Categorisation, Bounded Rationality and Bayesian Inference

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia 5005, Australia

Abstract

A Bayesian interpretation of the Generalised Context Model (GCM) is introduced. In this model, it is assumed that participants sample a set of exemplars from memory, and make classification decisions using similarity-based inference. It is shown that these assumptions give rise to a space of hypotheses about the category, with a prior distribution influenced by constraints on time and processing, as well as assumptions about the structure of the environment. Under this view, the response scaling parameter in the GCM plays a central role, since it defines the basic structure of the hypothesis space available in a categorisation task.

The notion of bounded rationality was introduced by Simon (1956), who argued that constraints on time and processing resources mean that it is inappropriate for an organism to integrate all of the information relevant to some decision. Rather, a limited search process is a more rational approach. This idea has been advocated more recently by Gigerenzer and Todd (1999), who propose "fast and frugal" heuristic models of human decision making. These heuristics should address three questions:

- How should a stimulus environment be searched for information?
- When should this search for information be terminated?
- Once the search has been terminated, what decision should be made given the available information?

In contrast to this view, rational Bayesian models emphasise the statistical inferences warranted by the information available to people. Rational Bayesian accounts have been successfully applied to a range of psychological phenomena such as memory (e.g., Anderson & Schooler 1991, Shiffrin & Steyvers 1997), categorisation (e.g., Anderson, 1990, 1991), generalisation (e.g., Shepard 1987, Tenenbaum & Griffiths 2001) and causal learning (e.g., Steyvers, Tenenbaum, Wagenmakers & Blum 2003), as well as problems such as inferring semantic representations (Griffiths & Steyvers 2002) and similarity models (Navarro & Lee 2003). Additionally, the Bayesian framework arguably remains the most coherent and powerful approach to statistical inference and model selection (e.g., Jaynes 2003). In this paper I argue that "bounded rationality" fits very elegantly into a Bayesian framework, and can be used to tie together a variety of ideas about human categorisation.

Exemplars and categorisation

I begin by outlining a classic process model of categorisation, the Generalised Context Model (GCM; Nosofsky, 1984, 1986). As it is standardly interpreted, the GCM is a "computational process model", in that it describes an algorithmic method of making categorisation decisions. It is not framed in Bayesian terms, nor is the process explicitly tied to some notion of fast and frugal decision making. However, the model has a long empirical and theoretical tradition, and is widely regarded as one of the most successful accounts of categorisation phenomena.

Suppose that one's prior experience of some category consists of n previously observed instances. According to the exemplar theory of classification, a category is represented by storing all n instances (Medin & Schaffer 1978, Nosofsky 1984, 1986). In the Generalised Context Model (GCM; Nosofsky, 1986), the probability $p(C \mid i)$ with which people decide that the ith stimulus belongs to the category C is proportional to the sum of the similarities s_{ij} to the individual exemplars. It is convenient to refer to this sum as the GCM similarity between the ith stimulus and the category, denoted $s_{iC}^{(\mathrm{GCM})}$,

$$s_{iC}^{(\mathrm{GCM})} = \sum_{j \in C} s_{ij}. \qquad (1)$$

Thus the model predicts that $p(C \mid i) \propto s_{iC}^{(\mathrm{GCM})}$. In calculating the similarity between two stimuli, it is usually assumed that stimuli are represented

as points in an m-dimensional Minkowski space), though this is not essential to the theory (e.g., Tversky 1977). Following Shepard (1987), similarity is assumed to decrease exponentially with distance. Correspondingly the similarity s_{ij} between the ith and jth stimuli is given by $s_{ij} = \exp(-\lambda d_{ij})$ where the attention-weighted distance d_{ij} is

$$d_{ij} = \left(\sum_{k=1}^{m} \left(w_k \left| x_{ik} - x_{jk} \right| \right)^r \right)^{1/r} \qquad (2)$$

and x_{ik} is the coordinate value of the ith stimulus on dimension k. In this equation, w_k denotes the proportion of attention applied to the kth dimension, r determines the metric, and λ denotes the steepness of the exponential decay, called the generalisation gradient.

A more recent version of the model (Ashby & Maddox 1993), GCM-γ, assumes that the summed similarities are raised to some power γ, adding an extra parameter and altering the probability of category membership. Thus,

$$s_{iC}^{(\text{GCM-}\gamma)} = \left(\sum_{j \in C} s_{ij} \right)^{\gamma} \qquad (3)$$

and the model now predicts that $p(C \,|\, i) \propto s_{iC}^{(\text{GCM-}\gamma)}$.

Bayesian Framework

A Bayesian framework for categorisation is simple enough to outline (see Anderson 1990, 1991). If people can entertain a number of distinct hypotheses $h \in \mathcal{H}$ about a category, each of which describes a distribution over stimuli in the environment, then the overall category density is found by marginalising over these hypotheses,

$$p(i \,|\, C) = \sum_{h} p(i \,|\, h) p(h).$$

Using Bayes theorem, the posterior probability of choosing category C is simply,

$$p(C \,|\, i) = \frac{p(i \,|\, C) p(C)}{p(i)}.$$

The focus in this paper is on the category density $p(i \,|\, C)$, and in understanding how a particular category becomes associated with a particular distribution over stimuli. In what follows, a Bayesian formulation of GCM-γ is presented, based on the fast and frugal approach of Gigerenzer and Todd (1999).

Similarity-Based Decisions

In a categorisation context, the set of previously encountered exemplars (assumed to be stored in memory) constitutes an environment that provides information about the category. An exemplar-based fast and frugal approach to categorisation would engage in some limited search through this environment, accumulating evidence required to make a decision. For the moment, assume that information about the category is retrieved stochastically from memory by sampling a set of γ exemplars with replacement. The outcome of such a sampling process can be expressed by the vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, where ϕ_j indicates the number of times that the jth exemplar is sampled, and where $\sum_j \phi_j = \gamma$. I refer to the outcome of this sampling process as the $\boldsymbol{\phi}$-mixture.

In this section I consider how decisions might on the basis of having sampled the $\boldsymbol{\phi}$–mixture. We can view this set of exemplars as having produced a hypothesis $h_{\boldsymbol{\phi}}$ about the category. Following the principles of Bayesian inference outlined earlier, we can treat this hypothesis as a distribution $p(i \,|\, h_{\boldsymbol{\phi}})$ over potential stimuli. Notice that this implies a probabilistic response to stimuli, reflecting some uncertainty about the environment. The issue is how to define this distribution appropriately.

It has frequently been argued that categorisation substantially relies on similarity (see Komatsu 1992). The idea was perhaps best expressed by Shepard (1994), who argued that stimulus representations are shaped by evolution or prior learning to reflect the structure in the environment. So our perceptions of similarity are viewed as reflecting the probability with which stimuli are expected to share consequences (e.g., Shepard 1987, Tenenbaum & Griffiths 2001). As a consequence, to the extent that categories reflect the same structures, we should expect similarity and categorisation to be closely related. While there are a number of respects in which categorisation and similarity are thought to differ (e.g., Rips 1989), similarity-based models of categorisation such as the GCM have generally been empirically successful over the years.

Given a set of sampled exemplars indicated by $\boldsymbol{\phi}$, the question that arises is how they should relate to $p(i \,|\, h_{\boldsymbol{\phi}})$. In the original GCM, similarities are assumed to combine additively. In contrast, prototype models such as those advocated by Smith and Minda (1998) tend to combine similarities in a multiplicative manner. In what follows, I will adopt a multiplicative similarity rule, which implies that,

$$p(i \,|\, h_{\boldsymbol{\phi}}) \;\propto\; \prod_{j} s_{ij}^{\phi_j}. \qquad (4)$$

The multiplicative similarity rule has previously been discussed by Nosofsky (1984). If stimuli are spatially represented, for instance, then multiplying similarities is equivalent to summing distances. Thus the comparison between the observed stimulus and the sampled exemplars takes place at a *representational* level, which agrees with the notion that the set of exemplars forms a single hypothesis. In contrast, summing similarities would imply that each member of the set is compared to the observed stimulus in isolation and *then* combined, suggesting a decision process rather than a representational one. Note that the effect of this assumption is to require that all the sampled exemplars have some similarity to the novel stimulus.

The normalising constant for this distribution is found by integrating the similarity measure over all the space of all possible stimuli, so the density function is given by,

$$p(i \,|\, h_{\boldsymbol{\phi}}) \;=\; \frac{\prod_j s_{ij}^{\phi_j}}{\int \prod_j s_{xj}^{\phi_j} \; dx},$$

where the integration over x is taken over all possible stimuli. For instance, if stimuli are represented spatially, then the integration is taken over the entire space. Obviously, this integral would be difficult to calculate in many cases, but as we will see, it drops out of the final expression for $p(i \,|\, C)$.

Searching Memory Quickly

From a bounded rationality perspective, it is assumed that human inference is based on a *limited* search of the environment. If the memory environment consists of the set of n previously observed exemplars, one might sample a set of γ exemplars independently with replacement from memory. The advantages to this process are speed and simplicity. Only a limited number γ of samples are required, and the samples are entirely independent of one another, and so could presumably be implemented in a straightforward manner. This sampling scheme implies that the probability of sampling the *hypothesis* $h_{\boldsymbol{\phi}}$ is equivalent to the probability of sampling the corresponding multiset $\boldsymbol{\phi}$ from a set of arbitrary tokens. This induces the following prior:

$$p(h_{\boldsymbol{\phi}}) \propto \binom{\gamma}{\phi_1, \ldots, \phi_n},$$

where $\binom{\gamma}{\phi_1, \ldots, \phi_n} = \frac{\gamma!}{\phi_1! \ldots \phi_n!}$. Notice that this choice answers the first two questions posed by Gigerenzer and Todd (1999): search is random, and terminates after γ samples have been drawn.

Although such a search is likely to be both fast and frugal, it does not exploit the structure of the environment in a manner consistent with Simon's (1956) view of bounded rationality. In particular, it does not capture the notion that the sampled exemplars are supposed to represent a *hypothesis about a category*. Rather, exemplars are treated as arbitrary tokens sampled independently with replacement. In practice, this seems implausible. For instance, a set of instances that includes three oranges and two lemons seems to be *a priori* more likely to form a useful hypothesis about a natural category than three oranges and two automobiles. The former seems to capture some compact structure, while the latter seems like a conjunction of two distinct hypotheses. In other words, the samples are not likely to be conditionally independent of one another. A set of interrelated exemplars forms a good hypothesis, while unrelated exemplars do not. This is a form of prior knowledge about the environment, which should have some impact on the prior beliefs expressed through $p(h_{\boldsymbol{\phi}})$.

In order to address this, consider the following argument. The nature of the sampling process places an inherent prior over a collection of arbitrary tokens. However, since these tokens have representational content, we have some additional evidence about the prior likelihood that those tokens refer to a category. If categorisation decisions are based on similarity, and since a set of very disparate exemplars cannot *all* be highly similar to anything, it is reasonable to assume that they are less likely to consitute a natural category. In short, the prior probability assigned to a set of exemplars ought to be related to their ability to be jointly similar to future stimuli. If we denote this additional evidence as $p(h \in \mathcal{C} \,|\, h_{\boldsymbol{\phi}})$, then this argument suggests that

$$p(h \in \mathcal{C} \,|\, h_{\boldsymbol{\phi}}) \propto \int \prod_j s_{xj}^{\phi_j} \; dx.$$

In this expression, $h \in \mathcal{C}$ denotes the observation that h refers to a category of some kind. An appropriate prior should incorporate this evidence through Bayes' rule, from which we obtain the prior,

$$\begin{aligned} p(h_{\boldsymbol{\phi}} \,|\, h \in \mathcal{C}) \;&\propto\; p(h \in \mathcal{C} \,|\, h_{\boldsymbol{\phi}}) p(h_{\boldsymbol{\phi}}) \\ &\propto\; \binom{\gamma}{\phi_1, \ldots, \phi_n} \int \prod_j s_{xj}^{\phi_j} \; dx. \quad (5) \end{aligned}$$

However, for the remainder of the paper, I will drop the "$h \in \mathcal{C}$" term, and refer to the quantity in Eq. 5 as $p(h_{\boldsymbol{\phi}})$.

At this point, it is worth reflecting on the implications of this prior. The prior incorporates two sources of evidence, namely the prior probability of sampling a collection of tokens, and the probability that the representational content of those tokens constitutes a viable hypothesis about a category. The "token sampling" component of the prior arises because people have limited resources (e.g., time, computation), and so can only afford to engage in a limited search of memory. On the other hand, the "representational" component arises because some collections of stimuli are less likely to form a natural category in the environment. So the prior distribution that I have discussed reflects constraints on time and processing, as well as (beliefs about) pre-existing structures in the environment. It is in this respect that the prior distribution is viewed as *boundedly rational*.

That said, there are some important caveats that attach to the prior. Firstly, it is deliberately chosen so as to ensure that the Bayesian model is equivalent to GCM-γ (see below). No strong claims are made as to whether this is "really" the right prior, merely that it is possible to induce a prior by thinking about contraints on time and processing, as well as the structure of the environment. Secondly, a strong (rather than merely pragmatic) committment to the notion of bounded rationality requires a fast heuristic that samples hypotheses with probability $p(h_{\boldsymbol{\phi}})$. To do this, we would need a process in which there exists some *dependency* between samples from memory, in which similar exemplars tend to be retrieved together. This seems quite plausible, given the literature on semantic priming (e.g., Meyer & Schvaneveldt, 1971). It would almost certainly be possible to devise such a sampling scheme, and it would be interesting to see how closely this process agrees with existing views of memory. However, such a discussion is beyond the scope of this paper.

Bounded Rationality of GCM-γ

The probability that category C is the appropriate one having observed stimulus i can be found using Bayes rule, so $p(C \,|\, i) \propto p(i \,|\, C)p(C)$. If all categories are equally likely a priori, then the $p(C)$ term is a constant, and Bayes' rule implies that $p(C \,|\, i) \propto p(i \,|\, C)$. Of course, this need not be true in general, so the prior probability of the category would need to be determined. Nevertheless, since the focus of this paper is on the category density $p(i \,|\, C)$, this complication will not be discussed further.

In any particular categorisation experiment, the sampled hypotheses are unobservable. All that we can observe are the categorisation decisions. So we can use the results from previous sections to express the (marginal) probability of observing stimulus i given the category C, as

$$\begin{aligned} p(C \,|\, i) &\propto p(i \,|\, C) \\ &= \sum_{\boldsymbol{\phi}} p(i \,|\, h_{\boldsymbol{\phi}}) p(h_{\boldsymbol{\phi}}) \\ &= \sum_{\boldsymbol{\phi}} \binom{\gamma}{\phi_1, \ldots, \phi_n} \prod_j s_{ij}^{\phi_j} \qquad (6) \\ &= \left(\sum_j s_{ij} \right)^{\gamma} = s_{iC}^{(\text{GCM-}\gamma)}. \end{aligned}$$

where the last step corresponds to the factorisation of a polynomial. In other words, the Bayesian model just derived is equivalent to GCM-γ.

Notice that there is a natural structure that arises because similarities are assumed to be exchangeable. Suppose that $n = \gamma = 4$. We then have four hypotheses that belong to the class that we usually refer to as "exemplars", namely $\boldsymbol{\phi} = (4, 0, 0, 0)$, $\boldsymbol{\phi} = (0, 4, 0, 0)$, $\boldsymbol{\phi} = (0, 0, 4, 0)$, and $\boldsymbol{\phi} = (0, 0, 0, 4)$. Equivalently, there is a single term that belongs a "quasi-prototype" class, namely $\boldsymbol{\phi} = (1, 1, 1, 1)$. Along the same lines, it is easy to see that $\boldsymbol{\phi} = (2, 1, 1, 0)$ belongs to the same class as $\boldsymbol{\phi} = (1, 1, 0, 2)$, but that $\boldsymbol{\phi} = (3, 1, 0, 0)$ belongs to a different class altogether. Psychologically, a hypothesis class includes all possible product terms that mix similarities in the same proportions. The exemplar class is the trivial mixture of only a single similarity, while the quasi-prototype class mixes all similarities in equal proportions. Thus, altering the value of γ leads to changes in the hypothesis classes that are considered by GCM-γ.

Taking all this together, we can view the GCM-γ category density as a weighted average of a number of distinct hypotheses about the underlying category, where the hypotheses can be naturally partitioned in terms of the mixture classes to which they belong. This is shown in Figure 1, in which there are four exemplars located in one dimensional space, with co-ordinates of 0, 0.1, 0.3, and 1. Setting $\gamma = 4$, the GCM-γ model predicts a kind of "circus tent" category distribution, shown at the back of the plot. While this category density does not look very principled, the Bayesian view allows us to decompose it into an average across a range of different hypotheses. To that end, Figure 1 also plots each of the individual hypothesis distributions, arranged by the mixture class to which they belong. Additionally, the size of each distribution is scaled in order to indicate how much each hypothesis contributes to the overall marginal distribution.

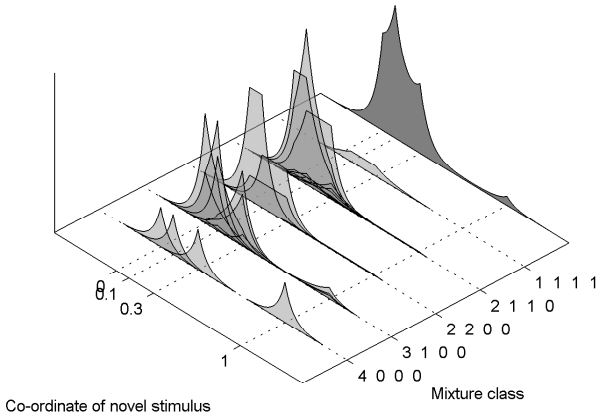

Figure 1: The decomposition of GCM-γ for a one-dimensional category with four exemplars and a smoothing parameter of $\gamma = 4$. Each of the terms in the expansion produces a unique generalisation gradient, and belongs to one of five natural classes.

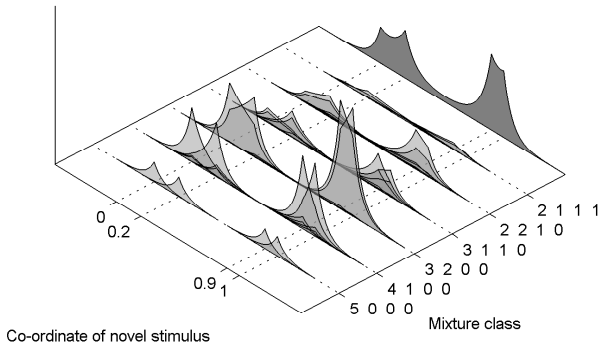

Figure 2: The decomposition of GCM-γ for a one-dimensional category with four exemplars and a smoothing parameter of $\gamma = 5$. Each of the terms in the expansion produces a unique generalisation gradient, and belongs to one of six natural classes.

If we restrict the prior distribution $p(h_{\boldsymbol{\phi}})$ to one based on the "token sampling" view, in which γ exemplars are sampled randomly, independently, and with replacement, then it is trivial to see that some hypothesis classes are more likely than others. For example, it is much easier to sample a $(2, 2, 0, 0)$ mixture than a $(4, 0, 0, 0)$ when $p(h_{\boldsymbol{\phi}}) \propto \binom{\gamma}{\phi_1, \ldots, \phi_n}$. However, in this sampling scheme all hypotheses *within* a class are equally likely. But when the prior incorporates the representational constraints discussed earlier, there is considerable variation within a hypothesis class. Hypotheses that combine distant (and hence dissimilar) stimuli receive very low weight, such as the $(2, 2, 0, 0)$ mixture that combines the stimulus at 0 with the stimulus at 1 in Figure 1. In fact, the weight assigned to that hypothesis is so small that it is almost invisible.

Looking at the Figure as a whole provides a nice theoretical interpretation of the "circus tent" distribution. There are a lot of hypotheses in the region between 0 and 0.3 that are assigned a high likelihood. So when $\gamma = 1$, the hypothesis space contains only the exemplars. When this is raised to $\gamma = 4$, the hypothesis space is expanded in such a way as to assign more weight to the regions between adjacent exemplars.

A similar effect is seen in Figure 2, in which there are still four exemplars, but γ is raised to 5. However, the exemplars are now located at 0, 0.2, 0.9 and 1. Because these now fall into two natural clusters, the hypotheses with high prior weight are those that mix the two small-valued stimuli or the two large-valued stimuli, producing a "twin peaked" distribution. Also, notice that since $\gamma \neq n$ in this case, there is no "quasi-prototype" in which all exemplars mix equally.

Discussion

It is interesting to observe that in order to translate GCM-γ into a more overtly probabilistic framework, the prior distribution $p(h_{\boldsymbol{\phi}})$ had to be chosen in a way that incorporated both a sampling process for $\boldsymbol{\phi}$ and rational considerations about the structure of natural categories. This may not be entirely coincidental, if one accepts the view that human cognition is conditionally rational given constraints on time and information processing (e.g., Anderson 1990, Gigerenzer & Todd 1999). In order to make timely categorisation decisions, people may only spend a limited amount of time drawing information from memory, as captured by the γ parameter in exemplar models.

That said, there is a sense in which this work is preliminary, and the approach could be expanded. In particular, the Bayesian inference introduced here integrates over a set of hypotheses that people might entertain, but the sampling process that underlies this hypothesis space suggests that only one hypothesis is considered at any given time. As a re-

sult, if people test hypotheses sequentially (as suggested by Nosofsky, Palmeri & McKinley's (1994) RULEX model, for instance), then the Bayesian interpretation of GCM-γ would predict sequential dependencies in categorisation performance. Furthermore, there is an inherent confound in GCM-γ, in that γ controls the diversity of the hypothesis space and acts as an evidence parameter that governs the way hypotheses are sampled from that space. It is not clear whether these two roles are distinct in human categorisation. If they are, then this may suggest possible refinements to exemplar models. Finally, since this paper has started from GCM-γ and worked towards a rational Bayesian interpretation, it would be interesting to consider the other direction. The GCM is implicitly built on Shepard's (1987) rational analysis, so it might be that one could start with the rational considerations and derive a categorisation model that behaves quite similarly to GCM-γ.

Acknowledgments

This work was supported by Australian Research Council grant DP-0451793. I thank Nancy Briggs, Simon Dennis, Michael Lee and Mark Pitt.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorisation. *Psychological Review, 98*, 409429.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

Ashby, F. G. & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorisation. *Journal of Mathematical Psychology, 37*, 372-400.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Griffiths, T. L. & Steyvers, M. (2002). Probabilistic models of semantic representation. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112*, 500-526.

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

Meyer, D. & Schvaneveldt, R. W. (1971). Facilitation in recognizing parts of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Navarro, D. J. & Lee, M. D. (2003). Combining dimensions and features in similarity based representations. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems, 15*, 67-74. Cambridge, MA: MIT Press.

Navarro, D. J., Pitt, M. A. & Myung, J. I. (submitted). Does response scaling cause the Generalized Context Model to mimic a prototype model? Submitted to *Psychonomic Bulletin and Review*.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 104-114.

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

Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53-79.

Rips, L. J. (1989). Similarity, typicality and categorisation. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 21-59). Cambridge: Cambridg University Press.

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

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*, 2-28.

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. (1956). Rational choice and the structure of environments. *Psychological Review, 63*, 129-139.

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.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J. & Blum, B. (2003). Inferring causal networks from observations and intervations. *Cognitive Science, 27*, 453-487.

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

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.