating them based on consistency with data. In this case, the hypotheses $h$ are consequential regions, the data $d$ consist of the observation that $\mathbf{0}$ is contained in some consequential region, and the function $f(h)$ that we want the expectation of is the indicator function that takes the value 1 if $\mathbf{x}$ is in the consequential region and 0 otherwise. The approximation to this expectation is then given by Equation 13.

The importance sampling approximation to Equation 17 is thus obtained by assuming that a set of hypotheses parameterized by centers and sizes $(c_j, s_j)$ are sampled from the prior and activated by the likelihood $\mathbf{1}(\mathbf{0} \in (c_j, s_j))\, 1/m(s_j)$, to give

$$p(\mathbf{x}|\mathbf{0}) \approx \frac{\sum_{j=1}^{m} \mathbf{1}(\mathbf{x}, \mathbf{0} \in (c_j, s_j)) \frac{1}{m(s_j)}}{\sum_{j=1}^{m} \mathbf{1}(\mathbf{0} \in (c_j, s_j)) \frac{1}{m(s_j)}}. \tag{18}$$

The numerator simplifies the product of the indicator function that we want the expectation of,

$\mathbf{1}(\mathbf{x} \in (c_j, s_j))$, with that in the likelihood, $\mathbf{1}(\mathbf{0} \in (c_j, s_j))$, to a single indicator function that takes the value 1 when both $\mathbf{x}$ and $\mathbf{0}$ are in the interval $(c_j, s_j)$. Since $c$ and $s$ are independent under the prior, we can also draw $m$ samples of each and then take the sum over all $m^2$ pairs of $c$ and $s$ values, reducing the number of samples that need to be taken from the prior.

The results of using this approximation with several different priors on the size of the consequential region are shown in Figure 4. The different priors that are used are those considered by Shepard (1987) in his original analysis of generalization behavior. The figure shows the generalization gradient – the probability of generalizing from $\mathbf{0}$ to $\mathbf{x}$ as a function of psychological distance – for these six prior distributions, together with approximations that vary the number of sampled hypotheses. In the one-dimensional case, the psychological distance between $\mathbf{0}$ and $\mathbf{x}$ is just the value of $\mathbf{x}$ (taking $\mathbf{0}$ as the origin), which is shown on the horizontal axis of each plot in the figure . Relatively small numbers of sampled hypotheses (20 and 100) are sufficient to produce reasonable approximations to the generalization gradients associated with all of these prior distributions.

### Simulation 3: Predicting the future

Remembering past events, like the local temperature in March in previous years, or the duration of red traffic lights, can help us make good predictions in everyday life. Griffiths and Tenenbaum (2006) studied people's predictions about a variety of everyday events, including the grosses of movies and the time to bake a cake, and found that these predictions corresponded strikingly well with the actual distributions of these quantities. In each case, people were asked to predict the total extent or duration of a quantity based on its current value, such as how much money a movie would make based on how much it has made so far, or how long a cake would be in the oven based on how long it has currently been in the oven. Predicting the future in this way can be analyzed as Bayesian inference, and approximated using an exemplar model.

As formulated in Griffiths and Tenenbaum (2006), the statistical problem that people solved

is inferring the total duration or extent of a quantity, $t_{total}$, from its current duration or extent, $t$. The goal is to compute the posterior median of $t_{total}$ given $t$. Unlike the mean, the median gives a robust estimate of $t_{total}$ when the posterior distribution is skewed, which is the case for many of these everyday quantities. The posterior median $t^*$ is defined to be the value such that $p(t_{total} > t^*|t) = 0.5$, where the posterior distribution is obtained by applying Bayes' rule with an appropriate prior and likelihood. The prior $p(t_{total})$ depends on the distribution of the everyday quantity in question, with temperatures and traffic lights being associated with different distributions. As in the previous example, the likelihood is obtained by assuming that the phenomenon is encountered at a random point drawn uniformly from the interval between 0 and $t_{total}$, with $p(t|t_{total}) = 1/t_{total}$ for all $t_{total} > t$.

Making correct predictions about everyday events requires knowing the prior distributions of the relevant quantities – the grosses of movies, the time taken to bake a cake, and so forth. While it is unlikely that we store these distributions explicitly in memory, the posterior median can be approximated using stored exemplars that are sampled from the prior $p(t_{total})$ using Equation 12. The posterior probability that a value of $t_{total}$ is greater than $t^*$ can be formulated as an expectation,

$$p(t_{total} > t^*|t) = E[\mathbf{1}(t_{total} > t^*)|t], \tag{19}$$

where $\mathbf{1}(t_{total} > t^*)$ is an indicator function taking the value 1 when its argument is true, and 0 otherwise, as in the previous example. This problem fits the schema for the general approximation to Bayesian inference given by Equation 13, with the hypotheses $h$ being values of $t_{total}$, the data $d$ being the observation $t$, and the function of interest $f(h)$ being the indicator function $\mathbf{1}(t_{total} > t^*)$. Consequently, the expectation given in Equation 19 can be approximated using an exemplar model in which exemplars $t_{total,j}$ are sampled from the prior $p(t_{total})$ and activated by the likelihood

$1/t_{total}$ if they are greater than $t$. This gives the approximation

$$p(t_{total} > t^* | t) \approx \frac{\sum_j \mathbf{1}(t_{total,j} > t^*, t_{total,j} > t)\frac{1}{t_{total,j}}}{\sum_j \mathbf{1}(t_{total,j} > t)\frac{1}{t_{total,j}}}. \quad (20)$$

The approximate median of the posterior distribution is the exemplar $t_{total,j}$ that has $p(t_{total} > t_{total,j}|t)$ closest to 0.5.

Considering limitations in memory capacity and computational power, we conducted two sets of simulations. In predicting the future, only values of $t_{total}$ that are greater than the observed value of $t$ are plausible, with all other values having a likelihood of 0. Consequently, sampling directly from the prior can be inefficient, with many samples being discarded. We can thus break the approximation process into two steps, with the first being generating a set of values of $t_{total}$ from memory, and the second being assigning those values of $t_{total}$ greater than $t$ a likelihood of $1/t_{total}$ and normalizing. Our simulations considered limitations that could apply to either of these steps. In the *memory-limited* case, the number of exemplars generated from memory is fixed. In the *computation-limited* case, the bottleneck is the number of exemplars that can be processed simultaneously, placing a constraint on the number of exemplars such that $t_{total} > t$. In this case, we assume that exemplars are generated from memory until they reach this upper limit.

Figure 5 shows the results of applying these different approximation schemes to the predicting the future task, varying the number of exemplars. We examined performance across seven prior distributions, corresponding to the baking time of cakes, human life spans, the grosses of movies, the duration of the reigns of pharaohs, the length of poems, the number of terms in the United States House of Representatives, and the runtime of movies, and for 5, 10, and 15 exemplars. The prior distributions were those used by Griffiths and Tenenbaum (2006), who collected data from online databases for each of these different quantities. In each case, we simulated the performance of 50 participants using the appropriate number of exemplars sampled directly from the prior (for the memory-limited case) or sampled from the prior but constrained to

be consistent with the observed value of $t$ (for the computation-limited case). In the memory-limited case, if none of the exemplars is larger than the observation, the observed value $t$ is taken as the only exemplar which results in $t^* = t$. The figure also shows the quality of the approximation produced by directly sampling exemplars from the posterior distribution, rather than generating from the prior. For each approximation scheme, 50 simulated participants' responses were generated. The plot markers indicate the median and the 68% confidence interval on the median (ie. the 16th and 84th percentiles of the sampling distribution), computed using a bootstrap with 1000 samples drawn from the responses of these participants.

For a quantitative measure of the success of the approximation, we computed the sum of the absolute value of the deviations for each of the median results shown in Figure 5 ($t^*_{ML}, t^*_{CL}, t^*_{SA}$ for memory-limited, computation-limited, and sampling respectively) to both the true function ($t^*_{Bayes}$) and to the median human responses ($t^*_{human}$). These error scores were then normalized by the difference in $t^*_{Bayes}$ for the lowest and highest values of $t$ for each prior, in order to compensate for the different scales of these quantities, and then summed across priors to produce the scores shown in Table 1. This quantitative analysis confirmed the trends evident from the figure. Approximation performance improved with more exemplars, but was already fairly good with only five exemplars when compared against the performance of the full Bayesian model considered by Griffiths and Tenenbaum (2006). The memory-limited case tended to perform worse than the other approximations for a given number of exemplars, since some of the exemplars generated from the prior would not enter into the approximation for the reasons detailed above.

The question of whether approximations based on a small number of exemplars might account for the results of Griffiths and Tenenbaum (2006) was independently raised by Mozer, Pashler, and Homaei (2008), who argued that a close correspondence to the posterior median could be produced by aggregating responses across a large number of participants who each had only limited knowledge of the appropriate prior, such as a handful of samples from that distribution. The original model considered by Mozer et al. (2008), which estimates $t^*$ as the minimum of the

set of exemplars greater than $t$, does not have an interpretation as importance sampling, and degenerates as an approximation as the number of exemplars increases, rather than improving. However, one of the variants on this model, called GTkGuess in their paper, is equivalent to our memory-limited importance sampling approximation provided at least one sampled exemplar is greater than $t$. Consistent with the results presented here, Mozer et al. (2008) demonstrated that this model produced a good correspondence with the results of Griffiths and Tenenbaum (2006) with only a small number of exemplars, considering both aggregate performance and the amount of variability produced by different approximation schemes.

One important difference between the analysis we present here and that of Mozer et al. (2008) is that we do not necessarily view using an exemplar model to approximate Bayesian inference as being related to having limited prior knowledge. For Mozer et al. (2008), the exemplars used in approximating Bayesian inference were taken to represent all that a given individual knew about a phenomenon. Since each participant in Griffiths and Tenenbaum (2006) made only a single judgment about each phenomenon, it was possible to accurately model the aggregate judgments by making this assumption. However, another possibility that is equally consistent with the data is that each individual has a large pool of exemplars available, and only samples a small number in making a given prediction. In this case, a small number of exemplars are used in order to make the Bayesian computation efficient, not because they represent the complete knowledge of the learner. These two possibilities can be differentiated by conducting an experiment in which individuals make multiple judgments about a given phenomenon. If participants only have access to a small number of exemplars, they produce very similar responses for a range of values of $t$, while if they are sampling different sets of exemplars on different trials, their responses should increase as a function of $t$ in a way that is consistent with applying Bayesian inference. Lewandowsky, Griffiths, and Kalish (in press) conducted such an experiment, and found support for the latter hypothesis.

### Simulation 4: Concept learning

The simulations we have presented so far correspond to cases where Bayesian inference is performed with a hypothesis space that contains only hypotheses that correspond to continuous quantities (formant values, the size of consequential regions, the extent or duration of everyday phenomena). However, Bayesian inference is also carried out with hypothesis spaces in which each hypothesis is discrete, and qualitatively different from other hypotheses. The "number game" of Tenenbaum (1999; Tenenbaum & Griffiths, 2001) is a good example. This game is formulated as follows: Given natural numbers from 1 to 100, if a number or set of numbers $x$ belongs to an unknown set $C$, what is the probability that another number $y$ also belongs to the same set? For example, if the numbers $\{59, 60, 61, 62\}$ all belong to an unknown set, what is the probability that 64 belongs to that set? What about 16?

The problem of determining whether $y$ belongs to the same set as $x$ is another instance of the problem of generalization, and can be answered using a similar Bayesian inference. Our data are the knowledge that $x$ belongs to the set $C$, and our hypotheses concern the nature of $C$. Since $C$ is unknown, we should sum over all possible hypotheses $h$ in the hypothesis space $\mathcal{H}$ when evaluating whether $y$ belongs to $C$,

$$p(y \in C|x) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|x) = \sum_{h \in \mathcal{H}} \mathbf{1}(y \in h)p(h|x), \tag{21}$$

where $\mathbf{1}(y \in h)$ is the indicator function of the statement $y \in h$, taking value 1 if this is true and 0 otherwise. In the analysis presented by Tenenbaum (1999; Tenenbaum & Griffiths, 2001), the likelihood $p(x|h)$ is proportional to the inverse of the size of $h$ (the "size principle") being $1/|h|$ if $x \in h$ and 0 otherwise. This corresponds to the uniform sampling assumption made in the previous two examples. A hypothesis space $\mathcal{H}$ containing a total of 6,412 hypotheses was used, including intervals of numbers spanning a certain range, even numbers, odd numbers, primes, and cubes.

The number game is challenging because any given number (say $x = 8$) is consistent with many hypotheses (not only intervals containing 8, but also hypotheses such as even numbers, cubic

numbers, number with final digit 8, etc.). Interestingly, the responses of human participants can be captured quite accurately with this Bayesian model (Figure 6 (a)). However, this involves instantiating all 6,412 hypotheses, calculating the likelihood for each rule and integrating over the product of the prior and likelihood. Such computations are challenging, so a mechanism that approximates the exact solution is desirable. Fortunately, the probability computed in Equation 21 is an expectation, and can be approximated by importance sampling and thus by an exemplar model.

The number game is another instance of a problem that requires the more general approximation scheme summarized in Equation 13. The hypotheses $h$ are candidates for the identity of the concept $C$, the data $d$ are the observation that $x$ belongs to $C$, and the function $f(h)$ that we want to evaluate the expectation of is the indicator function $\mathbf{1}(y \in h)$. We can approximate this expectation by sampling hypotheses $h_j$ from the prior $p(h)$, and reweighting those hypotheses by the likelihood $p(x|h)$, with

$$p(y \in C|x) \approx \frac{\sum_j \mathbf{1}(y \in h_j, x \in h_j) 1/|h_j|}{\sum_j \mathbf{1}(x \in h_j) 1/|h_j|}, \tag{22}$$

meaning that $p(y \in C|x)$ is just the ratio of the summed likelihoods of the hypotheses stored in memory that generate $y$ to the summed likelihoods of all hypotheses stored in memory.

Figure 6 (b) and (c) show generalization responses for different sets of numbers, $x$, for a single simulated participant. As in Simulation 3, we conducted simulations for both memory- and computation-limited approximations, with the latter case corresponding to generating sample hypotheses $h$ from the prior until a fixed number consistent with $x$ had been generated. The simulations used the same parameters as those in the full Bayesian model of Tenenbaum and Griffiths (2001), except the likelihood function assigns a small non-zero probability to all natural numbers from 1 to 100 for every hypothesis to ensure numerical stability. The results suggest that a small number of exemplars (20 and 50 for computation-limited and memory-limited respectively)

is sufficient to account for human performance. The memory-limited case needs more exemplars because not all exemplars are qualified hypotheses. Therefore, the effective number of exemplars, which determines the computational load, is small. The consistency of these results with the human judgments indicates that exemplar models provide a plausible mechanism that relies on reasonable memory and computational resources and can be used with highly structured hypothesis spaces.

To further evaluate the model, we compared the variance of the predictions produced by importance sampling with the variability among individuals on this task. Since the model predictions rely on a sample from the prior, there can be variability between simulated participants which we can compare with the variability among human participants. Moreover, we should expect to see specific simulated participants who produced behavior similar to that of specific human participants. Figure 7 (a) shows the variability among the eight participants analyzed by Tenenbaum (1999), together with the variability among 100 simulated participants (using the memory-limited case). Both human and simulated participants exhibit significant variability in their responses, particularly for the stimulus $x = \{60\}$. The patterns of responses also share some key features. For $x = \{60, 52, 57, 55\}$, since there is no specific numeric rule describing the set, most plausible hypotheses are intervals containing $x$. Therefore, we expect higher variability near the boundary of the set (ie. below 52 or greater than 60) and lower variability within the set. For $x = \{60, 80, 10, 30\}$, high variability in genealization to multiples of five and ten is observed in both human and simulated participants.

The variability seen in the human and simulated participants disagree in two respects. First, there is significant baseline variability in the human responses that is not captured by the model, especially for $x = \{60, 52, 57, 55\}$ and $x = \{60, 80, 10, 30\}$. After looking in detail at individual trials, we found that high baseline variability is partly due to inconsistent use of the rating scale (which ranged from 1-7) to express "low probability." For example, for $x = \{60, 52, 57, 55\}$, two out of eight participants gave minimum responses of 2 out of 7, while the other six used the full range and had minimum responses of 1. A second point of difference between the human and

simulated responses is in the use of the "square numbers" hypotheses with $x = \{81, 25, 4, 36\}$. The model displays greater variability than seen among human participants when generalizing to other squares from this set. This is due to the fact that the memory-limited exemplar model is not guaranteed to sample the "square numbers" rule in every trial, while the educated participants used by Tenenbaum (1999) consistently recognized this mathematical rule.

For a closer look at the way that variability manifests in the model, we examined whether it was possible to find patterns of predictions that matched the behavior of individual participants. Figure 7 (b) shows some close correspondences between human and simulated participants. Each row shows the responses of a different human participant, together with the closest-matching responses chosen from the 100 simulated participants used in our analysis of variability. In each case the correlation between human and simulated participants was greater than $r = 0.95$, and many of the details of the responses are in correspondence. For example, in the case of $x = \{60\}$, this individual evaluated multiple hypotheses such as intervals, multiple of 10 and multiples of 6, and a similar pattern appears in the model predictions.

### Simulation 5: Category effects on reconstruction from memory

Retrieving or reconstructing items from memory can also be formulated as a problem of statistical inference, with Bayes' rule being used to evaluate which item in memory might correspond to a particular cue (Anderson & Milson, 1989; Shiffrin & Steyvers, 1997; Hemmer & Steyvers, 2009; Huttenlocher et al., 2000). Examining how this kind of probabilistic inference can be approximated using an exemplar model has the potential to be particularly informative, since exemplar models themselves are based on memory. This creates an opportunity to consider how exemplars come to be stored in memory, and what role statistical inference plays in this process.

We will focus on the problem of reconstructing items from memory, and in particular on a study by Huttenlocher et al. (2000, Experiment 1) examining how the relative frequencies of items within a category can be used to improve accuracy in reproducing stimuli. In this study

participants learned the distribution associated with a novel one-dimensional stimulus (the width of a schematic fish). The form of this distribution varied across participants. Some participants learned a single category, which was associated with either a uniform or a Gaussian distribution on fish width. Other participants learned two categories, each of which was associated with one half of the uniform distribution used in the one category case (the categories thus corresponded to "slender" and "fat" fish). During training, participants were briefly shown a stimulus, and then asked to reproduce that stimulus from memory (having been provided with its category label). Reconstructions were produced by adjusting the size of a schematic fish until participants felt that they had matched the size of the original stimulus.

Reconstructing a stimulus from memory can be analyzed as a Bayesian inference. Returning to the very first example of Bayesian inference we considered in the paper, we might assume that the observed stimulus $x$ is taken as a noisily perceived instance of some true stimulus $x^*$, with the noise process described by the distribution $p(x|x^*)$. The prior distribution on $x^*$ is provided by the category $c$, which is associated with a distribution $p(x^*|c)$. The best reconstruction of $x^*$, in the sense of minimizing the squared error between the reconstruction and the true value, is the posterior expectation of $x^*$ given $x$ and $c$,

$$E[x^*|x,c] \quad = \quad \int x^* p(x^*|x,c) dx^*, \tag{23}$$

where the posterior distribution $p(x^*|x,c)$ is calculated using Bayes' rule. Huttenlocher et al. (2000) explicitly tested this model of reconstruction from memory, arguing that using category information to guide reconstruction should increase accuracy.

The problem of reconstruction from memory is of exactly the same form as the stimulus denoising problem we used to demonstrate the equivalence between importance sampling and exemplar models. The expectation in Equation 23 can be approximated by storing a set of exemplars $x_j^*$ in memory, sampled from the prior $p(x^*|c)$, and then activating those exemplars in

proportion to the likelihood $p(x|x^*)$. Huttenlocher et al. (2000) assumed that the likelihood was a Gaussian distribution with a mean at $x^*$, and explored several different prior distributions $p(x^*|c)$. In each case, the Bayesian inference required to reconstruct a stimulus from memory can be approximated using an exemplar model of the form specified in Equation 12.

Although this analysis of reconstruction from memory is similar to that for the perceptual magnet effect, there are two important differences. First, category labels are given explicitly in the case of reconstruction, but are unknown in the perceptual magnet effect. Second, and perhaps more importantly, the experiments conducted to explore these phenomena differ in how the relevant priors were acquired. The prior distribution on speech sounds was established before the experiment exploring the perceptual magnet effect, as a result of learning the distributions associated with these sounds in English. In contrast, the prior being used to reconstruct the stimuli in the experiment conducted by Huttenlocher et al. (2000) is learned on the fly, through the process of forming the reconstructions. The reconstruction produced on one trial might thus play the role of a stored exemplar on a later trial.

To explore the effects of incrementally building a set of exemplars over time, we conducted a series of simulations of this study in which we used a variant on the standard exemplar model. The reconstruction of the first stimulus seen by each simulated participant was taken to be exactly equal to that stimulus. Each subsequent stimulus was reconstructed using an exemplar model with the previous $n$ stimuli as exemplars (or all stimuli, if fewer than $n$ have been observed), including the observed value of the current stimulus. Following Huttenlocher et al. (2000), the likelihood $p(x|x^*)$ was taken to be a normal distribution with mean $x^*$ and variance $\sigma^2$. The resulting model has two parameters: the noise level $\sigma^2$, and the memory capacity $n$. Our simulations varied these two parameters, with $n = \{1, 2, 5, 10, \infty\}$ and $\sigma = \{1, \ldots, 10\}$ pixels.[4]

Figure 8 shows the results of these simulations. In each case, we plot the bias in reconstruction for stimuli of different widths, defined to be the difference between the width of the reconstruction and the width of the stimulus. In general, stimuli that are smaller than the mean of a

category show a positive bias and stimuli that are larger show a negative bias, consistent with reconstructions moving towards the mean of each category. This effect comes out in all of our models, being the basic prediction resulting from a Bayesian analysis of this problem. However, the results also show how the exemplar models capture some subtle characteristics of the data. For example, in the normal prior condition (the middle row of the figure), a full Bayesian model would predict that bias is a linear function of fish width. This prediction is quite clearly reflected in the results for $n = \infty$, which most closely approximates exact Bayesian inference. In contrast, both the human data and the models with smaller values of $n$ show a non-linear function, with bias reduced for more extreme stimuli. To understand this effect, we should note that the current observation $x$ is always included as an exemplar in producing the reconstruction of $x^*$. Thus, when $x$ takes an extreme value lying at the tails of the prior, it is often over-weighted since recent observations are unlikely to lie in proximity to this extreme value. In this case, the reconstruction of $x^*$ relies more on $x$ itself, resulting in smaller bias.

## General Discussion

The formal correspondence that we have shown to exist between exemplar models and importance sampling suggests a way to solve the computationally challenging problem of probabilistic inference using a common computational model of psychological processes. Our five simulations illustrate how this approach can be applied in a range of settings where probabilistic models have previously been proposed. Simulation 1 showed that exemplar models can be used to perform Bayesian inference for a simple speech perception problem, providing an account of the perceptual magnet effect that does not require parametric assumptions about the distribution of speech sounds associated with phonetic categories, or any form of learning of these distributions. Simulation 2 demonstrated that a similar approach could approximate the predictions of Shepard's (1987) classic analysis of generalization. Simulation 3 examined how exemplar models could be used in predicting the future. Simulation 4 extended our analysis to a case where hypotheses

represent discrete, qualitatively different accounts of observed data. Finally, Simulation 5 considered how exemplars might be recruited in the course of an experiment, and showed that this approach could account for the results of a study of reconstruction from memory.

In the remainder of the paper, we discuss three issues raised by these results. First, while our simulations show that exemplar models can be used to approximate Bayesian inference in a range of settings, this approach will not provide good approximations in all cases. The relationship with importance sampling makes it possible to clearly state in which cases we expect this to be an effective approximation scheme. Second, none of the cases we consider involve any kind of dynamics, with the hypothesis space remaining static over time. Since some cognitive problems require dealing with hypothesis spaces that change in size and content over time, we outline how our approach can be extended to accommodate this situation. Finally, we consider some of the broader implications of the correspondence between exemplar models and importance sampling that we have identified in this paper, viewing this result as just one instance of a more general approach towards connecting rational models of cognition with psychological processes.

*The limits of importance sampling*

While importance sampling is widely used to approximate probabilistic inference, it is not appropriate for all problems. As discussed above, the quality of the approximation provided by importance sampling depends on the relationship between the target distribution $p(y)$, the function $g(y)$ for which we want to find an expected value, and the proposal distribution $q(y)$. In particular, we want the proposal distribution to assign high probability to values of $y$ for which both $p(y)$ and the contribution of $g(y)$ to the expectation are large, and low probability to other values of $y$ (see footnote 2 for details). Otherwise, samples from the proposal distribution may not correspond to values of $y$ that make a large contribution to the expectation of $g(y)$.

The relationship between importance sampling and exemplar models that we have identified relies on the assumption that the exemplars are drawn from the prior (ie. that the prior is used as a

proposal distribution). This makes it easy to identify the limitations of this approach: Bayesian inference can only be approximated effectively using the kind of exemplar models we have considered in this paper when there is a reasonably close match between the posterior and the prior. This will be the case when the data are relatively uninformative, meaning that the posterior does not deviate significantly from the prior. Data can be uninformative because of small sample size, or because of a high level of uncertainty (as reflected in the likelihood). All of the settings we explored in our simulations met this criterion, requiring an inference to be made on the basis of only one or at most a handful of stimuli.

One way to extend the range of problems for which exemplar models yield approximations to Bayesian inference might be to remove the assumption that the exemplars are drawn from the prior. While we have focused on the equivalence between Equations 3 and 12, the exemplar-based computations represented by Equation 3 are also equivalent to those used in the more general formulation of the importance sampler in Equation 10. Thus, exemplar models can be used to approximate expectations over a distribution $p(x^*|x)$ when the exemplars are generated from any distribution $q(x^*)$, provided the similarity function used to activate each exemplar is proportional to $p(x^*|x)/q(x^*)$. When $q(x^*) = p(x^*)$, we obtain the class of models analyzed in this paper. However, relaxing this assumption broadens the range of proposal distributions that can be used, and may make it possible for exemplar models to produce efficient approximations to Bayesian inference across a wider range of problems.

*Approximating dynamic inferences*

A second limitation of the approach that we have presented in this paper is that it is only appropriate in cases where the hypothesis space is static, with the same hypotheses being used in multiple inferences. The simple strategy of using a stored set of hypotheses does not work in cases where the hypothesis space itself changes over time, and results in a particularly poor approximation when that hypothesis space grows with the number of observations. One example

where such a problem arises is dividing a set of observations into clusters, as in Anderson's (1990, 1991) rational model of categorization. In this model, the hypothesis space consists of all possible clusterings of a set of observations. This hypothesis space has to be revised with each new observation, reflecting all of the ways in which that observation could be added to the existing clusters. Not only does the hypothesis space change over time, but it grows super-exponentially in the number of observations.

While exemplar models are not appropriate for this situation, they are closely related to another Monte Carlo method that can be extremely effective for approximating dynamic inferences. This method, known as *particle filtering*, translates importance sampling into a dynamic setting. The basic idea is that the posterior distribution over hypotheses after $n$ observations should be closely related to the posterior distribution after $n+1$ observations, in the same way that the prior and posterior were closely related in the examples we considered above. The posterior after $n+1$ observations can thus be approximated by importance sampling, using a proposal distribution based on the posterior after $n$ observations. This idea can be applied recursively: while we may not know the posterior after $n$ observations, we can approximate this by importance sampling too, using a proposal distribution based on the posterior after $n-1$ observations, and so on. A particle filter thus consists of a set of samples that evolves through time, with samples from the posterior distribution after $n$ observations being used to generate samples from the posterior distribution after $n+1$ observations.

Particle filters share with the models that we have discussed in this paper the idea of approximating a probability distribution with a small number of samples. However, the models we have considered all assume that these samples are fixed exemplars stored in memory, while a particle filter dynamically constructs a set of samples in response to the information provided by a sequence of observations. Despite this difference, the basic components of a particle filter are very similar to the components of an exemplar model, requiring activation of hypotheses in proportion to their likelihood, normalization, and random selection. As a consequence, particle filters may

provide a psychologically plausible scheme for approximating Bayesian inference in dynamic settings. This idea has been explored in the context of the rational model of categorization by Sanborn et al. (2006), and similar models have been proposed as explanations of change point detection (Brown & Steyvers, 2009), associative learning (Daw & Courville, 2008), sentence processing (Levy, Reali, & Griffiths, 2009), and reinforcement learning (Yi, Steyvers, & Lee, in press).

*Rational process models*

Probabilistic models of cognition are typically expressed at Marr's (1982) computational level, analyzing learning, reasoning, and perception in terms of ideal solutions to abstract problems posed by the environment. This is at odds with much of the history of cognitive psychology, in which theories are typically expressed at the level of representation and algorithm. As Marr noted, these two levels should not be considered independent of one another: findings at one level provide constraints on theories at the other. However, despite a few notable exceptions (e.g., Kruschke, 2006), there has been little exploration of the relationship between probabilistic models of cognition and psychological process models.

The connection between importance sampling and exemplar models that we have established in this paper hints at a strategy that might help to establish a more general link between probabilistic models formulated at the computational level and psychological process models expressed at the algorithmic level. The computational challenges posed by probabilistic inference do not arise just as an obstacle for rational models of cognition: they also appear whenever a computer scientist or statistician wants to work with a probabilistic model. As a consequence, researchers in computer science and statistics have developed a variety of schemes for efficiently approximating probabilistic inference. Importance sampling is just one of these schemes, and the fact that it can be implemented in a psychologically plausible way suggests that there may be other approximate algorithms for probabilistic inference that are candidate explanations for how people

might address the computational challenges posed by rational models of cognition.

In embodying an effective solution to the problem of approximating probabilistic inference, and making use of psychological notions common in mechanistic process models, exemplar models are an instance of a "rational" process model. Such rational process models push the principle of rationality embodied in existing rational models of cognition a level deeper. Rational models of cognition apply the principle of rationality – the assumption that optimal solutions are informative about human behavior – at the computational level. Rational process models apply a similar principle at the level of representation and algorithm, assuming that the psychological processes that are used to approximate probabilistic inference represent efficient solutions to this problem. As noted above, particle filters are another instance of a rational process model, but the great diversity of efficient approximation algorithms for probabilistic inference suggests that there may be many other psychologically plausible mechanisms for solving this problem that are still to be discovered.

In providing a connection between abstract probabilistic models of cognition and psychological processes, rational process models also have the potential to help us understand the neural mechanisms that underlie probabilistic computation. For example, our analysis of the perceptual magnet effect revealed that approximating Bayesian inference by importance sampling resulted in a model that was extremely similar to a neural network model proposed by Guenther and Gjaja (1996). This connection is valuable in two ways: It shows how such a neural network could be used to approximate Bayesian inference, and it provides a high-level explanation of why this neural mechanism produces the perceptual magnet effect. We anticipate that similar connections will exist in other domains, particularly given the close correspondence between exemplar models and neural network architectures such as radial basis function networks (Kruschke, 1992; Shi & Griffiths, in press).

## Conclusion

We have presented both theoretical results and simulations showing that exemplar models provide a simple, psychologically plausible mechanism for performing at least some kinds of Bayesian inference. Our theoretical results indicate that exemplar models can be interpreted as a form of importance sampling, and can thus implement an approximation to Bayesian inference. Our simulations demonstrate that this approach produces predictions that correspond reasonably well with human behavior, and that relatively few exemplars are needed to provide a good approximation to the true Bayesian solution in at least five settings.

The approach that we have taken in this paper represents one way of addressing questions about the mechanisms that could support probabilistic inference. Our results suggest that exemplar models are not simply process models, but rational process models – an effective and psychologically plausible scheme for approximating statistical inference. This approach pushes the principle of optimality that underlies probabilistic models down to the level of mechanism, and suggests a general strategy for explaining how people perform Bayesian inference: Look for connections between psychological process models and approximate inference algorithms developed in computer science and statistics.

## References

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

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

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703-719.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49-67.

Daw, N., & Courville, A. C. (2008). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20.* Cambridge, MA: MIT Press.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968-986.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752-782.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, *5*, 171-189.

Gigerenzer, G., & Todd, P. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767-773.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik &

L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*, 1111-1121.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*, 189-202.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220-241.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*, 553-562.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563-607.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677-699.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606-608.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 937–944).

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (in press). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358-368.

Luce, R. D. (1959). *Individual choice behavior*. New York: John Wiley.

Ma, W. J., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432-1438.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.

Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, *40*, 342-347.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393-418.

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

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization.

In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika*, *27*, 124-140.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shi, L., Feldman, N., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

Shi, L., & Griffiths, T. L. (in press). Neural implementation of hierarchical Bayesian inference by importance sampling. In J. Lafferty & C. K. I. Williams (Eds.), *Advances in Neural Information Processing Systems 22.*

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Simon, H. A. (1957). *Models of man*. New York: Wiley.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*, 3-21.

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*. Unpublished doctoral dissertation, Massachussetts Institute of Technology, Cambridge, MA.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Yi, S. K. M., Steyvers, M., & Lee, M. D. (in press). Modeling human performance on restless bandit problems using particle filters. *Journal of Problem Solving*.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301-308.

Zemel, R. S., Dayan, P., & Pouget, A. (1997). Probabilistic interpretation of population codes. In *Advances in Neural Information Processing Systems 9.* Cambridge, MA: MIT Press.

## Author Note

This research was supported by grant number FA9550-07-1-0351 from the Air Force Office of Scientific Research. Preliminary results from Simulations 1 and 4 were presented at the 30th Annual Conference of the Cognitive Science Society (Shi, Feldman, & Griffiths, 2008). We thank Michael Lee, Mike Mozer, and an anonymous reviewer for their comments on a previous version of this manuscript.

## Notes

[1]Our analysis requires that this similarity measure has a finite integral, with $\int s(x,x^*)dx$ equal to a fixed constant for all $x^*$. This assumption is satisfied by similarity functions such as the exponential or Gaussian that are typically used in exemplar models.

[2]If the function $g(y)$ takes on its largest values in regions where $p(y)$ is small, the variance of the simple Monte Carlo estimate can be large. An importance sampler can have lower variance than simple Monte Carlo if $q(y)$ is chosen to be complementary to $g(y)$. In particular, the asymptotic variance of the sampler is minimized by specifying $q(y)$ as

$$q(y) \propto |g(y) - E_p[g(y)]|\, p(y). \tag{24}$$

This is not a practical procedure, since finding this distribution requires computing $E_p[g(y)]$, but the fact that the minimum variance sampler need not be $p(y)$ means that importance sampling can provide a better estimate of an expectation than simple Monte Carlo.

[3]Because listeners only hear noisy speech sounds $S$, they may not have direct access to a sample from $T$. Storing samples from $S$ instead of $T$ produces the same qualitative effect, though the computation is no longer optimal. Alternatively, listeners may be able to bootstrap a sample from $T$ by using multiple cues to reduce the amount of noise and by using subsequent percepts to update stored values. We return to the problem of recruiting exemplars during inference in Simulation 5.

[4]We also conducted simulations in cases where perceptual noise was considered and reconstructed stimuli, instead of original stimuli, were taken as exemplars. All of these variations produced similar results.

Table 1
*Comparison of Approximation Schemes with Exact Bayes and Human Data*

| Error score | Number of exemplars | | |
|---|---|---|---|
| | 5 exemplars | 10 exemplars | 50 exemplars |
| $\sum \lvert t^*_{ML} - t^*_{Bayes} \rvert$ | 4.2003 | 2.3333 | 1.2366 |
| $\sum \lvert t^*_{ML} - t^*_{human} \rvert$ | 8.3023 | 7.0858 | 6.6757 |
| $\sum \lvert t^*_{CL} - t^*_{Bayes} \rvert$ | 3.5601 | 1.8620 | 1.0798 |
| $\sum \lvert t^*_{CL} - t^*_{human} \rvert$ | 7.8566 | 6.8283 | 6.1023 |
| $\sum \lvert t^*_{SA} - t^*_{Bayes} \rvert$ | 1.4706 | 1.7449 | 2.3050 |
| $\sum \lvert t^*_{SA} - t^*_{human} \rvert$ | 6.8043 | 6.0633 | 6.5741 |

Note: Subscripts correspond to memory limited (ML), computation limited (CL), sampling from the posterior (SA), and true Bayesian and human estimates of $t^*$. Error scores were summed across values of $t$ for each prior, normalized as described in the text, and then summed across priors. The error score for the full Bayesian model, $\sum \lvert t^*_{human} - t^*_{Bayes} \rvert$, was 6.2626.

## Figure Captions

*Figure 1.* Approximating Bayesian inference by importance sampling using the prior $p(x^*)$ as the surrogate distribution. The true value of a stimulus $x^*$ is recovered from a noisy observation $x$ (represented by the gray dot). (a) Exemplars $x_j^*$ are sampled from the prior $p(x^*)$. Each bar marks the location of an exemplar, and the solid black line shows the prior. (b) The $x_j^*$ are weighted by a Gaussian likelihood function $p(x|x_j^*)$. Since the Gaussian is symmetric in $x$ and $x^*$, the weights assigned to the exemplars fall off as a Gaussian function around $x$, here plotted as a solid gray line. (c) The expectation is the weighted average of the $x_j^*$. Compared with $x$, the estimate $E[x^*|x]$ is shifted towards a region that has higher probability under the prior.

*Figure 2.* The variance of the importance weights in approximating posterior expectations depends on how much probability mass is shared between prior and posterior. Different patterns are observed if posterior and prior distributions are (a) strongly overlapping, (b) non-overlapping or (c) partially overlapping. In these figures, the importance weights have been normalized to make it clear what proportion of the expectation depends on each sample. Greater overlap between prior and posterior results in lower variance in the importance weights, use of a larger set of samples, and consequently a better approximation.

*Figure 3.* Locations of stimuli in perceptual space from Iverson and Kuhl's (1995) multidimensional scaling data and from a single hypothetical subject (open circles) and the middle 50% of hypothetical subjects (solid lines) using an exemplar model in which perception is based on (a) ten and (b) fifty exemplars. The labels $\mu_{/i/}$ and $\mu_{/e/}$ show the locations of category means in the model. Parameter values were those used by Feldman, Griffiths, and Morgan (2009).

*Figure 4.* Exemplar models approximate generalizations functions for six different prior distributions on the size of consequential regions, corresponding to the six priors originally considered by Shepard (1987). Each generalization function shows how the probability of

generalizing a property from an observed stimulus **0** to a new stimulus **x** decreases with the psychological distance between the stimuli. In this one-dimensional case, if we take **0** to be the origin, the psychological distance corresponds directly to the value of **x**. Prior distributions are shown as inset shaded curves, reproducing Figure 3 of Shepard (1987). Analytical results for the form of the generalization function are provided on the top of each inset prior, and are plotted in the dotted curve. An approximating exponential generalization function is plotted as a smooth curve. Exemplar models using 20 and 100 hypotheses sampled from the prior (corresponding to circles and asterisks respectively) provide a good approximation to these theoretical predictions.

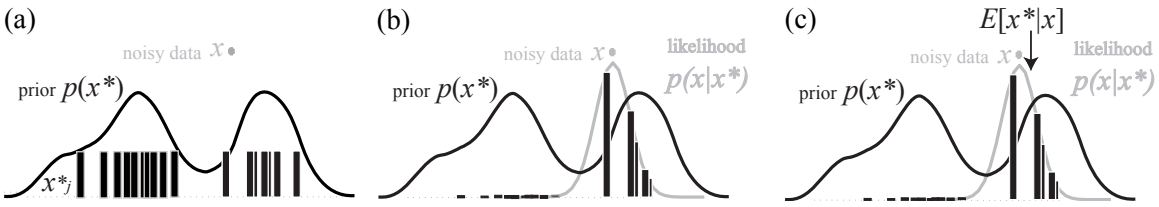
*Figure 5.* Simulations of prediction on everyday cognition, data from Griffiths and Tenenbaum (2006). The first row is the prior distribution of each dataset. The second to fourth rows are simulations with 5, 10 and 50 exemplars for memory-limited and computation-limited exemplar models, as well as sampling from the posterior. The solid line shows the optimal responses given the prior distribution, and the black dots are the responses of human participants. For both simulations and human data, the plot markers indicate the median response across a population of 50 simulated participants. Error bars show a 68% confidence interval computed by 1000 sample bootstrap.

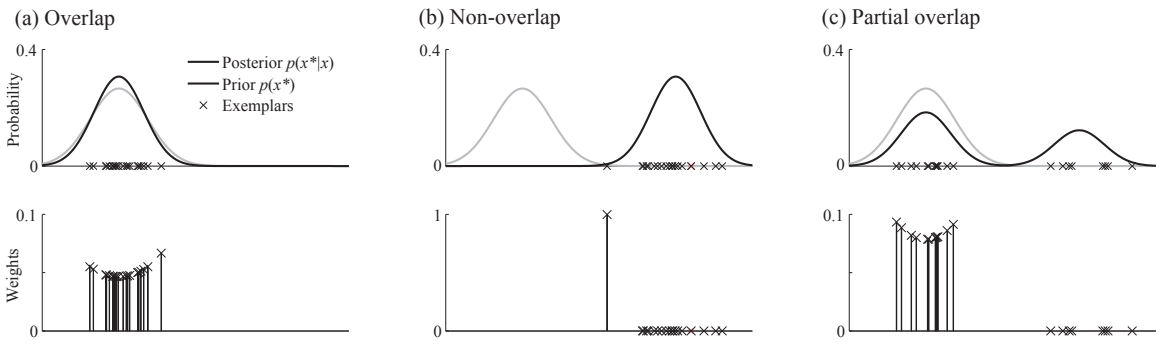
*Figure 6.* Simulations (dashed line) and behavioral data from Tenenbaum (1999) (gray bars) for the number game. The full Bayesian model uses 6,412 hypotheses. Results of computation-limited (20 exemplars) and memory-limited (50 exemplars) exemplar models are based on a single simulated participant with a set of hypotheses (exemplars) sampled from the prior. Models are tested under conditions suggesting single point generalization $x = 60$, a consecutive interval $x = \{60, 52, 57, 55\}$, multiples of 10 $x = \{60, 80, 10, 30\}$ and squares $x = \{81, 25, 4, 36\}$.

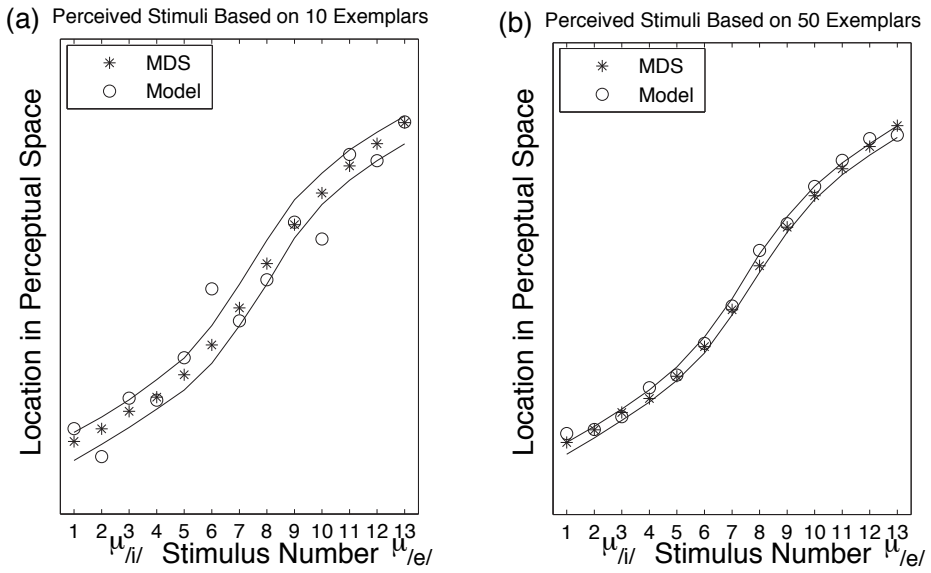
*Figure 7.* Variability across human and simulated participants in the number game. (a) The standard deviation of the ratings produced by eight human participants in the number game

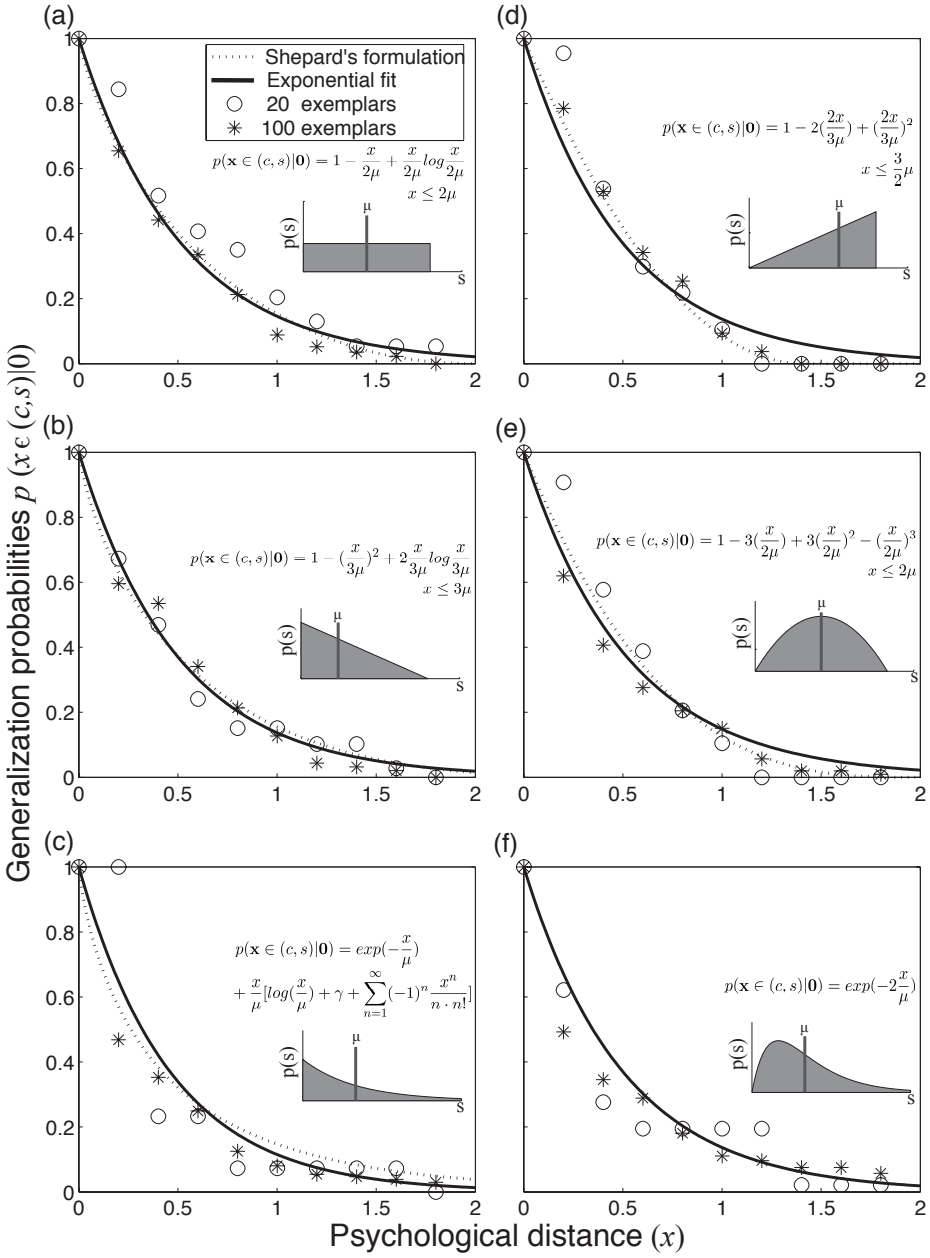
(denoted with asterisks) is compared with the standard deviation of the posterior probabilities produced by 100 simulated subjects. (b) Responses from four human participants, compared with the closest matching simulated participants from the pool of 100 used in evaluating variability.

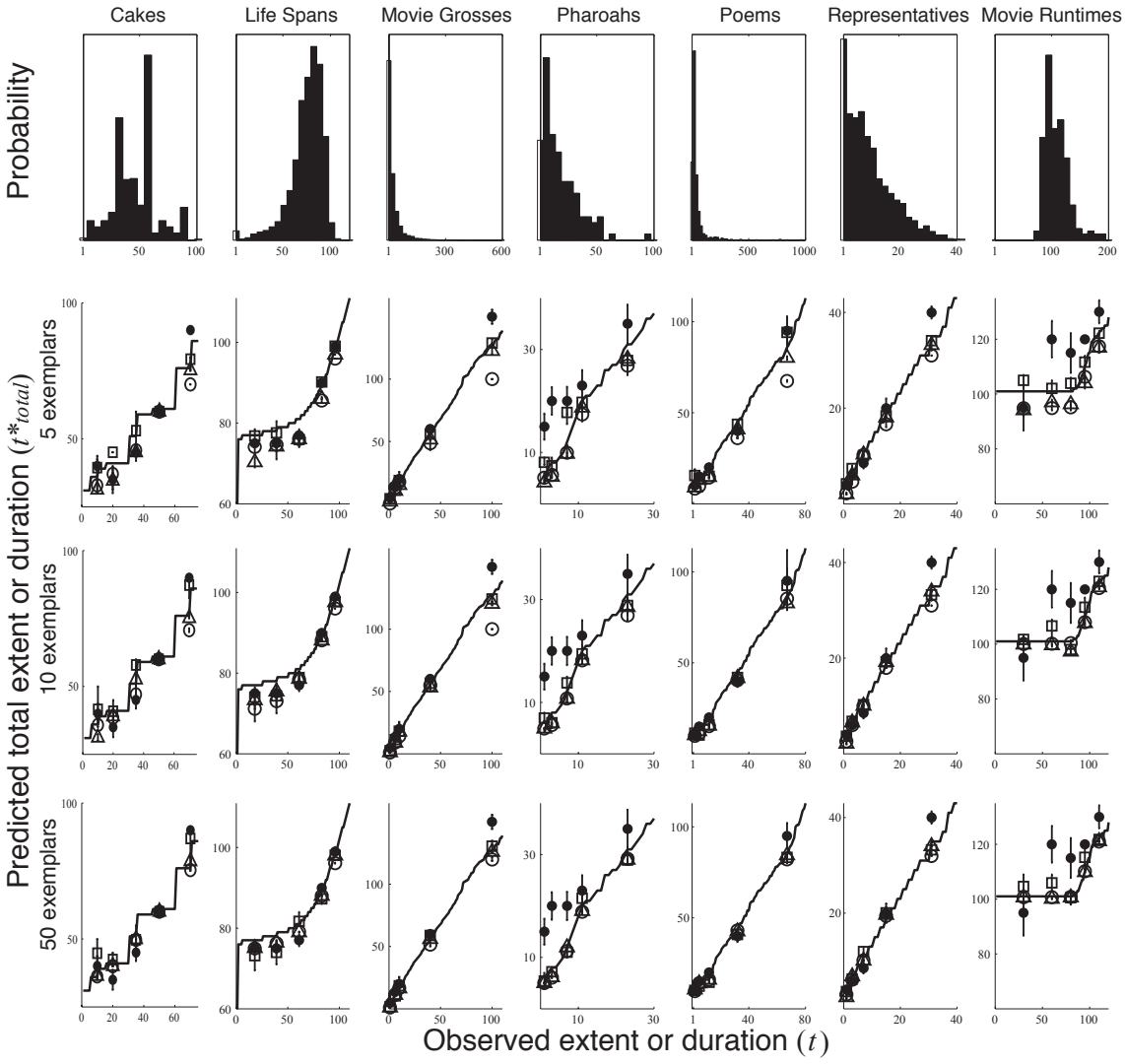
*Figure 8.* Reconstruction from memory with online recruitment of exemplars. (a) The left column shows the average bias in the reconstructed stimuli produced by participants (measured as the difference between the actual and reconstructed width of fish, in pixels) as a function of actual width. The rows show reconstructions produced for three prior distributions: a single category following a uniform and a normal distribution, and two categories following uniform distributions. Data are from Huttenlocher et al. (2000, Experiment 1). The remaining columns show simulations using exemplar models with a memory capacity of 1, 2, 5, 10 and $\infty$ exemplars. Data were generated in a way that was consistent with the original experiment, and the results show an average across 10 simulated participants with 192 trials per participant. The only free parameter, the assumed noise level $\sigma^2$, is specified by minimizing mean squared error (MSE) in each case. (b) Sensitivity of the results to memory capacity and recall noise. In the upper panel, memory capacity (in number of exemplars) is fixed and $\sigma^2$ is chosen to minimize MSE. Interestingly, MSE grows with increasing memory capacity, suggesting that a limited memory model ($< 10$ exemplars) is consistent with human behavior. In the lower panel, the effect of different noise levels $\sigma^2$ is examined, optimizing memory capacity. For all three priors, the error curves have concave bell shape and share a region of minimum error, suggesting that a single assumed noise level can account for results in all three conditions.

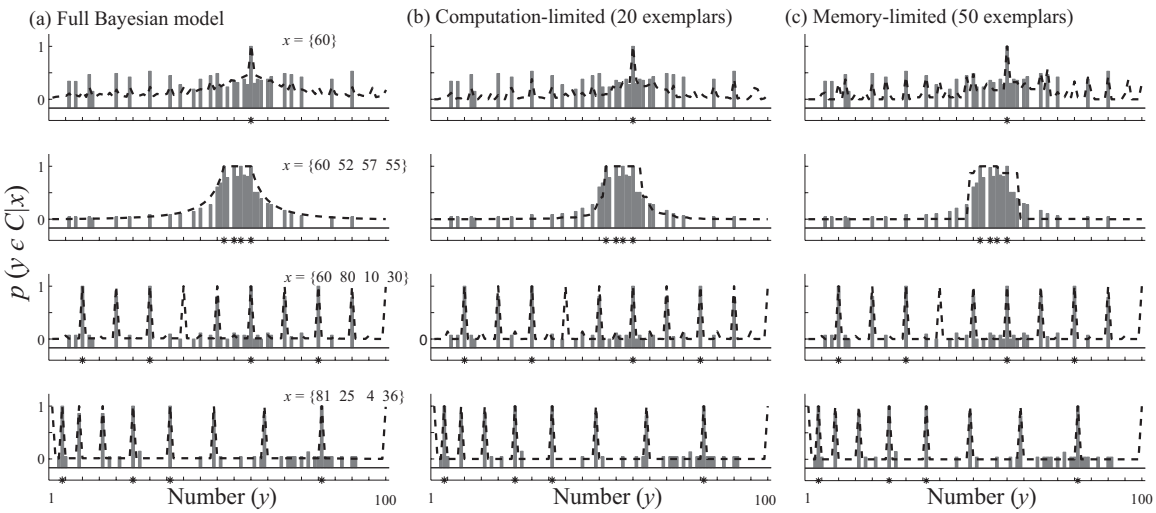

(a)
noisy data $x$
prior $p(x^*)$
$x^*_j$
(b)
noisy data $x$
likelihood
$p(x|x^*)$
prior $p(x^*)$
(c)
$E[x^*|x]$
noisy data $x$
likelihood
$p(x|x^*)$
prior $p(x^*)$

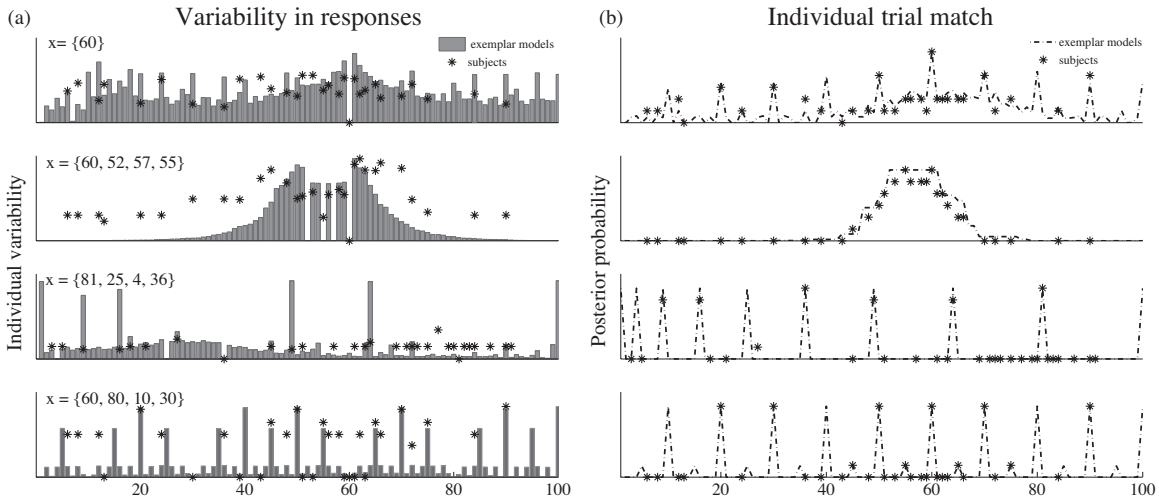

(a) Overlap
(b) Non-overlap
(c) Partial overlap
Probability
Weights
Posterior $p(x^*|x)$
Prior $p(x^*)$
Exemplars
0.4
0
0.1
1

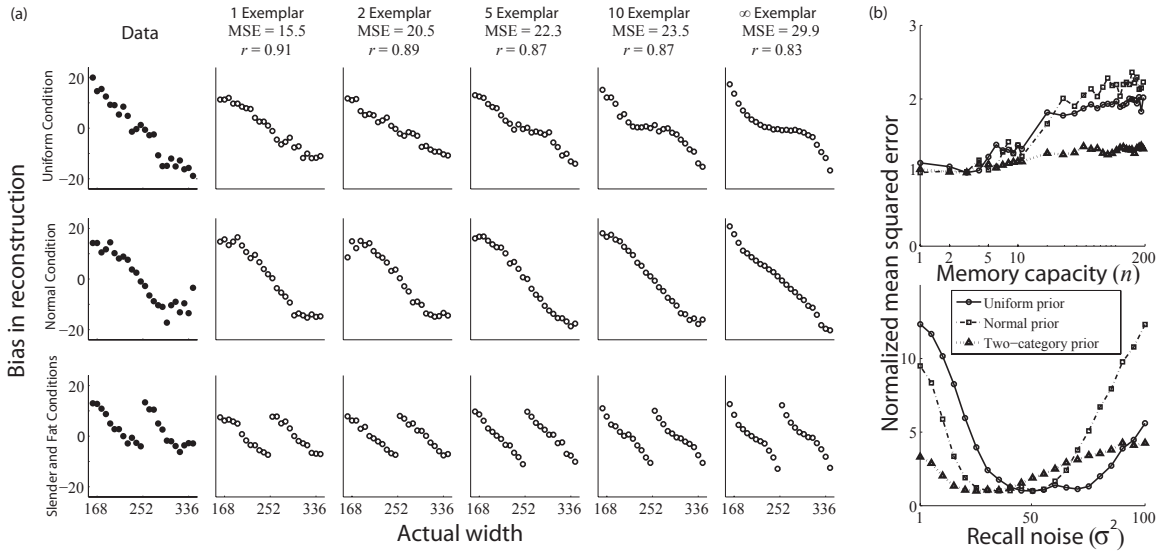

(a) Perceived Stimuli Based on 10 Exemplars
MDS
Model
Location in Perceptual Space
1 2 3 4 5 6 7 8 9 10 11 12 13
μ/i/ Stimulus Number μ/e/
(b) Perceived Stimuli Based on 50 Exemplars
MDS
Model
Location in Perceptual Space
1 2 3 4 5 6 7 8 9 10 11 12 13
μ/i/ Stimulus Number μ/e/


(a)
(b)
(c)
(d)
(e)
(f)
Shepard's formulation
Exponential fit
20 exemplars
100 exemplars
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = 1 - \frac{x}{2\mu} + \frac{x}{2\mu} log \frac{x}{2\mu}$
$x \leq 2\mu$
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = 1 - (\frac{x}{3\mu})^2 + 2\frac{x}{3\mu} log \frac{x}{3\mu}$
$x \leq 3\mu$
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = exp(-\frac{x}{\mu}) + \frac{x}{\mu}[log(\frac{x}{\mu}) + \gamma + \sum_{n=1}^{\infty}(-1)^n \frac{x^n}{n \cdot n!}]$
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = 1 - 2(\frac{2x}{3\mu}) + (\frac{2x}{3\mu})^2$
$x \leq \frac{3}{2}\mu$
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = 1 - 3(\frac{x}{2\mu}) + 3(\frac{x}{2\mu})^2 - (\frac{x}{2\mu})^3$
$x \leq 2\mu$
$p(\mathbf{x} \in (c,s)|\mathbf{0}) = exp(-2\frac{x}{\mu})$
p(s)
μ
s
Generalization probabilities $p\ (x \in (c,s)|0)$
Psychological distance $(x)$


Cakes
Life Spans
Movie Grosses
Pharoahs
Poems
Representatives
Movie Runtimes
Probability
Predicted total extent or duration ($t^*_{total}$)
5 exemplars
10 exemplars
50 exemplars
Observed extent or duration ($t$)
Prediction from true distribution
Human
Computation limited case
Memory limited case
Sampling


(a) Full Bayesian model
(b) Computation-limited (20 exemplars)
(c) Memory-limited (50 exemplars)
x = {60}
x = {60 52 57 55}
x = {60 80 10 30}
x = {81 25 4 36}
p (y ϵ C|x)
1
0
1
100
Number (y)


(a)
Variability in responses
(b)
Individual trial match
x= {60}
exemplar models
subjects
x = {60, 52, 57, 55}
x = {81, 25, 4, 36}
x = {60, 80, 10, 30}
Individual variability
Posterior probability
20
40
60
80
100


(a)
Data
1 Exemplar
MSE = 15.5
r = 0.91
2 Exemplar
MSE = 20.5
r = 0.89
5 Exemplar
MSE = 22.3
r = 0.87
10 Exemplar
MSE = 23.5
r = 0.87
∞ Exemplar
MSE = 29.9
r = 0.83
Bias in reconstruction
Uniform Condition
Normal Condition
Slender and Fat Conditions
20
0
−20
168
252
336
Actual width
(b)
Normalized mean squared error
Memory capacity (n)
Recall noise (σ²)
Uniform prior
Normal prior
Two−category prior
1
2
5
10
200
50
100