

# Unifying Rational Models of Categorization via the Hierarchical Dirichlet Process

**Thomas L. Griffiths (tom\_griffiths@berkeley.edu)**

Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650 USA

**Kevin R. Canini (kevin@cs.berkeley.edu)**

Department of Computer Science, University of California, Berkeley, Berkeley, CA 94720-1776 USA

**Adam N. Sanborn (asanborn@indiana.edu)**

Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA

**Daniel J. Navarro (daniel.navarro@adelaide.edu.au)**

School of Psychology, University of Adelaide, Adelaide SA 5005, Australia

## Abstract

Models of categorization make different representational assumptions, with categories being represented by prototypes, sets of exemplars, and everything in between. Rational models of categorization justify these representational assumptions in terms of different schemes for estimating probability distributions. However, they do not answer the question of which scheme should be used in representing a given category. We show that existing rational models of categorization are special cases of a statistical model called the hierarchical Dirichlet process, which can be used to automatically infer a representation of the appropriate complexity for a given category.

**Keywords:** rational analysis, categorization, Dirichlet process

Rational models of cognition aim to explain human behavior as an optimal solution to the computational problems posed by our environment (Anderson, 1990). Examining these computational problems provides a deeper understanding of the assumptions behind successful models of human cognition, and can lead to new models. In this paper, we pursue a rational analysis of category learning: inferring the structure of categories from a set of stimuli labeled as belonging to those categories. The knowledge acquired through this process can ultimately be used to make decisions about how to categorize new stimuli. Existing rational analyses of category learning (Anderson, 1990; Ashby & Alfonso-Reese, 1995; Rosseel, 2002) agree that the computational problem involved is one of *density estimation*: determining the probability distributions over stimuli associated with different category labels.

Viewing category learning as density estimation helps to clarify the assumptions behind the two main classes of psychological models: exemplar models and prototype models. Exemplar models assume that a category is represented by a set of stored exemplars, and categorization involves comparing new stimuli to the set of exemplars in each category (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). Prototype models assume that a category is associated with a single prototype and categorization involves comparing new stimuli to these prototypes (e.g., Reed, 1972). These approaches to category learning correspond to different strategies for density estimation, being nonparametric and parametric density estimation respectively (Ashby & Alfonso-Reese, 1995).

Despite providing insight into the assumptions behind models of categorization, existing rational analyses of cate-

gory learning leave a number of questions open. In particular, many categorization experiments have explored whether people represent categories with exemplars or prototypes. One desideratum for a rational account of category learning might be that it can indicate when a learner should choose to use one of these forms of representation over the other. The greater flexibility of nonparametric density estimation has motivated the claim that exemplar models are to be preferred as rational models of category learning (Nosofsky, 1998). However, nonparametric and parametric methods have different advantages and disadvantages: the greater flexibility of nonparametric methods comes at a cost of requiring more data to estimate a distribution. The decision as to which representation scheme to use should be determined by the stimuli presented to the learner, and existing rational analyses do not indicate how this decision should be made (although a similar argument is made by Briscoe & Feldman, 2006).

The question of how to represent categories is complicated by the fact that prototype and exemplar models are not the only options. A number of models have recently explored possibilities between these extremes, representing categories using clusters of several exemplars (Anderson, 1990; Vanpaemel, Storms, & Ons, 2005; Rosseel, 2002; Love, Medin, & Gureckis, 2004). The range of representations possible in these models emphasizes the importance of being able to identify an appropriate representation for a category from the stimuli themselves: with more options for the representation of categories, it becomes more important to be able to say which option a learner should choose.

Our goal in this paper is to build on previous rational analyses of category learning to provide not just a unifying *framework* which can be used to understand the assumptions behind existing models of categorization, but a unifying *model* of which these models are special cases. This model goes beyond previous unifying models of category learning (e.g., Rosseel, 2002; Vanpaemel et al., 2005) by providing a rational solution to the question of which representation should be chosen based purely on the structure of a category. These results are achieved by identifying connections between models of human category learning and ideas from nonparametric Bayesian statistics. In particular, we show that all of the models mentioned above can be viewed as variants of a stochastic

process called the *hierarchical Dirichlet process* (Teh, Jordan, Beal, & Blei, 2004).

Identifying the connection between models of human category learning and nonparametric Bayesian density estimation extends the scope of the rational analysis of category learning. It also provides a different perspective on human category learning. Rather than suggesting that people use one form of representation or another, our approach indicates how it might be possible (and, in fact, desirable) for people to switch between representations based upon the structure of the stimuli they observe, choosing the representation best justified by the available data. We illustrate this by modeling data from Smith and Minda (1998), in which people seem to shift from using a prototype representation early in training to using an exemplar representation late in training.

The plan of the paper is as follows. The next section summarizes exemplar and prototype models, and the idea of interpolating between the two. We then discuss existing rational models of categorization. This raises the question of how the models might be unified, which we address by turning to some ideas from nonparametric Bayesian statistics. Having established these ideas, we define a unifying rational model of categorization based on the hierarchical Dirichlet process, and show that this model can capture the shift from prototypes to exemplars in the data of Smith and Minda (1998).

## Exemplars and prototypes

Exemplar and prototype models were originally developed as accounts of the cognitive processes involved in categorization, incorporating different assumptions about how categories are represented and how this information is used. These models share the basic assumption that people assign stimuli to categories based on similarity. Given a set of  $N - 1$  stimuli with features  $\mathbf{x}_{N-1} = (x_1, x_2, \dots, x_{N-1})$  and category labels  $\mathbf{c}_{N-1} = (c_1, c_2, \dots, c_{N-1})$ , the probability that stimulus  $N$  with features  $x_N$  is assigned to category  $j$  is given by

$$P(c_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{\eta_{N,j} \beta_j}{\sum_c \eta_{N,c} \beta_c} \quad (1)$$

where  $\eta_{N,c}$  is the similarity of the stimulus  $x_N$  to category  $c$  and  $\beta_c$  is the response bias for category  $c$ . The key difference between the models is in how  $\eta_{N,c}$  is computed.

In an exemplar model (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), a category is represented by all of the stored instances of that category. The similarity of stimulus  $N$  to category  $j$  is calculated by summing the similarity of the stimulus to all stored instances of the category. That is,

$$\eta_{N,j} = \sum_{i|c_i=j} \eta_{N,i} \quad (2)$$

where  $\eta_{N,i}$  is a symmetric measure of the similarity between the two stimuli  $x_N$  and  $x_i$ . In a prototype model (e.g., Reed, 1972), a category  $j$  is represented by a single prototypical instance. In this formulation, the similarity of a stimulus  $N$  to

category  $j$  is defined to be

$$\eta_{N,j} = \eta_{N,p_j} \quad (3)$$

where  $p_j$  is the prototypical instance of the category and  $\eta_{N,p_j}$  is a measure of the similarity between stimulus  $N$  and the prototype  $p_j$ , as used in the exemplar model.

Realizing that these two models are opposite ends of a spectrum, Vanpaemel et al. (2005) observed that we can formalize a set of interpolating models by allowing the instances of each category to be partitioned into clusters, where the number of clusters  $K_c$  ranges from 1 to  $N_c$ . Then each cluster is represented by a prototype, and the similarity of a stimulus  $N$  to category  $j$  is defined to be

$$\eta_{N,j} = \sum_{k=1}^{K_j} \eta_{N,p_{j,k}} \quad (4)$$

where  $p_{j,k}$  is the prototype of cluster  $k$  in category  $j$ . When  $K_c = 1$  for all  $c$ , this is equivalent to the prototype model, and when  $K_c = N_c$  for all  $c$ , this is equivalent to the exemplar model. Thus, this generalized model, the Varying Abstraction Model (VAM), is more flexible than both the exemplar and prototype models, although it raises the problem of estimating which clustering people are actually using in a particular categorization task (for details, see Vanpaemel et al., 2005).

## Rational models of categorization

Following the methodology outlined by Anderson (1990), rational models of categorization explain human behavior in terms of adaptive solutions to a computational problem posed by the environment rather than the underlying cognitive processes. Existing analyses tend to agree that the basic problem is one of *prediction* – identifying the category label or some other unobserved property of an object using its observed properties (Anderson, 1990; Ashby & Alfonso-Reese, 1995; Rosseel, 2002). Focusing for the moment on the case of predicting category labels, as in most categorization experiments, the problem can be formulated as one of Bayesian inference: computing the probability that object  $N$  belongs to category  $j$  given the features and category labels of  $N - 1$  objects. Applying Bayes' rule, we can write

$$P(c_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{P(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) P(c_N = j | \mathbf{c}_{N-1})}{\sum_c P(x_N | c_N = c, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) P(c_N = c | \mathbf{c}_{N-1})} \quad (5)$$

with the posterior probability of category  $j$  being proportional to the product of the probability of an object with features  $x_N$  being produced from that category and the prior probability of choosing that category, taking into account the features and labels of the previous  $N - 1$  objects (assuming that only category labels influence the prior). Category learning, then, becomes a matter of determining these probabilities – a problem that is known as *density estimation*. Different rational models vary in how they approach this problem.

## Exemplar and prototype models

Ashby and Alfonso-Reese (1995) observed a connection between the Bayesian solution to the problem of categorization presented in Equation 5 and the way that choice probabilities are computed in exemplar and prototype models (i.e. Equation 1). Specifically,  $\eta_{N,j}$  can be identified with  $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$ , while  $\beta_j$  corresponds to the prior probability of category  $j$ ,  $P(c_N = j|\mathbf{c}_{N-1})$ . The difference between exemplar and prototype models thus comes down to different ways of estimating  $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$ .

The definition of  $\eta_{N,j}$  used in an exemplar model (Equation 2) corresponds to estimating  $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$  as the sum of a set of functions (known as “kernels”) centered on the  $x_i$  already labeled as belonging to category  $j$ , with

$$P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) \propto \sum_{i|c_i=j} f(x_N, x_i) \quad (6)$$

where  $f(x, x_i)$  is a probability distribution centered on  $x_i$ . This is a method that is widely used for approximating distributions in statistics, being a simple form of nonparametric density estimation (meaning that it can be used to identify distributions without assuming that they come from an underlying parametric family) called kernel density estimation.

The definition of  $\eta_{N,j}$  used in a prototype model (Equation 3) corresponds to estimating  $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$  by assuming that the distribution associated with each category comes from an underlying parametric family, and then finding the parameters that best characterize the instances labeled as belonging to that category. The prototype corresponds to these parameters. Again, this is a common method for estimating a probability distribution, known as parametric density estimation, in which the distribution is assumed to be of a known form but with unknown parameters.

## The Mixture Model of Categorization

The interpretation of exemplar and prototype models as different schemes for density estimation suggests that a similar interpretation might be found for interpolating models. Rosseel (2002) proposed one such model – the Mixture Model of Categorization (MMC) – in which it is assumed that  $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$  is a mixture distribution. Specifically, the model assumes that each object  $x_i$  comes from a cluster  $z_i$ , and each cluster is associated with a probability distribution over the features of the objects generated from that cluster. When evaluating the probability of a new object  $x_N$ , it is necessary to sum over all of the clusters from which that object might have been drawn, with

$$P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \sum_{k=1}^{K_j} P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})P(z_N = k|\mathbf{z}_{N-1}, c_N = j, \mathbf{c}_{N-1}) \quad (7)$$

where  $K_j$  is the total number of clusters for category  $j$ ,  $P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$  is the probability of  $x_N$  under cluster  $k$ , and  $P(z_N = k|\mathbf{z}_{N-1}, c_N = j, \mathbf{c}_{N-1})$  is the probability of

generating a new object from cluster  $k$  in category  $j$ . The clusters can either be shared between categories, or specific to a single category (in which case  $P(z_N = k|\mathbf{z}_{N-1}, c_N = j)$  is 0 for all clusters not belonging to category  $j$ ). It is straightforward to show that this reduces to kernel density estimation when each object has its own cluster and the clusters are equally weighted, and parametric density estimation when each category is represented by a single cluster. By a similar argument to that used for the exemplar model above, we can connect Equation 7 with the definition of  $\eta_{N,j}$  in the VAM (Equation 4), providing a rational justification for this method of interpolating between exemplars and prototypes.

## Anderson’s Rational Model of Categorization

The MMC elegantly resolves the question of how to define a rational model between exemplars and prototypes, but leaves open the issue of determining how many clusters are used in representing each category – a question about which of these kinds of representations might be more appropriate based on the available data. Anderson (1990) introduced a model that he called the Rational Model of Categorization (RMC), which presents a partial solution to this problem.

The RMC differs from the other models discussed in this section in assuming that category labels should be treated like features. Thus, the RMC specifies a joint distribution on features and category labels, rather than assuming that the distribution on category labels is estimated separately and then combined with a distribution on features for each category. As in the MMC, this distribution is a mixture, with

$$P(\mathbf{x}_N, \mathbf{c}_N) = \sum_{\mathbf{z}_N} P(\mathbf{x}_N, \mathbf{c}_N|\mathbf{z}_N)P(\mathbf{z}_N) \quad (8)$$

where  $P(\mathbf{z}_N)$  is a distribution over clusterings of the  $N$  objects. The key difference from the MMC is that this distribution allows the number of clusters to be unbounded, with

$$P(\mathbf{z}_N) = \frac{\alpha^K}{\prod_{i=0}^{N-1} [\alpha + i]} \prod_{k=1}^K (M_k - 1)! \quad (9)$$

where  $\alpha$  is a parameter of the distribution and  $M_k$  is the number of objects assigned to cluster  $k$ .<sup>1</sup> This is the distribution that results from sequentially assigning objects to clusters with probability

$$P(z_i = k|\mathbf{z}_{i-1}) = \begin{cases} \frac{M_k}{i-1+\alpha} & M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{i-1+\alpha} & M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \quad (10)$$

where the counts  $M_k$  are accumulated over  $\mathbf{z}_{i-1}$ . Thus, each object can be assigned to an existing cluster with probability proportional to the number of objects already assigned to that cluster, or to a new cluster with probability determined by  $\alpha$ .

<sup>1</sup>Due to space constraints, we have defined this distribution in the form associated with the Dirichlet process, rather than using the idea of a “coupling probability” from Anderson’s (1990) treatment (see Neal, 1998, and Sanborn, Griffiths, & Navarro, 2006, for details).

Despite having been defined in terms of the joint distribution of  $\mathbf{x}_N$  and  $\mathbf{c}_N$ , the assumption that features and category labels are independent given clusters makes it possible to write  $P(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$  in the same form as Equation 7. The probability of cluster  $k$  is simply

$$P(z_N = k | \mathbf{z}_{N-1}, c_N = j, \mathbf{c}_{N-1}) \propto P(c_N = j | z_N = k, \mathbf{z}_{N-1}, \mathbf{c}_{N-1}) P(z_N = k | \mathbf{z}_{N-1}) \quad (11)$$

where the second term on the right hand side is given by Equation 10. This defines a distribution over the same  $K$  clusters regardless of  $j$ , but the value of  $K$  depends on the number of clusters in  $\mathbf{z}_{N-1}$ . The RMC can thus be viewed as a form of the mixture model in which all clusters are shared between categories but the number of clusters is inferred from the data. However, the two models are not directly equivalent, because assuming that features and category labels are generated based on the clustering induces a dependency between the two, meaning that  $c_N$  depends on  $\mathbf{x}_{N-1}$  as well as  $\mathbf{c}_{N-1}$ , violating the (arguably sensible) assumption made by the other models and embodied in Equation 5.

The RMC thus comes close to our goal of specifying a unifying rational model of categorization, capturing many of the ideas embodied in other models and making it possible to infer a representation warranted by the data. However, the model is still significantly limited. First, the analysis given in the previous paragraph shows that the model assumes that every category is represented using the same set of clusters (and thus the same number), an assumption that is inconsistent with many models that interpolate between prototypes and exemplars (e.g., Vanpaemel et al., 2005). Second, the idea that category labels should be treated like other features has some odd implications, such as the dependency between features and category labels mentioned above. These limitations leave room for a model in which each category is directly represented by a different number of clusters, with the appropriate number being inferred from the data. We develop and test such a model in the remainder of the paper, by drawing on connections between the RMC and work in nonparametric Bayesian statistics.

## Dirichlet processes and beyond

The RMC defines a probability distribution as a mixture of an unbounded number of clusters. The same idea appears in nonparametric Bayesian statistics, in the form of the *Dirichlet process mixture model* (Antoniak, 1974; Neal, 1998). In fact, the distribution defined by the RMC is exactly the same as that defined by this model (Neal, 1998; Sanborn et al., 2006). This equivalence means that we can use recent results generalizing the Dirichlet process to identify a richer class of rational models of categorization.

Teh, Jordan, Blei, and Beal (2004) introduced a generalization of the Dirichlet process known as the *hierarchical Dirichlet process* (HDP). The basic idea is simple. Observations are divided into groups, and each group is modeled using a Dirichlet process. A new observation is first compared

	$\gamma \in (0, \infty)$ categories share clusters	$\gamma \rightarrow \infty$ categories share no clusters
$\alpha \rightarrow 0$ one cluster per category	HDP <sub>0,+</sub>	HDP <sub>0,<math>\infty</math></sub> (prototype)
$\alpha \in (0, \infty)$ intermediate number of clusters	HDP <sub>+,+</sub>	HDP <sub>+,<math>\infty</math></sub>
$\alpha \rightarrow \infty$ one stimulus per cluster	HDP <sub><math>\infty</math>,+}</sub> (RMC)	HDP <sub><math>\infty</math>,<math>\infty</math></sub> (exemplar)

Figure 1: Unifying rational models of categorization. Each model is specified as HDP <sub>$\alpha,\gamma$</sub> , where  $+$  is a value in  $(0, \infty)$ .

to all of the clusters in its group, with the prior probability of each cluster determined by Equation 10. If the observation is to be assigned to a new cluster, the new cluster is drawn from a second Dirichlet process that compares the stimulus to all of the clusters that have been created across groups. This Dirichlet process is governed by parameter  $\gamma$ , analogous to  $\alpha$ , and the prior probability of each cluster is proportional to the number of times that cluster has been selected by any group, instead of the number of observations in each cluster. The new observation is only assigned a completely new cluster if both Dirichlet processes select a new cluster.

The HDP provides a way to model probability distributions across groups of observations. Each distribution is a mixture of an unbounded number of clusters, but the clusters can be shared between groups. Furthermore, the number of clusters in each group can vary independently. A priori expectations about the number of clusters in a group and the extent to which clusters are shared between groups are determined by the parameters  $\alpha$  and  $\gamma$ . When  $\alpha$  is small, each group will have few clusters, but when  $\alpha$  is large, the number of clusters will be closer to the number of observations. When  $\gamma$  is small, groups are likely to share clusters, but when  $\gamma$  is large, the clusters in each group are likely to be unique.

## A unifying rational model

We can now define a unifying rational model of categorization, based on the HDP. If we identify each category with a “group” for which we want to estimate a distribution, the HDP instantly becomes a model of category learning, providing us with a way to formulate models in which the number of clusters in each category is learned, and subsuming all previous rational models through different settings of  $\alpha$  and  $\gamma$ . Figure 1 identifies six models we can obtain by considering limiting values of  $\alpha$  and  $\gamma$ .<sup>2</sup>

Three of the models shown in Figure 1 are exactly isomorphic to existing models. HDP <sub>$\infty,\infty$</sub>  is an exemplar model, with one cluster per object and no sharing of clusters. HDP<sub>0, $\infty$</sub>  is a prototype model, with one cluster per category and no sharing of clusters. HDP <sub>$\infty$ ,+}</sub> is the RMC, provided that category labels are treated as features. In HDP <sub>$\infty$ ,+}</sub>, every object has its own cluster, but those clusters are generated from the higher-level Dirichlet process. Consequently, group membership is

<sup>2</sup>The case of  $\gamma \rightarrow 0$  is omitted, since it simply corresponds to a model in which all observations belong to the same cluster across all categories, for all values of  $\alpha$ .

ignored and the model reduces to a Dirichlet process.

There are also several new models.  $HDP_{0,+}$  makes the same basic assumptions as the prototype model, with a single cluster per category, but makes it possible for different categories to share the same prototype – something that might be appropriate in an environment where the same category can have different labels. However, the most interesting models are  $HDP_{+,+}$  and  $HDP_{+,\infty}$ . These models are essentially the MMC, with clusters shared between categories or unique to different categories respectively, but the number of clusters in each category can differ and be learned from the data. Consequently, these models make it possible to answer the question of whether a particular category is best represented using prototypes, exemplars, or something in between, simply based on the structure of that category. In the remainder of the paper, we show that one of these models –  $HDP_{+,\infty}$  – can capture the shift that occurs from prototypes to a more exemplar-based representation in a recent categorization experiment.

### Modeling the prototype-exemplar transition

Smith and Minda (1998) argued that people seem to produce responses that are more consistent with a prototype model early in learning, later shifting to exemplar-based representations. The models discussed in the previous section potentially provide a rational explanation for this effect: the prior specified in Equation 9 prefers fewer clusters and is unlikely to be overwhelmed by small amounts of data to the contrary, but as the number of objects consistent with multiple clusters increases, the representation should shift. These results thus provide an opportunity to compare the HDP to human data.

We focused on the non-linearly separable structure explored in Experiment 2 of Smith and Minda (1998). In this experiment, 16 participants were presented with six-letter nonsense words labeled as belonging to different categories. Each letter could take one of two values, producing the binary feature representation shown in Table 1. Each category contains one prototypical stimulus (000000 or 111111), five stimuli with five features in common with the prototype, and one stimulus with only one feature in common with the prototype, which we will refer to as an “exception”. No linear function of the features can correctly classify every stimulus, meaning that a prototype model will not be able to distinguish between the categories exactly. Participants were presented with a random permutation of the 14 stimuli and asked to identify each as belonging to either Category A or Category B, receiving feedback after each stimulus. This block of 14 stimuli was repeated 40 times for each participant, and the responses were aggregated into 10 segments of 4 blocks each. The results are shown in Figure 2 (a). The exceptions were initially identified as belonging to the wrong category, with performance improving later in training.

We tested three models: the exemplar model  $HDP_{\infty,\infty}$ , the prototype model  $HDP_{0,\infty}$ , and  $HDP_{+,\infty}$ . In all three models, we assumed that the features of each object are independent

given its cluster, meaning that we can write

$$P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \prod_d P(x_{N,d}|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$$

where  $x_{N,d}$  is the value of the  $d$ th feature of object  $N$ . Given the cluster, the value on each dimension is assumed to have a Bernoulli distribution (although other distributions can be used for continuous features). Integrating out the parameter of this distribution with respect to a  $\text{Beta}(\mu_0, \mu_1)$  prior, we obtain

$$P(x_{N,d} = v|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \frac{M_{k,v} + \mu_v}{M_k + \mu_0 + \mu_1} \quad (12)$$

where  $M_{k,v}$  is the number of stimuli with value  $v$  on the  $d$ th feature that  $\mathbf{z}_N$  identifies as belonging to cluster  $k$ .

All three models were exposed to the same training stimuli as the human participants, and used to categorize each stimulus after each segment of 4 blocks. The cluster structures for the prototype and exemplar models are fixed, so the probability of each category is straightforward to compute. However, since  $HDP_{+,\infty}$  allows arbitrary clusterings, the possible clusterings need to be summed over when computing the probabilities used in categorization (as in Equation 7). We approximated this sum by sampling from the posterior distribution on clusterings using the Markov chain Monte Carlo (MCMC) algorithm described by Teh et al. (2004). Each set of predictions is based on an MCMC simulation with a burn-in of 1000 steps, followed by 100 samples separated by 10 steps each. The parameter  $\alpha$  was also estimated by sampling.

As in Smith and Minda’s original modeling of this data, a guessing parameter was incorporated to allow for the possibility that participants were randomly responding for some proportion of the stimuli. The guessing parameter was fit for each participant, being fixed across every instance of every stimulus for that participant. The values of  $\mu_0$  and  $\mu_1$  were also fit for each participant, with the restriction that  $\mu_0 = \mu_1$ , resulting in two free parameters for each of the models.

The predictions of the three models are shown in Figure 2, averaged across participants, and model fits appear in Figure 3. As might be expected, the prototype model does poorly in predicting the categories of the exceptions, while the exemplar model is more capable of handling these stimuli. We replicated the results of Smith and Minda (1998) in finding that the prototype model fit better early in training (for segments 1-4), and the exemplar model better later in training. However, we also found that  $HDP_{+,\infty}$  provided an equivalent or better account of human performance than the other two models from segment 4 onwards. In particular, only this model captured the shift in the treatment of the exceptions

Table 1: Categories A and B from Smith and Minda (1998)

	Stimuli
A	000000, 100000, 010000, 001000, 000010, 000001, 111101
B	111111, 011111, 101111, 110111, 111011, 111110, 000100

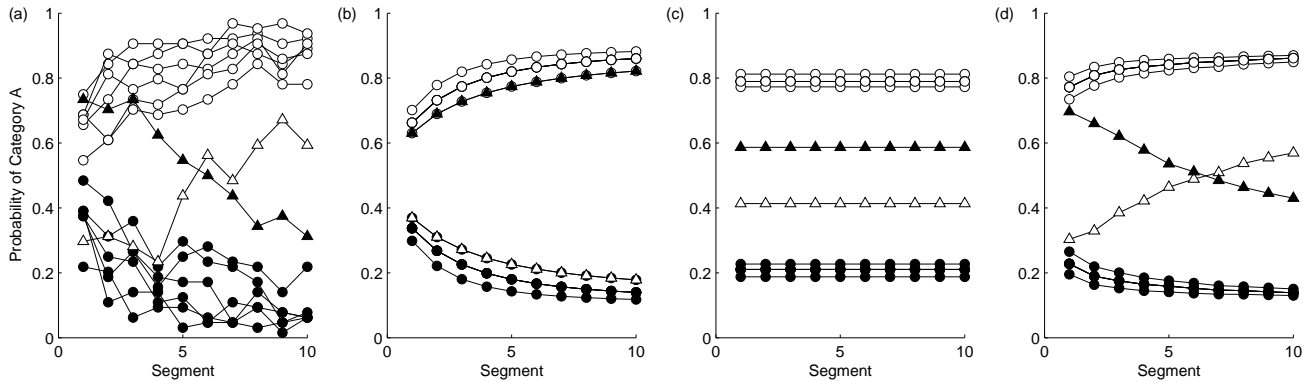


Figure 2: Human data and model predictions. (a) Results of Smith & Minda (1998, Experiment 2). (b) Prototype model,  $HDP_{\infty,0}$ . (c) Exemplar model,  $HDP_{\infty,\infty}$ . (d)  $HDP_{+, \infty}$ . For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure (111101 and 000100 respectively).

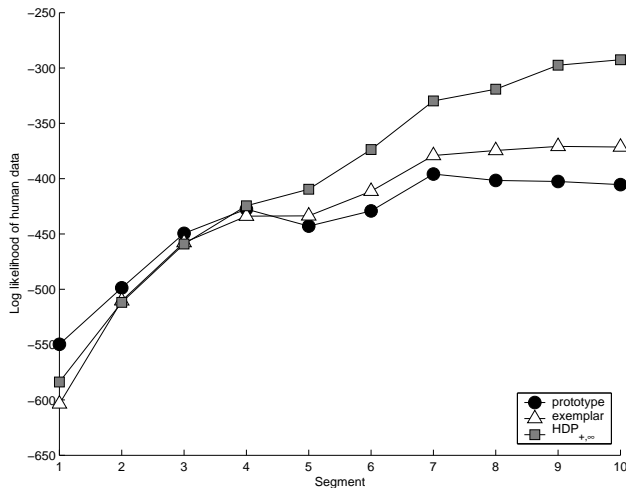


Figure 3: Model fits, measured as log-likelihood of the human responses under each model, for each segment of training.

over training. This shift occurred because the number of clusters in the HDP changes around segment 4: categories are initially represented with one cluster, but then become a more complex two cluster representation, with one for the stimuli close to the prototype and one for the exception.

## Conclusion

One of the most valuable aspects of rational models of cognition is their ability to establish connections across different fields. Here, we were able to exploit the correspondence between Anderson's (1990) Rational Model of Categorization and the Dirichlet process to draw on recent work in nonparametric Bayesian statistics that allowed us to define a more general rational model, based on the hierarchical Dirichlet process. This model subsumes previous rational analyses of human category learning, and provides a general solution to the problem of selecting the number of clusters to represent a category. The result is a picture of human categorization in which people do not use a fixed representation of categories

across all contexts, but instead select a representation whose complexity is warranted by the available data.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2, 1152-1174.
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216-233.
- Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias-variance tradeoff. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.
- Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, 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. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 218-247). Oxford: Oxford University Press.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 393-407.
- Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, 46, 178-210.
- 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.
- Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411-1436.
- Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press.
- Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.