Individual Differences in Attention During Category Learning

Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100 USA

Ruud Wetzels (wetzels.ruud@gmail.com)
Department of Psychology, University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam

Abstract

A central idea in many successful models of category learning—including the Generalized Context Model (GCM)—is that people selectively attend to those dimensions of stimuli that are relevant for dividing them into categories. We use the GCM to re-examine some previously analyzed category learning data, but extend the modeling to allow for individual differences. Our modeling suggests a very different psychological interpretation of the data from the standard account. Rather than concluding that people attend to both dimensions, because they are both relevant to the category structure, we conclude that it is possible there are two groups of people, both of whom attend to only one of the dimensions. We discuss the need to allow for individual differences in models of category learning, and argue for hierarchical mixture models as a way of achieving this flexibility in accounting for people's cognition.

Keywords: Selective attention, Category learning, Generalized Context Model, Individual differences, Hierarchical Bayesian modeling

Figure 1: Condensation category structure "B" from Kruschke (1993).

Introduction

Selective attention is one of the most compelling theoretical ideas in the study of human category learning. The basic idea is that, to learn a category structure, people selectively attend those dimensions of the stimuli that are relevant to distinguishing the categories. Nosofsky's (1984) landmark paper showed that, for stimuli represented in terms of underlying continuous dimensions, selective attention could help explain previously puzzling empirical regularities in the ease with which people learn different category structures (Shepard, Hovland, & Jenkins, 1961).

The Generalized Context Model (GCM: Nosofsky, 1984, 1986) incorporates an attention process that has proven enormously helpful in accounting for human category learning behavior. Kruschke (1992) developed a natural extension of the GCM that was able to learn selective attention weightings on a trial-by-trial basis for dimensional stimuli, and Lee and Navarro (2002) showed that the same approach worked equally well for stimuli represented in terms of discrete features rather than continuous dimensions.

In this paper, we raise the possibility that different people might apply selective attention differently when learning t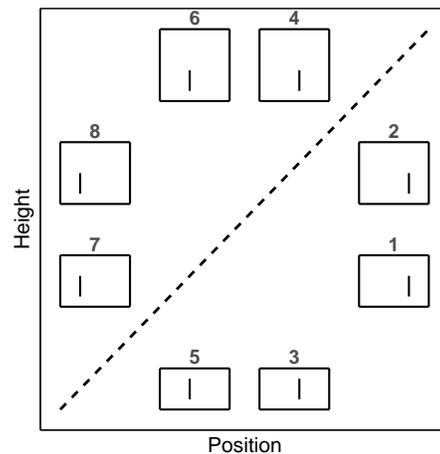he same category structure. We re-analyze human performance on a single task conducted by Kruschke (1993), using the GCM, but allowing for individual differences. We find evidence that one group of people attended primarily to one dimension of the stimuli, while a second group of people attended primarily to the other dimension. This finding runs counter to a standard analysis that does not allow for individual differences, and shows a distribution of attention across both dimensions.

Category Learning Data

The data we use in our re-analysis comes from Kruschke (1993), who studied the ability of ALCOVE to account for human learning across four category structures. Each structure involved the same eight stimuli—consisting of line drawings of boxes with different heights, with an interior line in different positions—but divided them into two groups of four stimuli in different ways. The category structure we use is the so-called "Condensation B" structure, which is shown in Figure 1. The eight stimuli are arranged by their heights and positions, and the four above and to the left of the dividing line belong to Category A. The stimuli are numbered 1–8 in the figure, for ease of reference later when we present modeling results.

Kruschke (1993) collected data from a total of 160 participants, with 40 attempting to learn each category structure. The task for each participant was, over eight consecutive blocks within which each stimulus was presented once in a random order, to learn the correct category assignment for each stimulus, based on corrective feedback provided for every trial. With the aim of analyzing human performance using the GCM—which means trial-by-trial learning is not being modeled—the data can be represented by d_{ik}, the number of times the ith stimulus was categorized as belonging to Category A by the kth participant, out of the $t = 8$ trials on which it was presented. In an analysis that does not consider individual differences, the behavioral data can be further summarized as $d_i = \sum_k d_{ik}$, the total number of times all participants classified the ith stimulus into Category A, out of $t = 40 \times 8$ total presentations.

Generalized Context Model Analysis

In this section, we present a standard version of the GCM, show how it can be formulated as a graphical model to enable fully Bayesian statistical inference[1], and present its application to the current data.

The Standard GCM

The GCM assumes that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. For the current data, there are only two dimensions, so the ith stimulus is represented by the point (p_{i1}, p_{i2}). The first dimension has an attention weight, w with $0 \leq w_d \leq 1$, and the second dimension then has an attention weight $(1-w)$. These weights act to 'stretch' attended dimensions, and 'shrink' unattended ones. Formally, the psychological distance between the ith and jth stimuli is $d_{ij}^2 = w(p_{i1} - p_{j1})^2 + (1-w)(p_{i2} - p_{j2})^2$.

The GCM assumes classification decisions are based on similarity comparisons with the stored exemplars, with similarity determined as a nonlinearly decreasing function of distance in the psychological space. We follow Nosofsky (1986) and model the similarity between the ith and jth stimuli as $s_{ij} = \exp(-c^2 d_{ij}^2)$, where c is a generalization parameter. The GCM also assumes that categories are represented by individual exemplars. This means that, in determining the overall similarity of a presented stimulus i to Category A, every exemplar in that category is considered, so that the overall similarity is $s_{iA} = \sum_{j \in A} s_{ij}$. Final categorization response decisions are based on the Luce Choice rule, as applied to the overall similarities. We assume an unbiased version of the choice rule, so that the probability that the ith stimulus

[1]Note that this does *not* mean we are proposing a "Bayesian" or "rational" version of the GCM (cf. Griffiths, Kemp, & Tenenbaum, 2008). We are simply using Bayesian statistics, rather than traditional model-fitting methods and frequentist statistical approaches, to make inferences about GCM parameters from data. That is, we are using Bayesian inference as statisticians do, and as psychologists should do, to relate models to data.
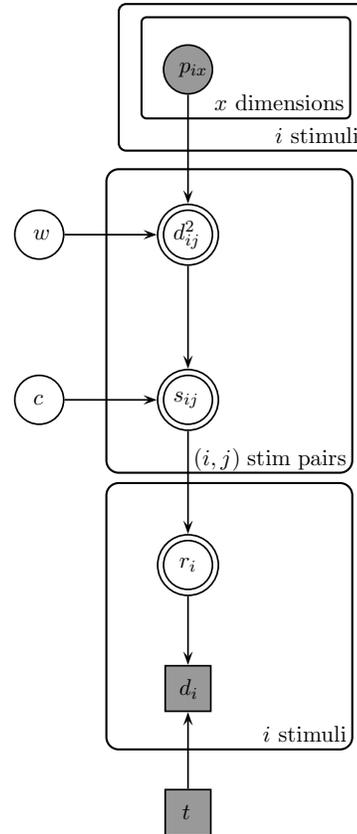

Figure 2: Graphical model implementation of the GCM.

will be classified as belonging to Category A, rather than Category B, is modeled as $r_i = s_{iA} / (s_{iA} + s_{iB})$. The observed decision data themselves are then simply modeled as $d_i \sim \text{Binomial}(r_i, t)$, meaning that each of the t presentations of the ith stimulus has a probability r_i of being categorized as belonging to Category A.

Graphical Modeling Implementation

Our analyses are implemented using the formalism provided by graphical models. A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. Details and tutorials aimed at cognitive scientists are provided by Lee (2008) and Shiffrin, Lee, Kim, and Wagenmakers (2008). The practical advantage of graphical models is that sophisticated and relatively general-purpose Markov Chain Monte Carlo (MCMC) algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data. Our analyses rely on WinBUGS (Spiegelhalter, Thomas, & Best, 2004), which is easy-to-learn software for implementing and analyzing graphical models (see Lee & Wagenmakers, 2010).

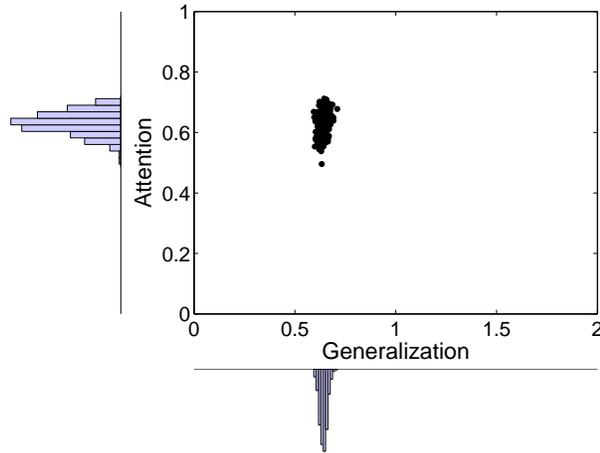A graphical model implementation of the GCM is shown in Figure 2. The known stimulus locations p_{ix}, to-

Figure 3: Joint and marginal posterior distributions over attention w and generalization c parameters of the GCM, when applied to the condensation data.

gether with the attention parameter w generate the pairwise distances d_{ij}^2. These distances, together with the generalization parameter c generate the pairwise similarities. These similarities, in turn, lead to response probabilities r_i which generate the observed data d_i.

Results

Our results are based on 3 chains of 5,000 samples each, with a burn-in of 1,000 samples, whose convergence was checked using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997).

The key result is shown in Figure 3, which plots the joint posterior distribution of the generalization and attention parameters (as a scatterplot), as well as their marginal distributions (as histograms). The marginal posterior for the attention parameter w—which gives the weight for the position dimension—lies between about 0.55 and 0.7. This result can be interpreted as showing that people give significant attention to both dimensions, although they are probably focusing a little more on the line position than the rectangle height. In condensation tasks, both stimulus dimensions are relevant to determining how stimuli belong to categories, and so the shared attention result makes sense. In other words, the standard application of the GCM produces a psychologically reasonable inference about selective attention, and it is tempting to view this analysis as the end of the story.

Individual Differences Analysis

The standard analysis assumes, however, that all people used exactly the same parameterization of the GCM to guide their category learning. But an examination of the individual learning curves in the current data suggests a large degree of variation between subjects, and raises the possibility that there are psychologically meaningful individual differences.

Types of Individual Differences

Figure 4 gives a schematic picture of four different assumptions about individual differences. Each panel shows a data space, containing the possible outcomes of an experiment. In the No Differences panel, there is a single true point, represented by the circular marker, corresponding to one parameterization of a cognitive process. The gray circles show the variety of behavioral data that might actually be produced in an experiment. The assumption of no individual differences means the goal of inference would be to find the circular marker from the gray points, and corresponds to the standard analysis of the GCM we have presented.

In the Continuous Differences panel there are many true points, again shown by circular markers. Each of these points could correspond to an individual subject's data from an experiment. The individuals are not identical (i.e., there is no longer a single point), but nor are they unrelated (i.e., their points are not spread across the entire data space). This sort of individual differences can be accommodated by hierarchical or multi-level models, in which there is a single hierarchical group distribution over the parameters of the individuals (e.g., Rouder & Lu, 2005).

In the Discrete Differences panel there are two true points, shown by a circular and a square marker. Each of these points could correspond to the data from different individuals, or from different subgroups, each with multiple individuals, in an experiment. The two points correspond to fundamentally different parameterizations of a cognitive process, or even to fundamentally different cognitive processes, and so the overall data is a mixture of two different cognitive processes. Mixture models are typically used to accommodate this sort of individual differences (e.g., Lee & Webb, 2005).
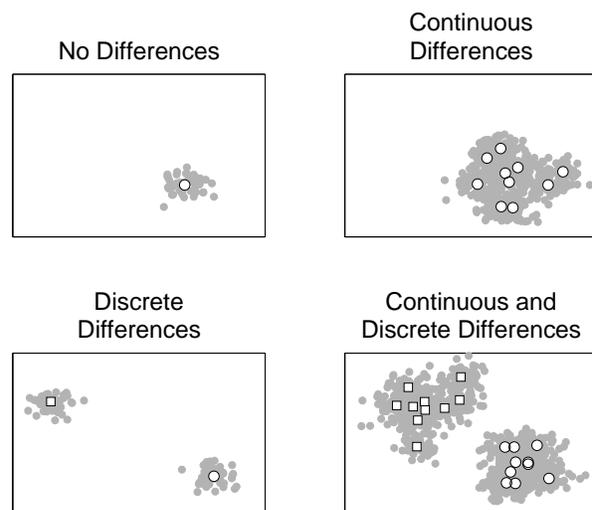

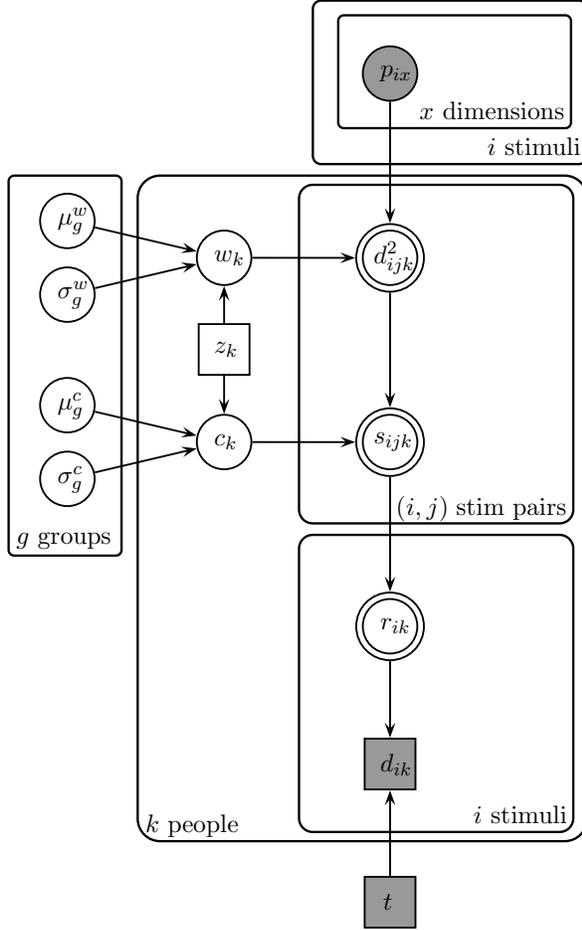Figure 4: Four different assumptions about individual differences.

Figure 5: Graphical model for the GCM with individual differences.

Graphical Model Implementation

Figure 5 shows the graphical model that extends the GCM to allow for continuous and discrete individual differences. There is now a plate for the participants, so that the kth participant has attention w_k and generalization c_k parameters. These are drawn hierarchically

from one of a set of Gaussian distributions depending on their group membership z_k. Formally, this means $w_k \sim \text{Gaussian}(\mu_{z_k}^w, \sigma_{z_k}^w)$ and $c_k \sim \text{Gaussian}(\mu_{z_k}^c, \sigma_{z_k}^c)$.

Statistically, this is a hierarchical (or "random-effect") mixture model. Psychologically, people belong to different qualitative groups, given by z_k, and their attention and generalization parameters are sampled from a continuous Gaussian distribution corresponding to their group.

We put standard vague priors on the group means and standard deviations, and on the latent assignment indicator variables. We then applied this extended GCM model to the current condensation data, assuming there were two groups of participants.

Results

Once again, our results are based on 3 chains of 5,000 samples each, with a burn-in of 1,000 samples, whose convergence was checked. Our key findings are laid out in Figure 6. The top-most bar graph shows the inferred allocation of the 40 participants into the two groups, as measured by the posterior expectation of the z_k variable. There are unambiguous assignments for 36 participants, with 24 belonging to Group 1 and 12 belonging to Group 2. This lack of uncertainty in mixture model latent assignment is usually an indication that there are multiple groups.

The attention and generalization properties of the two groups, in the form of the joint and marginal posterior distributions of μ_g^w and μ_g^c, are shown in the next two panels. Group 1 on the left has an attention weight above 0.8, while Group 2 on the right has an attention weight close to 0. The natural interpretation is that the first group of participants is primary attending to the position dimension, while the second group is almost exclusively attending to the height dimension.

Below the posterior distribution for the groups, a posterior predictive check of fit to the behavioral data is shown. For each of the 8 stimuli the posterior predictive distribution over the number of times it is classified as belonging to Category A is shown by the squares, with the area of each square being proportional to posterior predictive mass. The single thick line shows the average observed categorization behavior for those participants assigned to the group. The many thin lines show the individual participant behavior for the group. It is clear that Group 1 and Group 2 have participants showing qualitatively different patterns of categorizing the stimuli, and these differences are captured by the posterior predictive distributions.

The bottom-most panels in Figure 6 interpret the different category learning of the groups. The original stimulus space and category structure is shown, with bars showing the average number of times each stimulus was placed in Category A and Category B by members of the group. To understand Group 1, note that stimuli 4 and 5 are the ones least clearly categorized correctly. This is consistent with a focus on the position dimension, which would assign these two stimuli incorrectly. Similarly, for Group 2, stimuli 2 and 7 are categorized very poorly.

The obvious strategy for a more complete account of individual differences is to combine both Continuous and Discrete differences, as in the bottom-right panel on Figure 4. Here, there are two types of true points—indicated by circular and square makers—and constrained individual variation within each type. A combination of both hierarchical and mixture modeling naturally deals with these patterns of differences. The mixture component identifies the fundamentally different cognitive processes, and the hierarchical component captures the variation within each process. We are not aware of cognitive modeling that has adopted this approach, but it seems the most general and natural way to extend the GCM analysis.
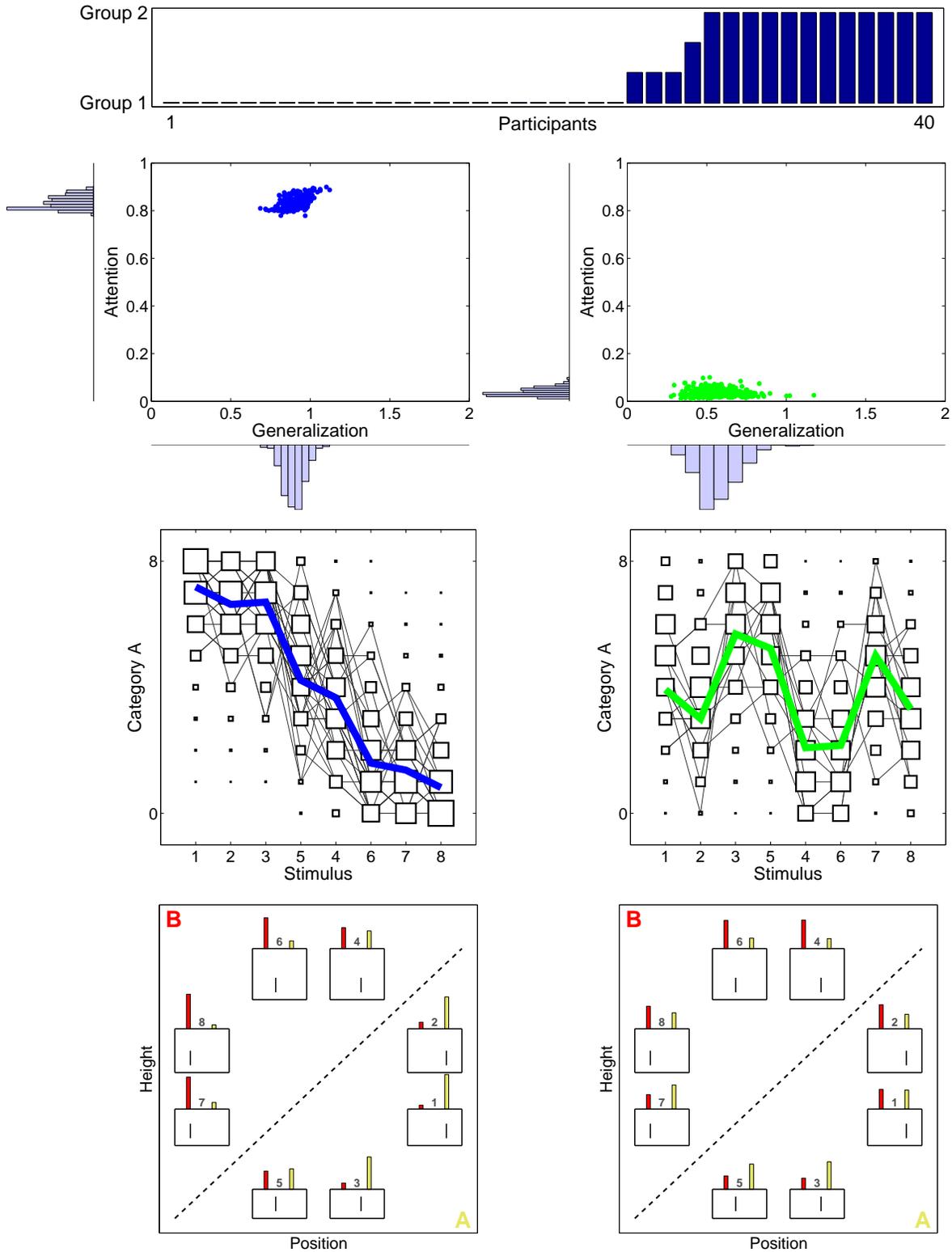
Figure 6: Results from GCM analysis assuming two groups of participants, showing the allocation of participants to groups, posterior and posterior predictive distributions for the groups, and the interpretation of the different groups in terms of the stimuli and category structure itself. See text for details.

This is consistent with a focus on the height dimension.

Finally, we compared a one-group to a two-group model, calculating the Bayes Factor using the Savage-Dickey method described by Wetzels, Grasman, and Wagenmakers (2010). This came out about 2.3 in favor the two-group model, meaning that the data are more than twice as likely to have come from two groups of participants than a single group. While this is far from conclusive evidence, it does suggest that the possibility there are two different groups of participants deserves serious consideration.

Discussion

Our extended analysis of Kruschke's (1993) condensation data, using a GCM with the ability to detect continuous and discrete individual differences, tells an interesting story. It suggests that there are two groups of participants, each of whom focus most of their attention on just one stimulus dimension while learning the category structure. The standard result of attention being distributed roughly evenly across both dimensions seems to be an artefact of failing to consider individual differences in modeling.

We realize that applying the GCM to the condensation data is non-standard, because the GCM is usually applied to category learning experiments with a training and a testing phase, rather than a single category learning sequence. Ideally, our modeling would be applied to transfer data collected after categories were learned to criterion, and it is possible the dynamics of learning provide a partial explanation for the individual differences we observe, although we do not think they can provide a full explanation. We also realize that there are many possible variations of the GCM that could be tried.

Accordingly, we certainly do not claim our single re-analysis automatically undermines the existing large and coherent body of work examining selective attention mechanisms in category learning. Systematic investigation of category learning across many tasks, looking for the presence of discrete and continuous individual differences, is needed to gauge the generality of our current results. We think this would be a worthwhile exercise, given the theoretical influence of selective attention mechanisms in the category learning literature.

We also think our analyses underscore a more general point, which is that it is important to consider and model individual differences in all of cognition. Finally, we think the ease with which very general assumptions about individual differences could be implemented to extend the standard GCM analysis shows the advantage of using Bayesian statistics to relate cognitive models to data.

References

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

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.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*(1), 43–58.

Lee, M. D., & Wagenmakers, E.-J. (2010). *A Course in Bayesian Graphical Modeling for Cognitive Science.* Course notes, University of California Irvine. [http://www.socsci.uci.edu/∼mdlee/bgm].

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

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

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573-604.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *13*.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis*, *54*, 2094–2102.