Modeling Individual Differences in Category Learning

Michael R. Webb (michael.webb@dsto.defence.gov.au)
Command and Control Division, Defence Science and Technology Organisation
Edinburgh, South Australia, 5111, AUSTRALIA

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

Many evaluations of cognitive models rely on data that have been averaged or aggregated across all experimental subjects, and so fail to consider the possibility that there are important individual differences between subjects. Other evaluations are done at the single-subject level, and so fail to benefit from the reduction of noise that data averaging or aggregation potentially provides. To overcome these weaknesses, we develop a general approach to modeling individual differences using *families* of cognitive models, where different groups of subjects are identified as having different psychological behavior. Separate models with separate parameterizations are applied to each group of subjects, and Bayesian model selection is used to determine the appropriate number of groups. We demonstrate the general approach in a concrete and detailed way using the ALCOVE model of category learning and data from four previously analysed category learning experiments. Meaningful individual differences are found for three of the four experiments, and ALCOVE is able to account for this variation through psychologically interpretable differences in parameterization. The results highlight the potential of extending cognitive models to consider individual differences.

Introduction

Much of cognitive psychology, as with other empirical sciences, involves the development and evaluation of models. Models provide formal accounts of the explanations proposed by theories, and have been developed to address diverse cognitive phenomena ranging from stimulus representation (e.g., Shepard 1980), to memory retention (e.g., Anderson & Schooler 1991; Estes 1997), to category learning (e.g., Ashby & Perrin 1988; Berretty, Todd, & Martignon 1999; Kruschke 1992; Tenenbaum 1999). One recurrent shortcoming of these models, however, is that (whether intentionally, or as an unintended consequence of methodology) humans are usually modeled as 'invariants', and not as 'individuals'. This occurs because, most often, models are evaluated against data that have been averaged or aggregated across subjects, and so the modeling assumes that there are no individual differences between subjects.

The potential benefit of averaging data is that, if the performance of subjects really is the same except for 'noise' (i.e., variation the model is not attempting to explain), the averaging process will tend to remove the noise, and the resultant data will more accurately reflect the underlying psychological phenomenon. When the performance of subjects has genuine differences, however, it is well known (e.g., Estes 1956; Myung, Kim, & Pitt 2000) that averaging produces data that do not accurately represent the behavior of individuals, and provide a misleading basis for modeling.

Even more fundamentally, the practice of averaging data restricts the focus of cognitive modeling to issues of how people are the same. While modeling invariants is fundamental, it is also important to ask how people are different. Experimental data reveal individual differences in cognitive processes, and in the psychological variables that control those processes, that also need to be modeled.

Cognitive modeling that attempts to accommodate individual differences usually assumes that each subject behaves in accordance with a different parameterization of the same basic model, and so the model is evaluated against the data from each subject separately (e.g, Ashby, Maddox, & Lee 1994; Nosofsky 1986; Wixted & Ebbesen 1997). Although this avoids the problem of corrupting the underlying pattern of the data, it also foregoes the potential benefits of averaging, and guarantees that models are fit to all of the noise in the data.

Another problem with individual subject analysis, from a model theoretic perspective, is that fitting each additional subject requires an extra set of free parameters, and so leads to a progressively more complicated accounts of the data as a whole. As has been pointed out repeatedly in the psychological literature recently (e.g., Myung & Pitt 1997; Pitt, Myung, & Zhang 2002), it is important both to maximize goodness-of-fit and minimize model complexity to achieve the basic goals of modeling. Unnecessarily complicated models that "over-fit" data often do not provide any insight or explanation of the cognitive processes they address, and are less capable of making accurate predictions when generalizing to new or different situations.

A better approach, therefore, is to partition subjects according to their individual differences, and model the averaged or aggregated data from each group. Under this approach, data are addressed by a set of models, called a model *family*, where a different parameterization is applied to each group of subjects. Where averaging is appropriate, within groups of subjects, it is applied. Where averaging is not appropriate, between groups of subjects, it is not applied.

In this paper, we apply these ideas to model individual differences in category learning, using Kruschke's (1992) well known, empirically successful, and widely used ALCOVE model. Our basic approach, however, is applicable to any model of category learning or, indeed, models of other cognitive phenomena.

Modeling Individual Differences in Category Learning

Formally, a model family $\mathcal{M}$ partitions the subjects S into G groups $S \rightarrow \{S_1, \ldots, S_G\}$, and so partitions the complete data D into G averaged data sets $D \rightarrow \{D_1, \ldots, D_G\}$. For the ith data set, a model family also specifies a model parameterization θ_i. Any possible partitioning of subjects can be considered, including the possibility that all subjects are in the same partition (corresponding to aggregating across subjects), or that each has their own partition (corresponding to a complete individual analysis). Differences in the category learning processes between groups are revealed by differences in the parameter values they use.

Because of the enormous flexibility allowed by model families, they can be made almost arbitrarily complicated, and could potentially fit any data set perfectly by adding new models, with extra parameters, to account for any remaining unexplained variation in data. It is necessary, therefore, for model fitting methods to use model selection criteria that balance goodness-of-fit and model complexity. The application of Bayesian model selection criteria (e.g., Pitt *et al.* 2002) is most easily pursued by specifying a probabilistic account, in the form of a likelihood function, of the relationship between a parameterized model family and empirical data.

To develop a likelihood function for category learning, suppose, under a proposed partitioning of subjects, the ith partition has k_i subjects, and that the n category learning trials are divided into blocks, with the jth block having b_j trials. Choosing one block with $b_1 = n$ corresponds to an analysis of the average response probabilities over all trials. Choosing n blocks with all $b_j = 1$ corresponds to a trial-by-trial analysis.

In a two category learning experiment, the data take the form of counts, d_{ij}, of the number of correct responses made by all of the subjects in the ith partition on the jth block of learning trials. Suppose also that a category learning model M, with its parameterization θ_i, predicts a correct response probability of γ_{ij} at the ith group of subjects on the jth block. Then the likelihood of the data arising under the model is given by the binomial distribution: $p\left(d_{ij} \mid M_i, \theta_i\right) = \binom{b_j k_i}{d_{ij}} \gamma_{ij}^{d_{ij}} \left(1-\gamma_{ij}\right)^{b_j k_i - d_{ij}}$. The likelihood of a model family simply extends this result to consider every block of trials and every partition, so that

$$p\left(D \mid \mathcal{M}\right) = \prod_i \prod_j \binom{b_j k_i}{d_{ij}} \gamma_{ij}^{d_{ij}} \left(1-\gamma_{ij}\right)^{b_j k_i - d_{ij}}. \quad (1)$$

The extension of this likelihood function to more general category learning experiments with more than two possible category responses, using a multinomial distribution, is straightforward.

Having defined the likelihood function, the Bayesian Information Criterion (BIC: Schwarz 1978) can be applied to balance goodness-of-fit with the complexity of a model family. The BIC is given by:

$$\mathrm{BIC} = -2 \ln p\left(D \mid \theta^*\right) + P \ln N, \quad (2)$$

where P is the number of parameters in the model family (i.e., the sum of all the parameters used by the models for each group), N is the total number of data, and θ^* is the maximum likelihood parameterization over all the models. Different possible model families, corresponding to different groupings of subjects, can be compared in terms of their BIC values, with the minimum BIC corresponding to the most likely account of the data.

Demonstration Using ALCOVE

Kruschke's (1993) Study

ALCOVE is a model of category learning that uses an exemplar-based stimulus representation, similarity-based generalization that is mediated by selective attention, and error-based learning from external feedback. The standard ALCOVE model Kruschke (1992) uses four free parameters. These control the rate of learning for attention weights (λ_a), the rate of learning for the associations between stimulus representations and category responses (λ_w), the gradient of the generalization function that measures stimulus similarity (c), and the way in which different levels of evidence for category alternatives are mapped onto response probabilities (ϕ).

Kruschke (1993) considered the ability of ALCOVE to model human category learning for filtration and condensation Categorization tasks (Garner 1974). The results of four separate experiments were reported, covering two filtration tasks (called position-relevant and height-relevant, due to the nature of the stimuli) and two condensation tasks (called condensation A and

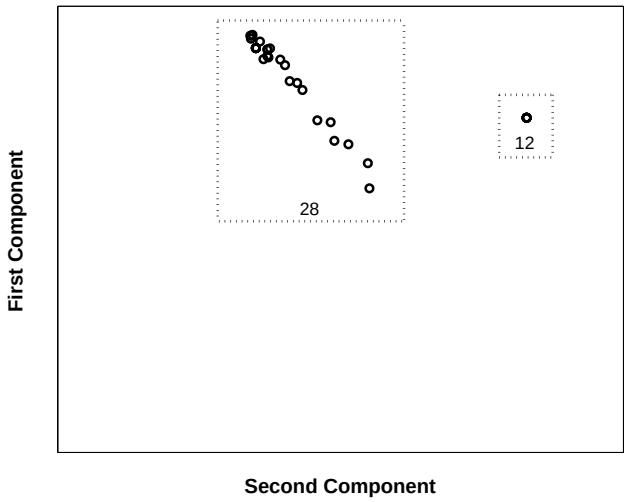

Figure 1: The application of the heuristic for partitioning subjects to find two groups for the position-relevant filtration data.

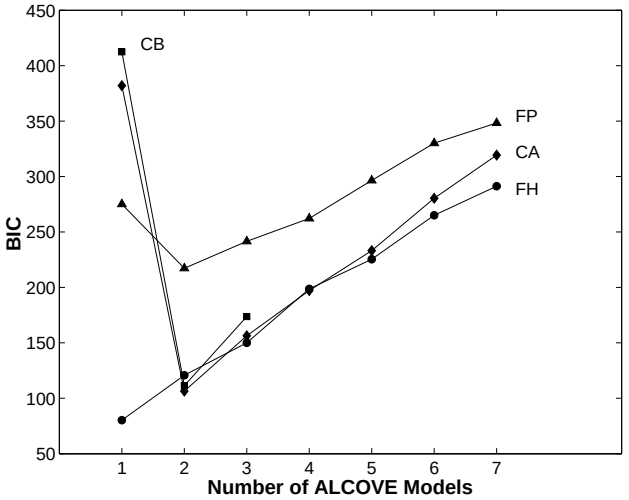

Figure 2: The pattern of change in BIC values for each clustering of the position-relevant filtration (FP), heigh-relevant filtration (FH), condensation A (CA) and condensation B (CB) category learning data.

condensation B). The data involved a total of 160 subjects, with 40 completing each task. Kruschke (1993) fit ALCOVE to all four sets of experimental results simultaneously, using trial-by-trial data formed by averaging across all 40 subjects. An examination of the individual learning curves in the raw data, however, reveals a large degree of variation between subjects within each experiment, and raises the possibility that there are psychologically meaningful individual differences in category learning.

Heuristic for Partitioning Subjects

In classification and clustering, an essential requirement for the determination of homogenous classes is a calculable similarity or distance measure between objects being compared (Gordon 1999). For category learning, the objects are the individual experimental observations for each subject, (i.e., each subject's learning curve). A candidate measure for describing the similarities between these curves is the correlation coefficient, which we used in a two-stage heuristic. In the first stage, singular value decomposition is applied to produce an ordered eigenvector-based representation of the similarities between the learning curves of subjects. In the second stage, a simple k-means clustering algorithm is applied to this representation to find clusters of subjects.

For each of Kruschke's (1993) four category learning tasks, this heuristic was applied to produce a range of partitions of the data, from a single group with all 40 subjects, to seven groups with differing numbers of subjects in each group. As a concrete example of this process, the clusters found when the subjects were divided into two groups for the position-relevant filtration task are shown in Figure 1. Each circle represents the learning curve of a subject, represented according to their values along the first two component eigenvectors. The two groups of subjects identified by k-mean clustering are superimposed using broken lines. One cluster on the left encompasses 28 of the subjects, while a much tighter cluster on the right encompasses the remaining 12 subjects.

Model Fitting and Evaluation

For each of the clusterings for each task, maximum likelihood fits of ALCOVE were found using a different parameterization for each group according to Eq. (1). BIC values were then calculated for each model family using Eq. (2), giving the results[1] shown in Figure 2. It is clear that the minimum BIC for three of the four tasks (position-relevant filtration, condensation A and condensation B) is achieved when two separate groups of subjects are considered, while the height-relevant filtration data are best modeled by considering all of the subjects as learning in the same way.

Figures 3 and 4 give more detailed results for, respectively, the position-relevant filtration and condensation

[1]The full range of BIC values for the CB task is not shown because, when four or more groups are considered, at least one of the groups contains only subjects who become less accurate as learning blocks progress. ALCOVE is qualitatively unable to accommodate the decrease in the averaged learning curve for this type of group, leading to very poor fit, and very large BIC values. We have omitted these values.

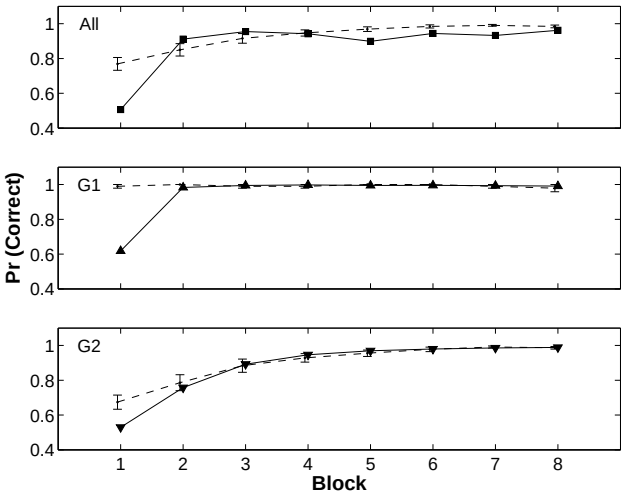

Figure 3: The change in accuracy across learning blocks for the subjects (broken lines) and ALCOVE (solid lines), for the one group ("All") and two group ("G1" and "G2") model families on the position-relevant filtration task.

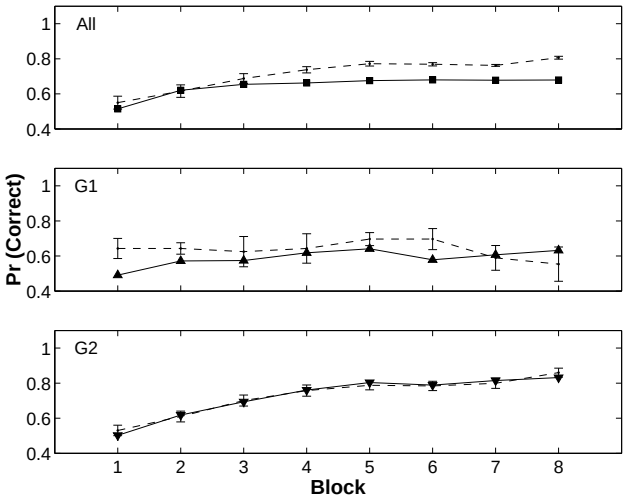

Figure 4: The change in accuracy across learning blocks for the subjects (broken lines) and ALCOVE (solid lines), for the one group ("All") and two group ("G1" and "G2") model families on the condensation A task.

A tasks. In both of these figures, the top panel, labeled "All", shows the average accuracy of all subjects across the eight learning blocks, and the maximum likelihood fit of ALCOVE to these data. The middle and bottom panels show the first (G1) and second (G2) groups of subjects proposed by the two-group model family that is prefered by the complexity analysis. These panels show the average accuracy for both groups of subjects separately, together with the maximum likelihood ALCOVE learning curve.

Figure 3 shows that the moderate learning evident when treating the subjects as having no individual differences is better modeled as coming from two distinct groups of subjects. Some subjects, in the first group, maintain near-perfect accuracy throughout the category learning task. Other subjects, in the second group, learn more gradually, only achieving near-perfect accuracy in the last few learning blocks. Figure 3 shows that, with the exception of the rapid achievement of accuracy in the first block for the first group of subjects, ALCOVE is able to model both of these patterns of learning[2].

In a similar way, Figure 4 shows that the gradual increase in accuracy, evident when treating the subjects as having no individual differences, is better modeled as coming from two distinct groups of subjects. The first group exhibits almost no learning, while the second learns at a moderate rate. Once again, ALCOVE is able to model both of these patterns of learning. In fact, ALCOVE has more difficulty accommodating the learning data resulting from averaging across all of the subjects. What the individual differences analysis developed here suggests is that this inability may not indicate a fundamental weakness in ALCOVE, but rather that the averaging process involved in summarizing human performance has masked important individual differences, and corrupted the underlying learning patterns in the original data.

Table 1 shows the maximum likelihood parameter values for each group of subjects in the model family with the lowest BIC value, for all four learning tasks. These parameter values are generally interpretable in terms of the different learning behavior revealed by the individual differences analysis. For example, for the position-relevant filtration task, the first group of subjects have a greater λ_w value than the second group, consistent with their more rapid learning. For this task, both groups have high ϕ values, consistent with their decisiveness (or 'confidence') in mapping evidence into response probabilities. Both groups of subjects in the condensation A task, however, have much lower ϕ values, consistent with their inferior learning performance, and the first group in this task, who basically fail to learn, have a very low ϕ value. Other comparisons of this type, both within and across tasks, generally have meaningful and useful interpretations,

[2]It is possible the application of one of ALCOVE's descendents, such as RASHNL (Kruschke & Johansen 1999) or the unified mixture of experts model (Kruschke 2001), which emphasize rule-oriented learning and incorporate a rapid attention shifting capability (Kruschke 1996), could overcome the deficiency.

Table 1: Maximum likelihood parameter values for each group of subjects in the model family with the lowest BIC value, for all four learning tasks. FP=position-relevant filtration, FH=height-relevant filtration, CA=condensation A, CB=condensation B.

Task	Group	λ_w	λ_a	c	ϕ
FP	G1	0.38	0.49	1.68	3.20
	G2	0.06	27.0	6.83	2.66
FH	All	0.23	0.58	1.56	1.00
CA	G1	0.47	1.14	2.53	0.27
	G2	0.24	0.38	7.52	0.93
CB	G1	0.41	0.32	0.79	0.31
	G2	0.17	0 .02	3.37	1.09

and highlight the ability of ALCOVE to represent psychologically important variations in category learning through its free parameters.

Discussion

There are at least two conclusions that can be drawn from modeling individual differences in Kruschke's (1993) category learning data using ALCOVE. The first is that there is strong evidence for large and meaningful differences in the learning behavior of groups of subjects for three out of the four tasks. Previous analyses, adopting the standard cognitive modeling practice of considering all of the subjects as a single group, are insensitive to these potentially important patterns of variation. The second conclusion is that, for these data, the basic ALCOVE model is generally able to capture the individual differences in learning, when asked to model appropriate groups of subjects. It does this by applying different psychologically meaningful parameterizations to accommodate variations in learning behavior. In this sense, what the results presented here demonstrate is that accounting for individual differences using model families has the potential to extend and increase the usefulness of existing cognitive models significantly.

From this promising start, there are a number of directions in which the basic approach described here can be refined and extended. Most generally, the extension to other cognitive phenomena provides a rich set of opportunities for future research. As with category learning, there is evidence of individual differences in the similarity data used to model stimulus representations (e.g., Ashby *et al.* 1994), and in the curves of forgetting used to model memory retention (e.g., Anderson & Tweney 1997; Heathcote, Brown, & Mewhort 2000; Myung, Kim, & Pitt 2000; Wixted & Ebbesen 1997), and in a range of other data from which cognitive models have been developed.

Considering a broader range of cognitive phenomena highlights the possibility of extending individual difference accounts to incorporate fundamentally different models to capture between-subject variation, rather than relying solely on parametric variation within the same basic model. In memory retention, for example, one group of subjects could be modeled using a power function while another group is modeled using an exponential decay function. For stimulus representation, some groups of subject could be modeled using a featural representation while others use a dimensional representation. In the category learning context considered here, it may make sense to model some subject groups using ALCOVE or its descendants, but apply a very different category learning model to others, such as the fast and frugal account provided by Categorization-By-Elimination (Berrety *et al.* 1999).

One of the weaknesses of the demonstration presented here is the reliance on the BIC to compare different competing individual differences models. While the BIC is conceptually and computationally straightforward, it is insensitive to the complexity effects arising from the functional form of parametric interaction within the individual models (Myung & Pitt 1997). This is a potentially important shortcoming, especially if fundamentally different models are used to explain performance for different subject groups. There are, for example, many competing models of retention that use two parameters (Rubin & Wenzel 1996), with different complexities that the BIC is unable to distinguish. The obvious remedy for this problem is to use more sophisticated model selection criteria that are sensitive to all of the components of model complexity. These include measures such as the Stochastic Complexity Criterion (SCC: Rissanen 1996) and Normalized Maximum Likelihood (NML: Rissanen 2001). For cognitive models that resist the formal analysis needed to derive these measures, an alternative is to use numerical methods, such Markov Chain Monte Carlo (e.g., Gilks, Richards, & Spiegelhalter 1996) to approximate the Bayesian posterior distributions that compare model families.

A final possibility for refining the approach demonstrated here is to use a more principled optimization approach to determine the groupings of subjects. The method used here, based on k-means clustering of correlations, is a sensible heuristic one. It is particularly well suited to a model like ALCOVE that requires considerable computation effort when finding maximum likelihood parameter values. The clustering heuristic is designed to identify good partitions of the subjects into groups, and only requires parameter fitting to be done once for each possible number of subject groups. For other models, however, such as analytic models of memory retention, finding maximum likelihood parameterizations is straightforward. In these cases, a more explicit optimization approach to finding partitions could be adopted, because repeated pa-

rameter fitting is possible. For example, a stochastic hill-climbing procedure could be used to find subject groups that minimize the BIC, SCC or NML of the model family.

Collectively, these possibilities describe a principled and general approach for building and evaluating cognitive models, using a variety of basic models and numbers of parameterizations, to accommodate individual differences. It is a more general approach to cognitive modeling than one that averages data, assuming there are no individual differences. It is a more powerful and succinct approach than one that uses subject-by-subject analysis. While much of the work to realize this potential remains to be done, the demonstration presented here, using multiple ALCOVE models to capture differences in category learning, provides a good concrete example of its potential. It shows how using model families, and relying on principled model selection criteria, can be used to develop detailed and interpretable accounts of both how people are cognitively the same, *and* how they are different.

Acknowledgments

This research was supported by Australian Research Council Grant DP0451793.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science 2*(6), 396–408.

Anderson, J. R., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition 25*, 724–730.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science 5*(3), 144–151.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review 95*(1), 124–150.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination. In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart*, pp. 235–254. New York: Oxford University Press.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin 53*(2), 134–140.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review 104*(1), 148–169.

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gordon, A. D. (1999). *Classification* (Second ed.). London: Chapman & Hall/CRC Press.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review 7*(2), 185–207.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review 99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science 5*, 3–36.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition 22*, 3–26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology 45*, 812–863.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition 25*(5), 1083–1119.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition 28*(5), 832–840.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review 4*(1), 79–95.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*(1), 39–57.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review 103*(4), 734–760.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science 210*, 390–398.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*, Cambridge, MA, pp. 59–65. MIT Press.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition 25*(5), 731–739.