

# Modeling category intuitiveness

Emmanuel M. Pothos ([e.m.pothos@swansea.ac.uk](mailto:e.m.pothos@swansea.ac.uk)), Amotz Perlman ([amotz@bgu.ac.il](mailto:amotz@bgu.ac.il)), Darren J. Edwards  
([225088@swansea.ac.uk](mailto:225088@swansea.ac.uk))

Department of Psychology, Swansea University, Swansea SA2 8PP, UK

Todd M. Gureckis ([todd.gureckis@nyu.edu](mailto:todd.gureckis@nyu.edu))

Department of Psychology, New York University, New York, NY 10003, USA

Peter M. Hines ([peter.hines@cs.york.ac.uk](mailto:peter.hines@cs.york.ac.uk))

Department of Computer Science, University of York, York YO10 5DD, UK

Nick Chater ([n.chater@ucl.ac.uk](mailto:n.chater@ucl.ac.uk))

Department of Psychology, UCL, London WC1E 6BT, UK

## Abstract

We asked 169 participants to spontaneously categorize nine sets of items. A category structure was assumed to be more intuitive if a large number of participants consistently produced the same classification. Our results provide a rich empirical framework for examining models of unsupervised categorization—and illustrate the corresponding profound modeling challenge. We provide a preliminary examination comparing two models of unsupervised categorization: SUSTAIN (Love, Medin, & Gureckis, 2004) and the simplicity model (Pothos & Chater, 2002).

**Keywords:** unsupervised categorization; simplicity; SUSTAIN.

## Introduction

In unsupervised categorization there is no pre-determined assignment of objects to categories. The participant is free to decide which classification makes more sense, typically with no or minimal constraints. Research into unsupervised categorization consists of several specific themes, for example, the circumstances under which unidimensional classification might be observed, the effect of stimulus format and procedure on classification strategies, and the role of general knowledge in category coherence (e.g., Milton & Wills, 2004; Yang & Lewandowsky, 2004). The focus of the present work is category intuitiveness, i.e. our ability to recognize certain groupings of objects as intuitive. To pick a trivial example, most people consider the grouping of all instances of chairs into one category as very intuitive. However, a category consisting of dolphins, babies born on Tuesdays, and the Eiffel Tower would be considered nonsensical. Ideally, we would be able to express these intuitions in mathematical terms.

Modeling category intuitiveness is an important objective for research into unsupervised categorization, but its study has been problematic because of the very large number of possible classifications for a set of items (for as few as 10 items there are 100,000 possible classifications). Category intuitiveness is closely related to Murphy and Medin's (1985) notion of category coherence. A possible difference is that category intuitiveness can be established on the basis

of purely perceptual considerations, whereas category coherence is theory-laden. The relation between the two is complex and we will not further address it presently.

Several sophisticated modeling frameworks have been proposed for modeling category intuitiveness and an exhaustive comparison would be impractical. We consider SUSTAIN (Love, Medin, & Gureckis, 2004) and the simplicity model (Pothos & Chater, 2002). There are several reasons why this comparison is interesting. First, while drawing from radically different formal specifications, both models suggest that a simplicity principal may guide category intuitiveness. In SUSTAIN, this is accomplished via an incremental coverage principle, whereby new knowledge structures are created when items are encountered which do not fit well into any existing structure. In the simplicity model, a categorization is favored to the extent that it provides a 'simplification' (in a formal, algorithmic sense) of the similarity structure of the presented items. Classifications that are highly complex or irregular for a set of items, would not allow much simplification and are discouraged. In addition, both models incorporate a similarity constraint: categories should be more intuitive and easier to learn to the degree that they respect the similarity structure of the items to be categorized.

Despite these similarities, there are important differences as well: SUSTAIN is parametric, in that it assumes that items have certain positions in psychological space. In contrast, the input to the simplicity model is the set of relative similarities, and so its operation is independent of exact information about item representation. Also, SUSTAIN's operation is guided by a number of free parameters while the simplicity model has typically no parameters; for a given input, it produces a prediction of what should be the most intuitive classification. Finally, SUSTAIN is a process model of trial-by-trial learning, whereas simplicity assumes all items are presented concurrently (although note that both models can be adapted to carry out their non-native form of categorization). In a way, the distinction between these two models reflects the

difference between theories developed at Marr’s algorithmic level (SUSTAIN) and at the computational/normative level (the simplicity model).

The goal of the present article is two-fold. First, we present results examining the types of category structures human participants prefer, when asked to spontaneously categorize sets of stimuli. Second, we examine SUSTAIN and simplicity as formal accounts of the empirical results.

## **Incremental Coverage vs. Information-Theoretic Simplicity: Comparing SUSTAIN and the Simplicity Model**

### **SUSTAIN**

SUSTAIN is a trial-by-trial clustering model of category acquisition, aiming to capture the full continuum between supervised and unsupervised categorization. Clusters in the model correspond to psychologically meaningful groupings of items. For example, when learning about categories of birds, a single cluster in the model might represent highly similar species such as robins and blue-jays separate from highly dissimilar examples such as ostriches. SUSTAIN is initially directed towards classifications involving as few clusters as possible, and only adds complexity as needed to explain the structure of a category. Two key aspects of SUSTAIN’s account are the role of *similarity* and *surprise* in directing category discovery. First, SUSTAIN favors clusters organized around perceptually or psychologically similar items. Second, new clusters are created in memory when the existing ones do a poor job of accommodating a new instance. Thus, SUSTAIN adjusts its category representations in a trial-by-trial fashion to accommodate the similarity structure of the items it has experienced.

When a to-be-categorized item is first presented to the model, it activates each existing cluster in memory, in a way based on the similarity of the item to each cluster. In addition, learned attention weights in the model can bias this activation in favor of dimensions which are more predictive for categorization. Clusters that are more activated are more likely to be selected as the “winner” for the item. If there are many highly activated clusters for a particular item, then confidence in the winning cluster is reduced—i.e., there is cluster competition (regulated by a parameter). In the unsupervised learning situations considered here, if the current input item fails to activate any existing cluster above some threshold level, then a new cluster is created for the item. This is the key mechanism of ‘surprise’ in SUSTAIN: new clusters are created in response to surprisingly novel stimuli that do not fit with existing knowledge structures. The threshold parameter ( $\tau$ ) controls what level of activation is considered ‘surprising’ enough, so that this parameter effectively determines the number of clusters the model creates ( $\tau$  is analogous to the coupling parameter in the rational model; Anderson, 1991).

Quantitative fits of SUSTAIN have shown that the model’s operation is not too dependent on exact parameter

values (Love et al., 2004). As a result, in the simulations reported here we recycle a single set of global parameters used from previous studies and only manipulate the setting of the threshold parameter ( $\tau$ ). Given that SUSTAIN is a trial-by-trial learning model, in modeling a free sorting task where multiple items are simultaneously presented, SUSTAIN’s fits are derived by running the model thousands of times on different stimulus orderings in order to create a distribution of plausible classifications: more psychologically intuitive classifications are considered to be the ones more frequently generated.

### **The Simplicity Model**

The simplicity model is effectively an implementation of Rosch and Mervis’s (1975) intuition about categorization, within an information-theoretic framework. Rosch and Mervis (1975) suggested that basic level categories maximize within- and minimize between-category similarity. In Pothos and Chater’s (2002) information-theoretic instantiation of this idea, classifications are considered descriptions of the similarity structure of a set of items. Where these descriptions afford an economical encoding of the similarity structure, they should be preferred. This is Occam’s razor (the simplicity principle), which has been argued to have psychological relevance (Chater, 1999; Feldman, 2000) and is congruent with Bayesian approaches in cognitive science (Tenenbaum et al., 2006).

More specifically, the simplicity model first computes the information content of all the similarity relations between a set of items. This is done by assuming that the similarity for each possible pair of items is compared to the similarity of every other pair. For example, is a banana and an apple more/less similar to a banana and an orange? Each such comparison is worth one bit of information (ignoring equalities). A classification for the items is defined as imposing constraints on the similarity relations: all similarities between objects in the same category are defined to be greater than all similarities between objects in different categories. Therefore, a classification can be evaluated in terms of how many correct constraints it provides—erroneous constraints need to be identified and corrected. Overall, taking into account the constraints imposed by a classification, the (information-theoretic) cost of correcting errors, and another cost term for specifying the classification, we compute the simplification provided by a particular classification. The prediction is that the greater this simplification, the lower the *codelength* of the similarity information of the items (when described with the classification) and the more psychologically intuitive the classification should be.

The above approach has proved adequate for small datasets (Pothos & Chater, 2002). For larger datasets, additional assumptions are required. First, some subclustering may occur. Following Rosch and Mervis (1975), we considered the initial prediction of the simplicity model as a basic level categorization. Categories in this

basic level categorization can be broken down into subordinates, by considering the items in each cluster as a new dataset and examining whether their classification affords additional simplification. This subclustering may corroborate or compete with the *final* (see later) classification. Second, for stimuli composed of more than one dimensions, classification may proceed on the basis of one dimension or both. The simplicity model has no parameters for attentional weighting, therefore dimensional selection has to take place automatically. Dimensional selection depends on whether the classification along either dimension (classification(x) or classification(y); for simplicity, call these x and y) is more intuitive than classification on the basis of both dimensions (call this xy; Pothos & Close, in press). In the present study, participants were asked to produce a two dimensional classification (an xy one). Therefore, the final classification (i.e., the basic level classification plus any further subclustering) has to be an xy one; ‘final classification’ will denote the classification the simplicity model predicts for a dataset. This can be achieved in two ways. Participants (or the model) may produce an xy classification straightaway. Or participants may first produce an x or a y classification, and then produce an xy one, by subclustering (we assume this is the only way in which subclustering can affect the form of the final classification).

In sum, in describing the results with the simplicity model, an assumption is that the final classification is xy, i.e., most participants will look for an xy classification. However, not all participants will produce the optimal xy classification. Why would they not do this? Because there might be competition from salient subclusters along either x or y. A subclustering is considered to compete with the final classification if it occurs either along x or y, or if it is in xy but it is susceptible to noise. Susceptibility to noise means that by introducing a little bit of noise in the similarity structure of the items the classification changes. This is rarely the case with a basic level categorization, i.e., such classifications are typically very stable against noise. However, when subclustering a cluster, introducing a little bit of noise (not more than 10% in psychological space positions), often leads to alternative classifications, if the items in the cluster are close to each other. When a (sub)clustering is susceptible to noise, we consider it as competing with the final classification.

In sum, subclustering and noise may lead to competition, which increases classification variability. A competition term is computed as the best codelength of the competing subclusters. Finally, competition may also arise if there is

more than one salient xy final categorization (this only happens in the 5202 dataset below).

Simplicity model predictions are specified as a percentage. We consider the codelength for the similarity structure of the items, without any clusters. Since we have 16 items in each dataset, this is 7140 bits (in each case). Note that this codelength does not take into account any regularity in the similarity structure of the items at all. We then consider the final codelength for a particular classification; this final codelength would take into account any non-competing subclustering. In other words, this is the codelength of the similarity structure of the items, when similarity information is encoded using categories. For example, in the case of the dataset labeled as 3585, the best possible classification is associated with a codelength of 3585 bits. Therefore, the simplicity prediction for this dataset would be expressed as  $3585/7140*100$  or 50.2%, to reflect the fact that only about 50% of the original codelength is required for describing the similarity structure of the items of this dataset *with* categories, compared to the situation where no categories are used. The lower this percentage, the greater the simplification afforded by the classification, and the more intuitive the corresponding classification is considered. The qualification to this conclusion relates to the ‘competition’ term: competition terms are computed in a way analogous to the above, and they correspond to how intuitive ‘competing’ subclusters are. Accordingly, the lower the competition term, the more intuitive competing subclusters are, and the less frequently the optimal classification should be produced.

## Experimental investigation

### Materials

We created nine datasets of 16 items each, reflecting a range of intuitions about unsupervised categorization (Figure 1). Each dataset is indexed by its codelength (with no subclustering or dimensional selection). Items were instantiated as spider-like images (Figure 2), so that length of ‘legs’ (after the joint) corresponded to the horizontal dimension in Figure 1 and length of body to the vertical dimension. By choosing such stimuli, both dimensions of physical variation were lengths, and so a Weber fraction in mapping the Figure 1 values to physical values could be safely assumed (8%%). In results not reported here, we collected similarity ratings and carried out multidimensional scaling to verify that our representational assumptions are valid (they are). Stimuli were printed individually and laminated.

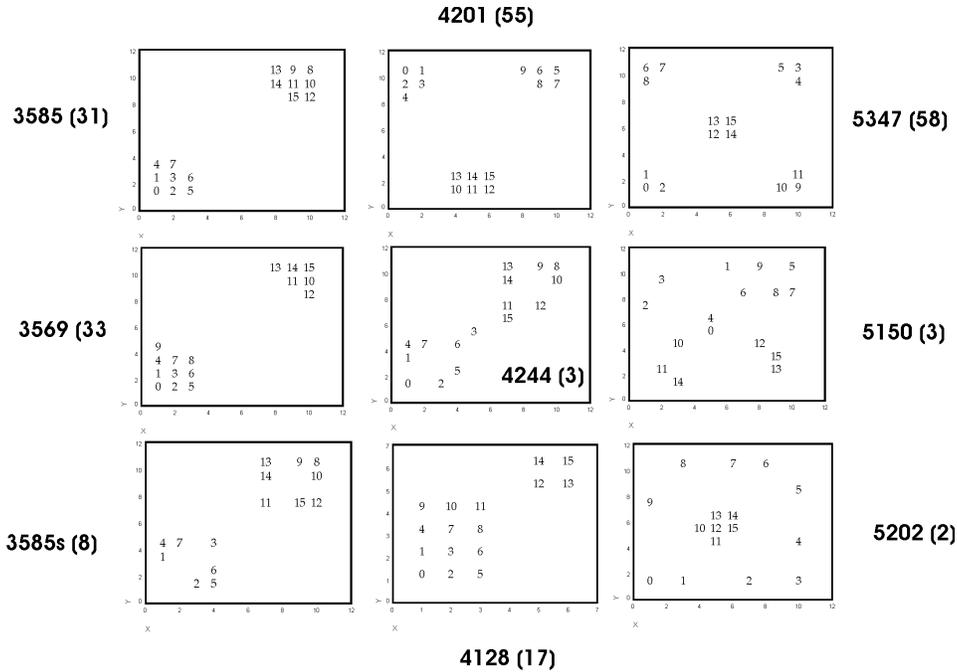


Figure 1: The nine datasets utilized in the present study. In parentheses is shown the number of times the most popular classification was produced.

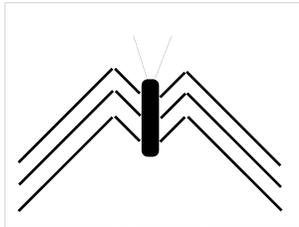


Figure 2: An example of the stimuli used.

### Participants and procedure

Participants were 169 students at Swansea University, who took part for a small payment. They received each set of items in a pile. They were asked to spread the items in front of them, and classify the items in a way that seemed natural and intuitive, using as many groups as they wanted, but not more than necessary. The two dimensions of variation were described and presented as equally important. There was an alternative set of instructions, where the stimuli were described as spiders in the Amazon; this was a ‘general knowledge’ manipulation, which, however, had no effect. Participants indicated their classification by arranging the stimuli into piles. Each participant went through all nine datasets, in a random order.

### Results

In an experiment of this sort there is clearly a wealth of data, more than 1500 participant responses. We focus on what we consider one of the key issues in unsupervised categorization, namely predicting category intuitiveness.

Note that the majority of clustering/ categorization models would be successful in predicting the most popular solution in each dataset. However, the key empirical issue and modeling challenge is to appreciate why in some datasets there were as few as 84 distinct classifications, while in others 160.

We suggest that category intuitiveness can be operationalized as classification variability, i.e., the number of distinct classifications produced for a set of items. If there are fewer distinct classifications, then the psychological intuitions about certain classifications being more obvious should be stronger. Alternatively, we can count the frequency of the most popular classification for a dataset. If the most popular classification has a high frequency, then it should be the case that this classification is considered more obvious. In our dataset, the two measures are *equivalent* (correlation: 0.99), therefore, we shall consider only the latter (of course, in general this may not be the case). Table 1 shows the empirical results and illustrates the complexity of research into unsupervised categorization. Even with datasets which were created to represent a simple two-category structure, there are over 100 unique classifications.

Table 1: ‘Fr of most popular’ refers to the number of participants who produced the most popular classification, ‘Distinct’ to the number of distinct classifications. Codelength/ competition values refer to the simplicity model fit.

Dataset	Fr. most popular/ distinct	Codelength (%)/ competition (%)
---------	-------------------------------	------------------------------------

3585	31/ 124	50.2/ 66
3569	33/ 116	50/ 62.4
3585s	8/ 152	50.2/ 52.9
4128	17/ 141	57.8/ 61
4201	55/ 104	43.5/ 68.2
4244	3/ 160	59.4/ 52.9
5150	3/ 159	72/ 60.2
5202	2/ 164	60.6/ 69.7
5347	58/ 84	57/ 100

### SUSTAIN Results

Following previous simulations of unsupervised sorting tasks with SUSTAIN (Gureckis & Love, 2002; Love, Medin, & Gureckis, 2004), the model was applied to the sorting task in a trial-by-trial fashion. In order to approximate the free-sorting task with SUSTAIN, we make the reasonable assumption that subjects consider each stimulus one at a time but that the order of item consideration is idiosyncratic. Accordingly, SUSTAIN was given 5 blocks of training, each block consisting of a different random ordering of all stimuli. Stimuli were represented to the model as coordinate pairs. Input values along each dimension were scaled between 0.0 and 1.0. Attention for both dimensions was set to an initial value of  $\lambda = 1.0$ . but during the learning phase SUSTAIN could adjust this value. Since subjects were encouraged to use both dimensions while sorting we assumed attention was equally allocated for both  $x$  and  $y$  (akin to the  $xy$  bias in the simplicity model). After the learning phase, we examined the structure of SUSTAIN's clusters by probing which items the model assigned to the same clusters in memory (i.e., items that strongly activated the same cluster were considered to be psychologically grouped).

Figure 3 show the results of SUSTAIN's basic predictions (scaled by multiplying probability of classification by 196). In order to account for the variability of responses by subjects, in the simulations reported here we assumed that  $\tau$  parameter varied from person to person following a roughly normal distribution (mean and SD were treated as free parameters for each dataset). Remember that the  $\tau$  parameter determines how dissimilar an item has to be from an existing cluster in order to warrant creating a new cluster in memory. All other free parameters in the model were recycled from a single global set of parameters used in previous studies. SUSTAIN provides a good account of the results (Figure 3). For example, SUSTAIN (like simplicity) correctly predicts that dataset 4201 and 5347 should have the most agreement while also predicting little consistency in responding for problems 4244, 5150, and 5202.

### Simplicity Model Results

We illustrate the simplicity model fit with the 3585 dataset and highlight aspects of its account for the other datasets. Observing Figure 1 for item id numbers, the basic level categorization in  $xy$  is (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15), with a codelength of 50.2%; the  $x$ ,  $y$

basic level categorizations are the same, so we select as  $xy$  the basic level categorization (since it is assumed that participants are biased to produce  $xy$  classifications). Furthermore, since no further subclustering is possible in  $xy$  (subclusters have very poor codelength, 92.5%), the final, predicted, classification for this dataset is (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15). With respect to competition for this solution, there are very good subclusters along  $x$  or  $y$ , each one of which is associated with a codelength of 66%; so, the competition term for this dataset is 66%. (recall, the competition value is the lowest codelength corresponding to any subclustering; these subclusters are competing, rather than part of the predicted classification, because we assume that the predicted classification is  $xy$ ). So, even though in this case we have a very intuitive final classification (codelength of only 50.2%), there is a rather high competition value as well, suggesting there would be some noise.

The simplicity model can account for the superiority of the 4201 and 5347 datasets because in both cases the basic level categorization is initially 1D. Therefore,  $xy$  subclustering provides additional simplification, rather than competition (recall, the final classification has to be  $xy$ ). For the 5202 dataset the basic level categorization is also one dimensional and  $xy$  subclustering provides additional simplification; however, there is also an alternative, competing final  $xy$  classification with comparable codelength. Table 1 provides a list of codelengths and competition values for the datasets. A linear regression analysis with codelength and competition as the independent variables, and frequency of the most popular classification as the dependent variable, was significant ( $F(2, 6) = 14.5$ ,  $p = .005$ ,  $R^2 = .83$ ); however, the correlation is not perfect, indicating that there is room for improvement in the simplicity approach. Also, the balance between competition and gain was governed by two (regression) parameters, but in the future it would be desirable to specify this balance automatically within the simplicity framework.

### Discussion

With 16 items there are well over 100,000 potential classifications. The immense size of this space, along with the fact that few constraints given to participants in our spontaneous classification task, suggests that idiosyncratic variation (assumptions about the stimuli, processing biases etc.) have plausibly played a significant role in determining the distribution of classifications. However, despite this variability, there were datasets for which more than 30% of participants agreed on which classification is the best and datasets for which no more than *three* participants agreed on an optimal

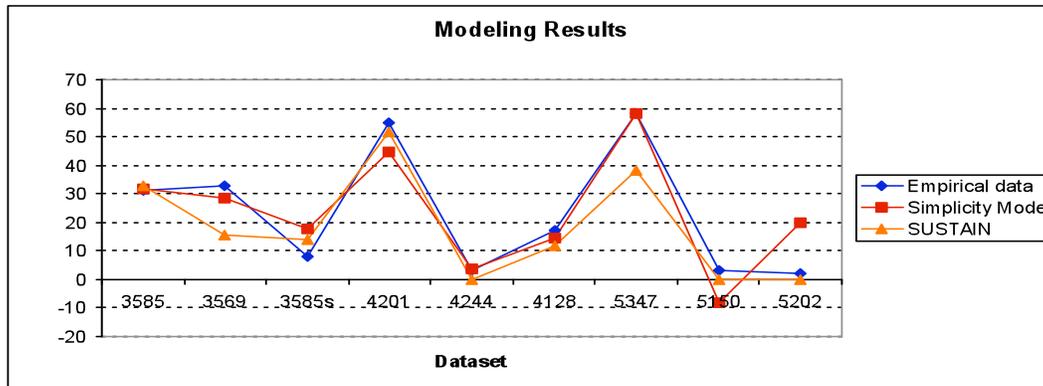


Figure 3: The y-axis show the frequency of the most popular classification in different datasets. The frequencies correspond either to the observed participant results, or to the predictions of the computational models.

classification. We consider this variability extremely interesting and an exciting, novel, and important challenge for models of unsupervised categorization.

In our preliminary analysis, we examined two models of unsupervised category construction that draw from somewhat different formalisms. While both SUSTAIN and the simplicity models have broad empirical support, neither has been tested against such an extensive range of unsupervised categorization data. The fact that both models provide a reasonable account of the classification behavior of human subjects in our task is encouraging and argues favorably for the relevance of simplicity and similarity as appropriate constraints in unsupervised categorization. Moreover, both models appear to have more difficulty over the same range of datasets (5347, 5150, 5202). It is possible that SUSTAIN and the simplicity model reflect different ways of computationally implementing (at the algorithmic and computational level respectively) simplicity/d similarity in unsupervised categorization? Much work remains before this potentially important conclusion can be confirmed.

With respect to the simplicity model, the roles of subclustering and stability against noise need to be better integrated with the main foundation of the model. Likewise, in terms of SUSTAIN, more work is needed to understand the full distribution of preferred groupings. For example, in some cases, SUSTAIN correctly predicted the relative prevalence of the most popular solution generated by human participants, however in a few occasions the model showed a bias towards alternative groupings that participants did often not select.

In general, as has been the case in supervised categorization (e.g., Nosofsky, 2000), we hope that comparative studies like the present one will help guide the development of computational models in unsupervised categorization as well.

### Acknowledgments

This research was supported by ESRC grant R000222655 to EMP and NIMH training grant T32 MH019879-12 to TMG. We would like to thank Brad Love for his help.

### References

- Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. *Psychological Review*, 98, 409-429.
- Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *Quarterly Journal of Experimental Psychology*, 52A, 273-302.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407, 630-633.
- Gureckis, T.M., Love, B.C. (2002). Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions. In *Proceedings of the 24<sup>th</sup> Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum: Hillsdale, NJ.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.
- Milton, F. & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 407-415.
- Murphy, G. L. & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92, 289-316.
- Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1735-1743.
- Pothos, E. M. & Chater, N. (2002). A Simplicity Principle in Unsupervised Human Categorization. *Cognitive Science*, 26, 303-343.
- Rosch, E. & Mervis, B. C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, 573-605.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10, 309-318.
- Yang, L. & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1045-1064.