Psychonomic Bulletin & Review
2005, 12 (4), 605-621

Modeling individual differences in cognition

MICHAEL D. LEE
University of Adelaide, Adelaide, Australia

and

MICHAEL R. WEBB
Defence Science and Technology Organisation, Edinburgh, Australia

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 of 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 have developed a general approach to modeling individual differences using families of cognitive models in which 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 evaluate this individual differences approach in a simulation study and show that it is superior in terms of the key modeling goals of prediction and understanding. We also provide two practical demonstrations of the approach, one using the ALCOVE model of category learning with data from four previously analyzed category learning experiments, the other using multidimensional scaling representational models with previously analyzed similarity data for colors. In both demonstrations, meaningful individual differences are found and the psychological models are able to account for this variation through interpretable differences in parameterization. The results highlight the potential of extending cognitive models to consider individual differences.

Much of cognitive psychology, like 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 (see, e.g., Shepard, 1980; Tversky, 1977) to memory retention (e.g., Anderson & Schooler, 1991; Estes, 1997; Laming, 1992) 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, not as individuals. This occurs because most often, models are evaluated against data that have been averaged or aggregated across subjects, 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 that the model is not attempting to explain), the averaging process will tend to remove the effects of 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 (see, e.g., Estes, 1956; Myung, Kim, & Pitt, 2000) that averaging produces data that do not accurately represent the behavior of individuals and provides 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. Although 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, so the model is evaluated against the data from each subject separately (see, 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 forgoes 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 param-

This research was supported by Australian Research Council Grant DP0451793. We thank Helen Braithwaite, Douglas Vickers, and Matthew Welsh for feedback and Robert Nosofsky and several anonymous reviewers for detailed and constructive comments. Correspondence relating to this article may be sent to M. D. Lee, Department of Psychology, University of Adelaide, SA 5005, Australia (e-mail: michael.lee@psychology.adelaide.edu.au).

eters, and so leads to progressively more complicated accounts of the data as a whole. As has been pointed out repeatedly in the psychological literature recently (e.g., by Myung & Pitt, 1997, and Pitt, Myung, & Zhang, 2002), it is important both to maximize goodness of fit and to minimize model complexity in order to achieve the basic goals of modeling. Unnecessarily complicated models that "over-fit" data often provide less insight and explanation of the cognitive processes they address and are less capable of making accurate predictions when generalized to new or different situations.

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

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 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 possibilities that all subjects are in the same partition (corresponding to averaging across subjects) or that each has a separate partition (corresponding to a complete individual analysis). Differences in psychological processes between groups are modeled by differences in the parameter values of the groups.

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 groups, with extra parameters, to account for any remaining unexplained variation in the data. It is necessary, therefore, for the fitting methods to use model selection criteria that balance goodness of fit and model complexity. This balance can be achieved in a principled way, through the application of Bayesian model selection criteria (see, e.g., Pitt et al., 2002).

In this article, we evaluate the individual differences approach in a simulation study and show that it is superior in terms of the key modeling goals of understanding and prediction. We then provide two practical demonstrations of the approach, one using the ALCOVE model of category learning with data from four previously analyzed category learning experiments, the other using multidimensional scaling representational models with previously analyzed similarity data for colors.

EVALUATION USING A SIMULATION STUDY

Consider a psychological experiment involving m participants, each of whom makes n independent binary decisions, designed to model the rate at which subjects make one decision rather than the other.[1] This framework is a fairly general one, incorporating the basic aspects of many memory and decision-making experiments. For example, in a memory retention task, one of the two alternatives may be the correct one, meaning that the cognitive modeling interest is in the rate of memory retention. Alternatively, in a more general decision-making task, neither choice may be "correct," but the interest is in the bias people have for one option over the other. These examples suggest that the binary observation experiment is a simple but realistic framework. In this section, we study how the average, individual, and group modeling approaches perform in achieving the basic goals of scientific modeling—to explain observed data in terms of existing theory and to predict future data—on plausible versions of the binary experiment.

Model Selection and Parameter Estimation

The raw data from the binary experiment take the form of a set of counts of one of the decisions (called "successes," in a possibly arbitrary way) made by each subject. We denote by k_i the number of successes for the ith subject, where $i = 1, \ldots, m$. Given these data, it is possible to compare the average, group, and individual models using Bayesian model selection in the following way.

Average approach. The average approach uses the model M_{ave}, which assumes that every subject has the same underlying rate of success, given by the parameter θ. This means that the probability of a subject's having j successes out of n trials is given by

$$p(j, n \mid \theta, M_{\text{ave}}) = \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

The likelihood of all of the data $D = (k_1, \ldots, k_m)$ under the average model with a particular rate θ is therefore given by the multinomial

$$p(D \mid \theta, M_{\text{ave}}) = \binom{m}{c_0 \ldots c_n} \prod_{j=0}^{n} \left[\binom{n}{j} \theta^j (1-\theta)^{n-j} \right]^{c_j},$$

where

$$\binom{m}{c_0 \ldots c_n}$$

is the multinomial coefficient $m!/(c_0! \ldots c_n!)$. The value c_j denotes the number of subjects with j successes, with $j = 0, \ldots, n$ and

$$\sum_{j=0}^{n} c_j = m.$$

We take the "best" value for the rate θ to be the one that maximizes this likelihood, so that

$$\theta^* = \operatorname*{argmax}_{\theta} p(D \mid \theta, M_{\text{ave}}).$$

For Bayesian model selection, the key quantity is the probability of the data under the model as a whole, which is found by integrating across all possible rate parameter

values, weighted by their prior probabilities. This marginal density can be found analytically as

$$p(D \mid M_{\text{ave}}) = \int_0^1 p(D \mid \theta, M_{\text{ave}})\, p(\theta \mid M_{\text{ave}})\, d\theta$$
$$= \binom{m}{c_0 \cdots c_n} \prod_{j=0}^{n} \binom{n}{j}^{c_j} \cdot \text{Beta}\left(\sum_{j=0}^{n} j c_j + 1, \sum_{j=0}^{n} (n-j) c_j + 1 \right),$$

where $\text{Beta}(u+1, v+1) = u!v!/(u+v+1)!$, and we have used the uniform prior $p(\theta \mid M_{\text{ave}}) = 1$. Since we know that both "zero" and "one" outcomes are possible in the experiment, this is the correct (and unique) choice of prior within the "objective Bayesian" framework for statistical inference (for details, see Jaynes, 2003, pp. 382–386).

Individual approach. The full individual differences approach uses the model M_{ind}, which assumes that every subject has a potentially different underlying rate of success, given by the parameter θ_i for the ith subject. The probability of all of the data D is therefore given by

$$p(D \mid \theta_1, \ldots, \theta_m, M_{\text{ind}}) = \prod_{i=1}^{m} \binom{n}{k_i} \theta_i^{k_i} (1-\theta_i)^{n-k_i}.$$

The maximum likelihood values for each of the rates are now given by

$$(\theta_1^*, \ldots, \theta_m^*) = \underset{(\theta_1, \ldots, \theta_m)}{\text{argmax}}\, p(D \mid \theta_1, \ldots, \theta_m, M_{\text{ind}}),$$

and the marginal density under uniform priors is

$$p(D \mid M_{\text{ind}}) = \prod_{j=0}^{n} \binom{n}{j}^{c_j} \prod_{i=1}^{m} \text{Beta}(k_i + 1, n - k_i + 1).$$

Group approach. The group approach assumes that the subjects should be partitioned into g groups, where each group has its own rate of success, given by the parameter θ_i for the ith group. This means that unlike in the average and individual approaches, a family of models $\mathcal{M}_{\text{grp}}$ is considered, with each model in the family corresponding to a group of subjects. It is obvious that the best allocation of subjects to groups will place together subjects with similar numbers of successes. This means that the only partitions that need to be considered are those having $g - 1$ boundaries that identify the number of successes used to place subjects into groups. We denote by l_i and h_i, respectively, the lowest and highest numbers of successes for the ith group and by m_i the number of subjects in the ith group. This means that the probability of the data D as a whole is given by

$$p(D \mid \theta_1, \ldots, \theta_g, l_1, \ldots, l_g, h_1, \ldots, h_g, \mathcal{M}_{\text{grp}}) =$$
$$\prod_{i=1}^{g} \left\{ \binom{m_i}{c_{l_i} \cdots c_{h_i}} \prod_{j=l_i}^{h_i} \left[\binom{n}{j} \theta_i^j (1-\theta_i)^{n-j} \right]^{c_j} \right\}.$$

The marginal density for a group model under uniform priors is

$$p(D \mid l_1, \ldots, l_g, h_1, \ldots, h_g, \mathcal{M}_{\text{grp}}) =$$
$$\prod_{j=0}^{n} \binom{n}{j}^{c_j} \prod_{i=1}^{g} \left\{ \binom{m_i}{c_{l_i} \cdots c_{h_i}} \cdot \text{Beta}\left(\sum_{j=l_i}^{h_i} j c_j + 1, \sum_{j=l_i}^{h_i} (n-j) c_j + 1 \right) \right\}.$$

The best group model, M^*_{grp}, corresponds to the choice of partition that maximizes this marginal density, and so is defined by

$$(l_1^*, \ldots, l_g^*, h_1^*, \ldots, h_g^*) =$$
$$\underset{(l_1, \ldots, l_g, h_1, \ldots, h_g)}{\text{argmax}}\, p(D \mid \theta_1, \ldots, \theta_g, l_1, \ldots, l_g, h_1, \ldots, h_g, \mathcal{M}_{\text{grp}}).$$

The maximum likelihood values for each of the group rates are found using this best group model, so that

$$(\theta_1^*, \ldots, \theta_g^*) =$$
$$\underset{(\theta_1, \ldots, \theta_g)}{\text{argmax}}\, p(D \mid \theta_1, \ldots, \theta_g, l_1^*, \ldots, l_g^*, h_1^*, \ldots, h_g^*, \mathcal{M}_{\text{grp}}),$$

which can be expressed more succinctly as

$$(\theta_1^*, \ldots, \theta_g^*) = \underset{(\theta_1, \ldots, \theta_g)}{\text{argmax}}\, p(D \mid \theta_1, \ldots, \theta_g, M^*_{\text{grp}}).$$

The best group model found in this way is the one with between 2 and $n - 1$ groups (i.e., those possibilities for which the best partitioning of subjects must be found). As was noted earlier, however, the group approach encompasses the average and individual approaches as special cases. Those possibilities are not included in the model family $\mathcal{M}_{\text{grp}}$, however, because there is no decision to be made about partitioning. Accordingly, the group approach should reduce to the average or the individual model in those cases in which one of these models makes the data more likely. Formally, this means that the final model used by the group approach is the one corresponding to the maximum of $p(D \mid M_{\text{ave}})$, $p(D \mid M^*_{\text{grp}})$, and $p(D \mid M_{\text{ind}})$.

Summary of the three approaches. A useful summary of the three modeling approaches is provided by Figure 1. It summarizes all of the data from a single trial of a binary experiment with $n = 20$ observations from $m = 50$ subjects by showing the relative numbers of subjects with 0, 1, . . . , 20 successes. Beneath the histogram, the maximum likelihood parameterizations assuming 1, 2, 3, 4, and 50 groups of subjects are shown by circles. For the 2-, 3-, and 4-group models, the partitions that correspond to the most likely model within the family are also indicated.

The average approach to modeling these data would use the single parameter provided by the 1-group model.

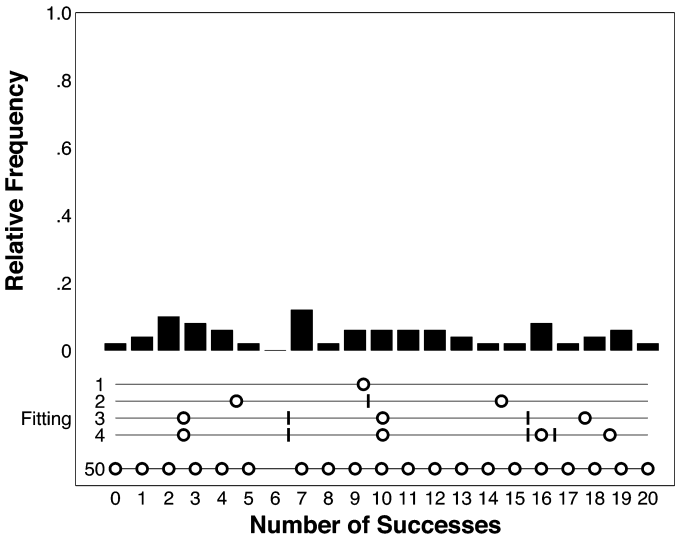

Figure 1. Data and model fitting for a single trial with $n = 20$ observations from $m = 50$ subjects. The data are summarized in the histogram, which shows the relative numbers of subjects with 0, 1, . . . , 20 successes. The maximum likelihood parameterizations using 1, 2, 3, 4, and 50 groups are shown below, together with the partition boundaries for the 2-, 3-, and 4-group models.

The individual approach would use the 50 different parameter values provided by the 50-group model. The group approach, as was explained earlier, would use the model for which the data provide maximal evidence [i.e., the maximum of $p(D|M_{\text{ave}})$, $p(D|M^*_{\text{grp}})$, and $p(D|M_{\text{ind}})$] and its associated parameter values. Intuitively, the additional possibilities considered by the group approach seem to be worthwhile for this example. The single parameter for the 1-group model does not seem to capture the distribution of the data, but 50 parameters seem to provide an overly complicated account. Something like the 3-group model, which effectively divides the subjects into low-, medium-, and high-success groups, seems to provide an appropriately simple but accurate account of the observed performance of the subjects.

Simulation Study

Basic assumptions. Any simulation study must create an artificial environment in which to study the phenomena of interest. It is important to be explicit about the assumptions made in doing this, because they determine the usefulness of the simulation. We have made three basic assumptions in creating binary experiments to generate data for testing the average, individual, and group modeling approaches. Our first assumption is the simplest and concerns realistic numbers of subjects and data in psychological experiments. We considered every possible combination of $m = 20$, 50, and 100 subjects with $n = 5$, 10, 20, and 40 observations.

Our second assumption concerns the nature of individual differences in psychological phenomena. Although there is clearly room for debate on this issue, we think a reasonable assumption is that there are some phenomena that have no meaningful individual differences, many phenomena that have a small number of qualitatively different possibilities, and some phenomena that are unique for all individuals. To comply with this assumption, for each binary experiment in the simulation study we chose at random a "true" number of individually different groups from the possibilities 1, 2, 3, 4, 5, and n. Once this choice was made, "true" rates were independently selected from the uniform distribution on [0, 1] for each group. We denote by γ_i the true rate for the ith group. The n subjects were randomly allocated to the groups, under only the constraint that each group must have at least 1 subject.

Our final assumption concerns the stability of individual differences over time. Suppose, for example, a memory retention task involves three meaningfully different groups of subjects, corresponding to people with low, moderate, and high probabilities of success in remembering. Whereas it is reasonable to expect the high achievers to perform well over repeated experiments, it seems unlikely that they will all remember with exactly the same rate of success or that any individual will remember with exactly the same rate of success in each replication of the experiment. A more plausible assumption is that the success rate of an individual on any particular experiment is drawn from a "high distribution" that is distinct from the moderate and low distributions. In other words, the important regularity is that there are three groups of people with broadly different retention

rates, but the rate for each individual within a group, or for the same individual across experiments, is subject to variation. To comply with this assumption, we selected the rate for a particular subject in the ith group from a Gaussian distribution with mean γ_i and variance σ^2 and did this independently for each experiment. We considered the values $\sigma = 0$, .05, and .10. Setting $\sigma = 0$, of course, corresponds to not making the third assumption, because it fixes rates across repeated experiments for each subject. Although we think that this situation is implausible, it nonetheless seemed worth evaluating to avoid the criticism that our assumptions were prejudiced against the individual approach, which clearly will benefit from the lack of variation.

Evaluation measures. Making these assumptions allowed experimental data to be generated that could be modeled by the average, individual, and group approaches. We evaluated the three approaches in terms of the basic modeling goals of understanding or explanation, on the one hand, and prediction or generalization, on the other.

The explanation provided by each approach was measured by the difference between the parameter values estimated by the models and the known true rates used to generate the data. For the average model, which uses a single parameter θ^* to explain all of the data, the mean difference between estimated and true parameters is

$$\frac{1}{m}\sum_{i=1}^{m}\left|\theta^* - \gamma_i\right| \equiv \frac{1}{m}\left\|\left(\theta^*,\ldots,\theta^*\right) - \left(\gamma_1,\ldots,\gamma_m\right)\right\|_1,$$

where $\|\cdot\|_1$ denotes the L_1 norm. For the individual model, which uses a separate parameter θ_i^* for each subject, the mean difference is

$$\frac{1}{m}\sum_{i=1}^{m}\left|\theta^* - \gamma_i\right| \equiv \frac{1}{m}\left\|\left(\theta_1^*,\ldots,\theta_m^*\right) - \left(\gamma_1,\ldots,\gamma_m\right)\right\|_1.$$

For the group model, subjects are allocated to g groups according to a mapping Γ, so if the ith subject belongs to the kth group, $\Gamma(i) = k$, and the kth group has parameter θ_k^*, the mean difference is

$$\frac{1}{m}\sum_{i=1}^{m}\left|\theta_{\Gamma(i)}^* - \gamma_i\right| \equiv \frac{1}{m}\left\|\left(\theta_{\Gamma(1)}^*,\ldots,\theta_{\Gamma(m)}^*\right) - \left(\gamma_1,\ldots,\gamma_m\right)\right\|_1.$$

To evaluate the ability of the three modeling approaches to make accurate predictions, we generated simulated data for two experiments. The model parameters were estimated using only the data from the first experiment, $D = (k_1, \ldots, k_m)$, but were assessed against the data from the second experiment, $D' = (k_1', \ldots, k_m')$. To make this assessment, we used the standard Bayesian approach of measuring the (quasi) posterior predictive densities (Gelfand & Dey, 1994). These measures basically assess how likely the second set of data are, given the model that has been learned from the first set, and are given by the conditional probabilities $p(D' \mid \theta^*, M_{\text{ave}})$, $p(D' \mid \theta_1^*, \ldots, \theta_g^*, M_{\text{grp}}^*)$, and $p(D' \mid \theta_1^*, \ldots, \theta_m^*, M_{\text{ind}})$ for the average, group, and individual approaches, respectively.

We compared the predictive densities on the log-odds scale, taking the density for the group approach as a reference. This means that the two final measures of comparative predictive ability are

$$\log\frac{p\left(D' \mid \theta_1^*,\ldots,\theta_g^*, M_{\text{grp}}^*\right)}{p\left(D' \mid \theta^*, M_{\text{ave}}\right)}$$

and

$$\log\frac{p\left(D' \mid \theta_1^*,\ldots,\theta_g^*, M_{\text{grp}}^*\right)}{p\left(D' \mid \theta_1^*,\ldots,\theta_m^*, M_{\text{ind}}\right)},$$

which measure the superiority (or inferiority, if negative) of the group model over the average and individual models, respectively.

Summary of the simulation study. A useful summary of the simulation study framework is provided by Figure 2. It shows the data-generating process for two experiments involving 5 subjects partitioned into two different groups. Subjects A and B are in the low-success group, and subjects C, D, and E are in the high-success group. The low-success subjects have probabilities of success p_A and p_B, drawn from the Gaussian distribution with mean γ_1 and variance σ^2 in Experiment 1, resulting in data k_A and k_B, which count their number of successes from n trials. Similarly, the high-success subjects have probabilities of success p_C, p_D, and p_E, drawn from the Gaussian distribution with mean γ_2 and variance σ^2 in Experiment 1, resulting in data k_C, k_D, and k_E. In Experiment 2, the subjects remain in the same groups but now have different success probabilities $p_A', \ldots, p_E'$ drawn from the same distributions, resulting in data $k_A', \ldots, k_E'$.

Model fitting was based on the data from Experiment 1, with the results for one, two, and five parameters being estimated for the average, group, and individual models, respectively. The evaluation of these models was then made in relation to the data from Experiment 2. In terms of explanation, the mean absolute difference across all subjects was found between the known mean of their Gaussian distribution and their corresponding parameter estimate. In terms of prediction, the posterior likelihood of the data $k_A', \ldots, k_E'$ was found for each model using its estimated parameter values.

Note that we did not evaluate the ability of the models to recover the "true" underlying numbers of groups. The obvious reason for not considering this is that the average and individual approaches would almost always be wrong, because they make assumptions that are contrary to those we made to generate the data. More fundamentally, we do not believe it is helpful to assess the ability of the group approach to recover the true number of groups.

A basic argument from advocates of the minimum description length approach to model evaluation (e.g., Grünwald, 1998; Rissanen, 2001) is that real data are generated by statistical processes that are not known and, indeed, are perhaps not knowable. Philosophically, this

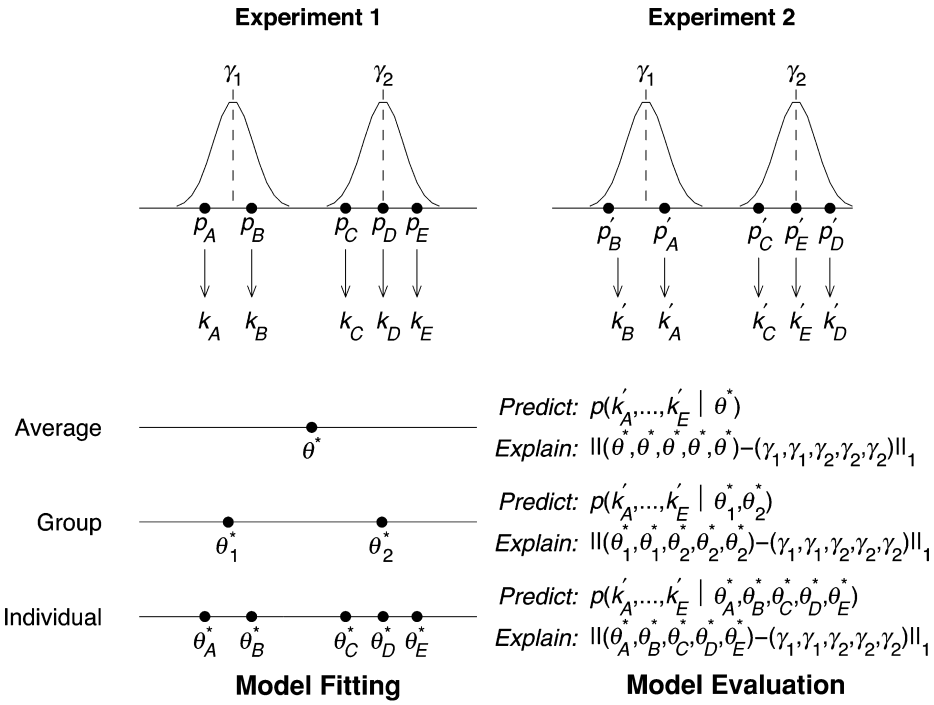

Figure 2. A summary of the simulation study framework for an example with 5 subjects divided into two groups, showing the method of data generation for the two experiments and how the models are fitted and evaluated.

means that models should aim to be "useful" rather than "true." Practically, this means that good models are those that capture regularities in data in ways that provide insight and facilitate prediction, regardless of whether they correspond to the putative truth among some limited set of models that are certain to be inadequate for expressing the statistical properties of the data.

Figure 3 makes this general point in a concrete way by showing a sample trial from the binary experiment we are considering. It involves 10 subjects partitioned into

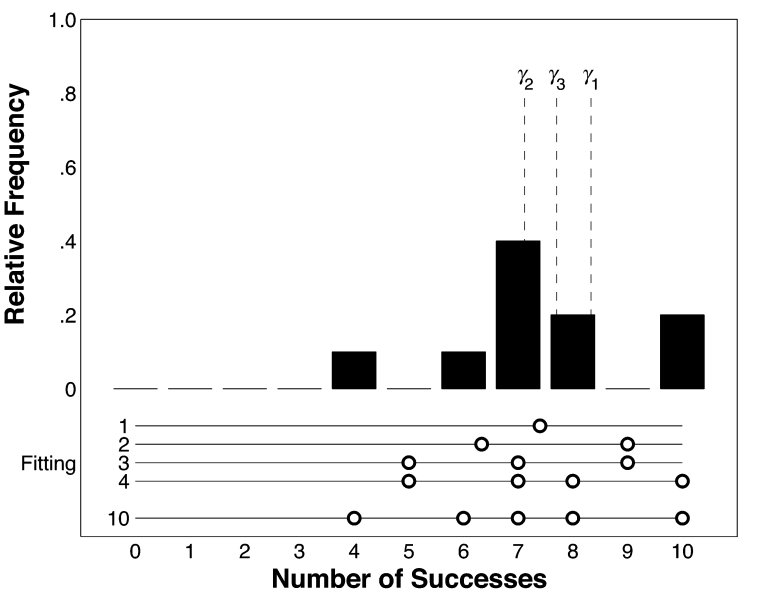

Figure 3. A sample trial with 10 subjects partitioned into three groups, each of which has a similar mean probability of success.

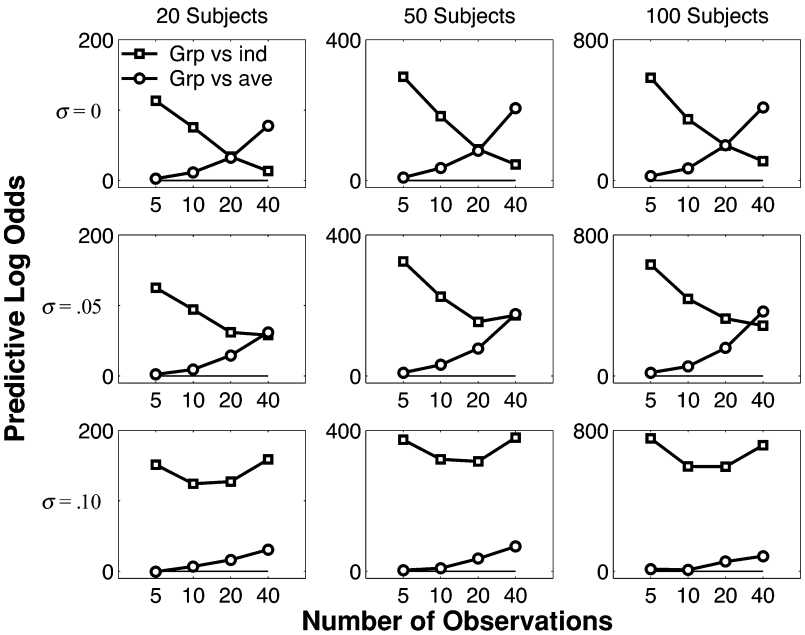

Figure 4. Posterior predictive log odds comparing the group approach with the individual and average approaches for different combinations of the numbers of observations (n = 5, 10, 20, and 40) and subjects (m = 20, 50, and 100), with variation in the probability of success for individuals within groups (σ = 0, .05, and .10).

three groups, but the randomly chosen means γ_1, γ_2, and γ_3 are very similar. These subjects are usefully understood, and their data will be well predicted, if the group model reduces to the average model and uses a single success parameter to model all subjects, even though this is not the "true" data-generating situation. Of course, it would be possible to place constraints on the way the different groups are defined, so that this sort of situation would not arise, but this would require making strong additional assumptions about the nature of individual differences in cognitive phenomena. In any case, additional constraints still would not overcome the inability of our statistical models to express the full range of regularities in real data. This is our justification for focusing on explanation and prediction in assessing the three modeling approaches, rather than on their abilities to recover artificially known data-generating processes.

Results

The results of the simulation study are shown in Figures 4 and 5. Figure 4 relates to prediction, with each graph showing the relative predictive log odds for the group versus individual and group versus average comparisons as the number of observations increases from $n = 5$ to $n = 40$. The graphs are arranged in rows corresponding to the group variance values, $\sigma = 0$, .05, and .10, and in columns corresponding to increasing numbers of subjects, $m = 20$, 50, and 100. Each data point shown is the average of 1,000 independent trials of the simulation study, and so represents performance for the range of possible "true" numbers of groups discussed earlier.

Figure 4 makes it clear that the predictive log odds always favor the group model, since they are always positive in both comparisons. More specifically, the relationship between the group and average models is easy to characterize. For all choices of m and σ, when few data are available the average model has predictive capabilities similar to those of the group model, but as more data become available the average model fares much worse. This makes sense, since few data are unlikely to reveal the individual differences that will be evident in larger data sets, meaning that the average model will become progressively less adequate as the sample size increases.

The relationship between the group and individual approaches is more complicated. When few data are available, the group approach is always clearly superior. This is because the individual approach over-fits the data, treating the stochastic variation in the observed counts as enduring regularities by employing separate parameters for each. When there is no variation in individual probabilities of success within groups (i.e., $\sigma = 0$), the predictive accuracy of the individual approach converges on that of the group approach as more data are observed. This is because large data samples reduce the stochastic variation that was being over-fit, and the individual approach mimics the group approach, with separate parameters for individuals converging on the same group value.

If there is individual variation in the probability of success, however, the individual approach again performs worse than the group approach, even when large numbers of data are available. This is because the noise introduced by $\sigma > 0$ is over-fit by the individual parameters, and they

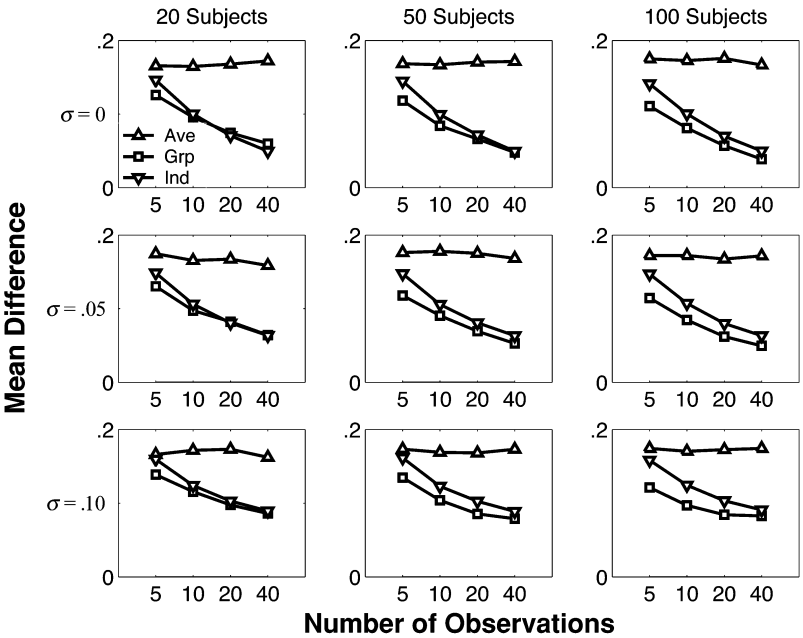

Figure 5. Mean absolute differences across subjects between known and estimated parameter values for different combinations of the numbers of observations (n = 5, 10, 20, and 40) and subjects (m = 20, 50, and 100), with variation in the probability of success for individuals within groups (σ = 0, .05, and .10).

no longer converge on the group values that offer the best predictive capability.

Figure 5 relates to understanding and shows the mean absolute difference across all subjects between the known and estimated parameter values. The average approach is always the least accurate and does not improve even as more data are available. The group and individual approaches are generally similar in accuracy, although with only a few very small exceptions, the group approach performs slightly better. In particular, the group approach seems better for small numbers of data, with both approaches improving with more data and becoming increasingly similar in accuracy.

Discussion

The results of the binary experiment simulation study demonstrate the clear superiority of the group approach. It always makes better predictions than the average and individual approaches. It is always much more accurate than the average approach in finding parameter values, and it is at least as accurate as the individual approach. Of course, these results depend on the three assumptions underlying the simulation study that were made explicit earlier. First, the superiority of the group approach over the average approach arises because of the assumed presence of individual differences in many of the trials. Second, the superiority of the group approach over the individual approach arises from assuming stochastic variation in the way data are generated, through both sampling and the underlying change in rate parameters. Third, the magnitude of these advantages depends on the numbers of subjects and data assumed to be typical of psychological experiments, although the qualitative trends would be the same under other assumptions.

Noting these correspondences shows that in a sense, the simulation studies just provide concrete confirmation of what should have been clear from the theoretical development of the group approach. The group approach was developed to deal with stochastic data-generating environments in which subsets of subjects have individual differences in basic parameters. If these features of the environment are assumed to be true, it should come as no surprise that the group approach performs best. The real advantage of the binary experiment evaluation is that it affords neat analytic solutions for maximum likelihood parameter estimation and Bayesian model selection for all three of the approaches, and so allows the differences in the approaches to be understood before more complicated models and real data are considered. Having established these differences, we now turn to demonstrating the application of the group approach to two practical problems.

AN APPLICATION TO CATEGORY LEARNING

Background

ALCOVE (Kruschke, 1992) 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 ex-

ternal feedback. The standard ALCOVE model 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* because of the nature of the stimuli) and two condensation tasks (called *condensation A* and *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 were psychologically meaningful individual differences in category learning.

There are two features of this application that make analysis more difficult than for the simulation study. First, the derivation of Bayes factors for families of ALCOVE models is not analytically tractable. This means that an approximate form of Bayesian model selection must be used. Second, optimizing the parameters of ALCOVE is not analytically tractable and is computationally costly using numerical methods. This means that identifying subsets of subjects for the group approach must be done approximately, using combinatorial optimization methods, since only a limited number of parameter optimizations are feasible for evaluating different partitions.

Model Selection

To develop a likelihood function for category learning, suppose that 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

$$p\left(d_{ij} \mid M, \theta_i\right) = \binom{b_j k_i}{d_{ij}} \gamma_{ij}^{d_{ij}} \left(1-\gamma_{ij}\right)^{b_j k_i - d_{ij}}.$$

The likelihood of a model family simply extends this result to consider every block of trials and every partition, so that

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

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, we use the Bayesian information criterion (BIC; Schwarz, 1978) as an approximate, easy-to-calculate means of Bayesian model selection. The BIC is given by

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

where P is the number of parameters in the model family (i.e., the sum of all of 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 of 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. As Kass and Raftery (1995) have noted, the "significance" of differences between BIC values can be assessed because they lie on a log-odds scale. Formally, for two models A and B with BIC values BIC_A and BIC_B, the approximation $2 \log [p(D \mid M_A) / p(D \mid M_B)] \approx \mathrm{BIC}_B - \mathrm{BIC}_A$ holds.

Inferring Partitions

Our original attempts to infer the partition with the lowest BIC from Kruschke's (1993) category learning data are detailed in Webb and Lee (2004). Originally, we relied on a two-stage heuristic that used singular value decomposition on the correlations between learning curves to find a low-dimensional representation of each subject's performance, and then we applied a version of k-means clustering to these representations to find clusters of subjects. A valid criticism of this approach is that it considers only one partition, which is found in a sensible but largely unprincipled way, and that it is entirely insensitive to the model being applied, since it operates solely on the raw data. These practices are inconsistent with the basic goals of developing cognitive models to act as simplified accounts of data that support prediction and generalization and provide meaning. The general approach to modeling individual differences developed in this article is most useful when groups of subjects are identified with respect to the low-dimensional parameterization of cognitive models that capture the constraints in empirical data.

Accordingly, we developed an improved method for inferring partitions that relies on an optimization process that closely relates the model to the data. In this method, for each possible number of groups, an initial partition-

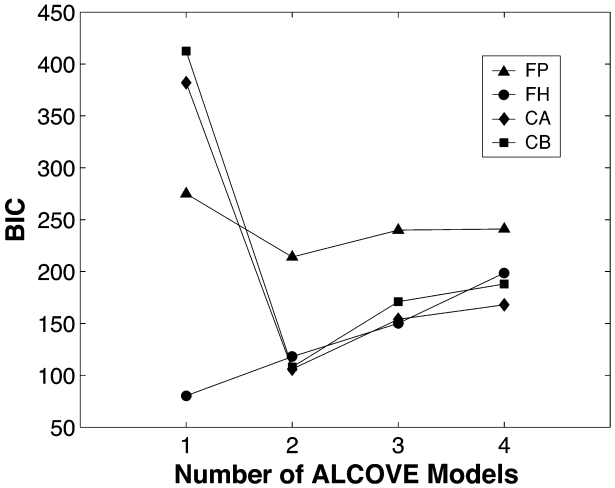

Figure 6. Pattern of change in BIC values for each clustering of the position-relevant filtration (FP), height-relevant filtration (FH), condensation A (CA), and condensation B (CB) category learning data.

ing of subjects is provided by the original heuristic method. A Nelder–Mead simplex algorithm (Nelder & Mead, 1965) is used to search for optimal parameterization of this initial partition, allowing its BIC to be evaluated. A combinatorial optimization process is then applied, based on subjects that are nearest (in the singular value decomposition representation) to the centroid of their neighboring group. All possible moves of these "nearest neighbors" into their adjacent group are considered, generating a list of candidate alternative partitions. For each of these alternatives, optimal parameterizations are also sought, allowing their BIC values to be calculated. Alternative partitions that improve the BIC are retained, and the process is repeated. Once no more nearest-neighbor moves lead to improvement, a partition that (locally) optimizes the BIC has been found and is retained as the best grouping of subjects.

Results

Figure 6 shows the results of the new method for inferring groups when applied to the four Kruschke (1993) tasks.[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, whereas the height-relevant filtration data are best modeled by considering all of the subjects as learning in the same way. These results are quantitatively extremely similar to those reported in Webb and Lee (2004), although the new method did improve the BIC slightly in a few cases. Qualitatively, the results are identical.

Figures 7 and 8 give more detailed results for, respectively, the position-relevant filtration and condensation 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 for the two-group model family that is preferred by the complexity analysis. These panels show the average accuracy for both groups of subjects separately, together with the maximum likelihood ALCOVE learning curve.

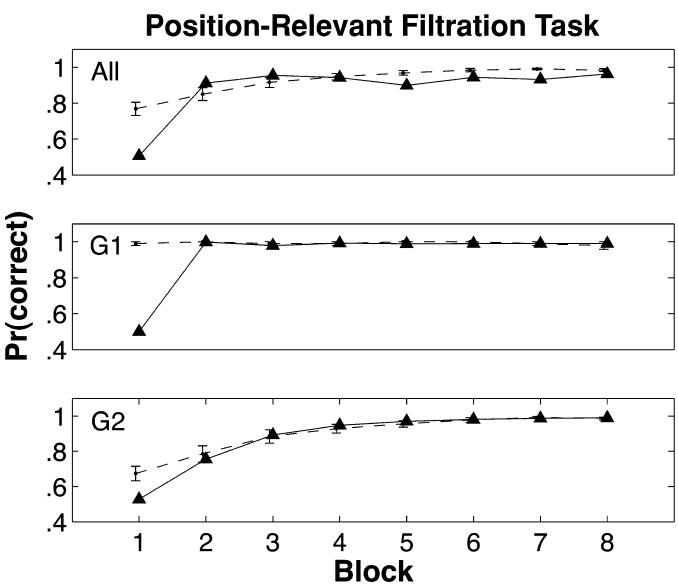

Figure 7. Change in accuracy across learning blocks for subjects (broken lines) and ALCOVE (solid lines) for the one-group (*All*) and two-group (G1 and G2) model families on the position-relevant filtration task. Error bars on the subject data represent one standard error in each direction.

Figure 7 shows that the moderate learning evident when the subjects are treated as having no individual differences is better modeled as coming from two distinct groups of subjects. Subjects in the first group maintain near-perfect accuracy throughout the category learning task, and subjects in the second group learn more gradually, achieving near-perfect accuracy only in the last few learning blocks. Figure 7 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.[3]

In a similar way, Figure 8 shows that the gradual increase in accuracy, evident when the subjects are treated as having no individual differences, is also better modeled as coming from two distinct groups of subjects. The first group exhibits almost no learning, and 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 has a greater λ_w value than the second group, which is consistent with their more rapid learning. For this task, both groups have high φ values, which are 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, in keeping with their inferior learning performance; in particular, the first group in this task, who basically failed to learn, have a very low φ value. Other comparisons of this type, both within and across tasks, generally have meaningful and useful interpretations and highlight the

Table 1
Maximum Likelihood Parameter Values for Each Group in the Model Family With the Lowest BIC Value for All of the Position-Relevant (FP) and Height-Relevant (FH) Filtration, Condensation A (CA), and Condensation B (CB) Category Learning Data

Task	Subject Group	λ_a	λ_w	c	φ
FP	G1	0.16	1.39	18.0	2.28
	G2	27.6	0.06	6.77	2.76
FH	All	0.58	0.23	1.56	1.00
CA	G1	1.14	0.47	2.53	0.27
	G2	0.38	0.24	7.52	0.93
CB	G1	0.34	0.24	0.80	0.40
	G2	0.07	0.16	3.59	1.14

Note—G1 and G2, Groups 1 and 2 in two-group model family; All, one-group model family.

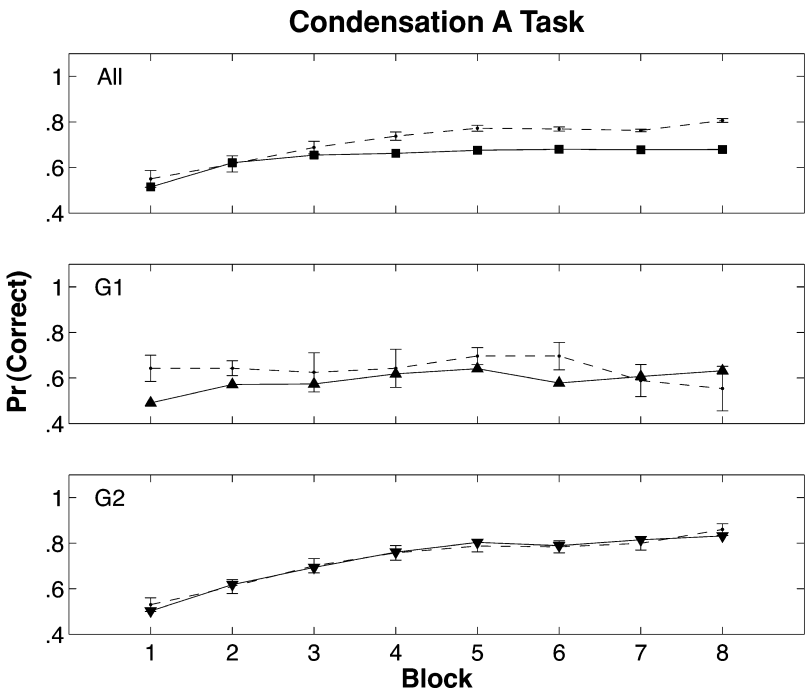

Figure 8. Change in accuracy across learning blocks for subjects (broken lines) and ALCOVE (solid lines) for the one-group (*All*) and two-group (G1 and G2) model families on the condensation A task. Error bars on the subject data represent one standard error in each direction.

ability of ALCOVE to represent psychologically important variations in category learning through its free parameters, once individual differences are considered.

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 strong evidence exists 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, have been 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 so by applying different psychologically meaningful parameterizations to accommodate variations in learning behavior. Although these points have been demonstrated previously for category learning models, including ALCOVE (see, e.g., Erickson, 1999; Lewandowsky, Kalish, & Griffiths, 2000; Nosofsky & Johansen, 2000; Nosofsky, Palmeri, & McKinley, 1994; Treat, McFall, Viken, & Kruschke, 2001; Yang & Lewandowsky, 2003), these other studies have not inferred the groupings by applying rigorous model selection criteria. What the results presented here demonstrate is that accounting for individual differences using model families learned from data has the potential to extend and increase the usefulness of existing cognitive models.

AN APPLICATION TO STIMULUS REPRESENTATION

Background

Helm (1959) collected dissimilarity data for $n = 10$ color stimuli from 14 subjects, 2 of whom repeated the experiment 4 weeks later. This gave a total of $m = 16$ symmetric dissimilarity matrices, $\mathbf{D}_1, \ldots, \mathbf{D}_m$, where $\mathbf{D}_k = [d_{ij}^k]$, with d_{ij}^k denoting the similarity between the ith and jth stimuli for the kth subject (or repeated subject).

Previous analyses (e.g., Borg & Groenen, 1997, pp. 359–370) of these data have considered multidimensional scaling representations, with a particular focus on the differences between the 5 matrices for known color-deficient subjects and the remaining 11 for color-normal subjects. The presence of these fundamental individual differences makes Helm's (1959) data interesting.

Group Multidimensional Scaling Representation

In multidimensional scaling (see, e.g., Cox & Cox, 1994; Shepard, 1987), stimuli are represented as points in a coordinate space, and their empirical dissimilarities are modeled by the distances between the points, usually according to one of the family of Minkowskian distance metrics. Applying our group approach simply means that collections of subjects use the same set of points to represent the stimuli. Formally, if the kth subject belongs to the gth group and this group represents the stimuli by points in an S_g-dimensional space, this subject's representation of the ith stimulus is the point $\mathbf{p}_i^g = (p_{i1}^g, \ldots, p_{iS_g}^g)$. This means that the empirical dissimilarity between the ith and jth stimuli for the kth subject, d_{ij}^k, is modeled by

$$\hat{d}_{ij}^k = \left[\sum_{s=1}^{S_g} \left| p_{is}^g - p_{js}^g \right|^r \right]^{1/r} + c^g,$$

where c^g is a nonnegative constant. The value of $r > 0$ determines the metric, with $r = 1$ (city block) and $r = 2$ (Euclidean) being common choices, corresponding to separable and integral stimuli, respectively (Garner, 1974; Shepard, 1991).

We follow Tenenbaum (1996; see also Lee, 2001; Lee & Pope, 2003) in assuming that the empirical dissimilarities follow Gaussian distributions with common variance σ^2. As has been argued by Lee (2001), the variance quantifies the precision of the data and plays an important role in determining the appropriate balance between fit and complexity. The repeated measures from 2 subjects in Helm's (1959) study provide exactly the sort of information that is needed to estimate this variance. Following Lee (2001, Equation 6), we calculated the within-subjects standard error for each dissimilarity comparison and averaged these to give an overall estimate of precision. These estimates were .0280 and .0206 for, respectively, the color-normal and the color-deficient subjects who provided repeated measures. Given the consistency of these estimates, we averaged them to produce a final estimate $\hat{\sigma} = .0243$ for the standard error of all dissimilarity judgments for all subjects.

Model Evaluation

The dissimilarity matrices $\mathbf{D}_1, \ldots, \mathbf{D}_m$, together with the $\hat{\sigma}$ estimate of their precision, constitute the data. A partition of the subjects into G groups selects a particular group model M_{grp} from the model family $\mathcal{M}_{\text{grp}}$. This model is parameterized by the points representing each group for each group, $\mathbf{p}_1^1, \ldots, \mathbf{p}_n^1, \ldots, \mathbf{p}_1^G, \ldots, \mathbf{p}_n^G$, together with the constants $c^1, \ldots, c^G$ and the metric parameter r. The likelihood function relating the data to the model is given by

$$p\left(\mathbf{D}_1,\ldots,\mathbf{D}_m,\hat{\sigma} \,\middle|\, \mathbf{p}_1^1,\ldots,\mathbf{p}_n^1,\ldots,\mathbf{p}_1^G,\ldots,\mathbf{p}_n^G,c^1,\ldots,c^G,r,M_{\text{grp}} \right) = \prod_g \prod_{i<j} \frac{1}{\left(\hat{\sigma}\sqrt{2\pi}\right)} \exp\left[-\frac{\left(d_{ij}^g - \hat{d}_{ij}^g\right)^2}{2\hat{\sigma}^2} \right].$$

Accordingly, up to a constant that does not depend on the model being considered, the (negative) log likelihood

function is simply the sum-squared error between each dissimilarity datum and its modeled value, scaled by the precision of the data, as follows:

$$\ln\left[p\left(\mathbf{D}_1,\ldots,\mathbf{D}_m,\hat{\sigma}\,|\,\mathbf{p}_1^1,\ldots,\mathbf{p}_n^1,\ldots,\mathbf{p}_1^G,\ldots,\mathbf{p}_n^G,c^1,\ldots,c^G,r,M_{\text{grp}}\right)\right] = \frac{1}{2\hat{\sigma}^2}\sum_g\sum_{i<j}\left(d_{ij}^g-\hat{d}_{ij}^g\right)^2+\text{constant}.$$

Since the data have m lots of $n(n-1)/2$ dissimilarities, the BIC is given by

$$\text{BIC}=\frac{1}{\hat{\sigma}^2}\sum_g\sum_{i<j}\left(d_{ij}^g-\hat{d}_{ij}^{g*}\right)^2+P\ln\frac{mn(n-1)}{2}+\text{constant},$$

where $\hat{d}_{ij}^{g*}$ now represents the best modeled value of the similarity between the ith and jth stimuli for subjects in group g, and P is the total number of coordinate location parameters. The representation of the gth group contributes about $n(S_g-1)+1$ parameters,[4] and P simply sums these parameter counts across all group representations.

Model Fitting

For a particular group model, where the partitioning of subjects into groups is known and a choice of metric r is made, finding multidimensional scaling representations is reasonably straightforward. For G groups, it involves G independent standard multidimensional scaling optimizations. For this, we used the obvious extension of the approach described in Lee (2001), using the BIC to determine the appropriate number of dimensions, but we fit the model to each subject in the group separately rather than to their averaged data. This is an important distinction, because it means that we did not average the data within groups, which would corrupt the data for the reasons described by Estes (1956). Rather, the same parameterization was applied to the raw data of every subject in the same group.[5]

Formally, for the gth group we found the best points to represent each stimulus and the additive constant

$$\left(\mathbf{p}_1^{g*},\ldots,\mathbf{p}_n^{g*},c^{g*}\right)=\underset{\left(\mathbf{p}_1^g,\ldots,\mathbf{p}_n^g,c^g\right)}{\text{argmax}}\sum_{\substack{k\in g\text{th}\\ \text{group}}}\sum_{i<j}\left(d_{ij}^k-\hat{d}_{ij}^g\right)^2$$

using a Levenberg–Marquardt approach to continuous optimization (see, e.g., More, 1977). This was done separately for dimensionalities $S_g=1,2,\ldots,$ up to a maximum chosen to be sufficiently large to ensure that the best dimensionality according to

$$\text{BIC}_g=\frac{1}{\hat{\sigma}^2}\sum_{\substack{k\in g\text{th}\\ \text{group}}}\sum_{i<j}\left(d_{ij}^k-\hat{d}_{ij}^{g*}\right)^2+\left[S_g(n-1)+1\right]\ln\frac{yn(n-1)}{2}$$

has been found, where y is the number of subjects in the gth group.

For the combinatorial optimization problem of finding the best partitioning of subjects into groups, we used a simple greedy-search algorithm that is similar in spirit to the method we used in the category learning application. In essence, it starts with a "seed" partition, finds multidimensional scaling representation for each group using the approach described above, and sums their BIC values. It then moves a subject, chosen at random, from one partition to another that is also chosen at random. Representations are found for this new partition, and the overall BIC is evaluated. If the change leads to an improvement (i.e., a decrease) in the BIC, it is accepted; otherwise, it is rejected. This process continues until a large fixed number of changes have been tried unsuccessfully. We applied this algorithm multiple times, using different seeding partitions, and chose the final group model to be the one with the best BIC of any of the results.

Results

In fitting the group model to Helm's (1959) data, we made the metric assumption $r=2$, because color is usually regarded as an integral stimulus domain. The results under this assumption are summarized in Figure 9. The top left panel shows the change in the best BIC value under the assumptions that there are one, two, or three different groups of subjects. It is clear that a two-group model is preferred, since it provides a much better account of the data than do one-group models, which assume no individual differences. To explore this finding,

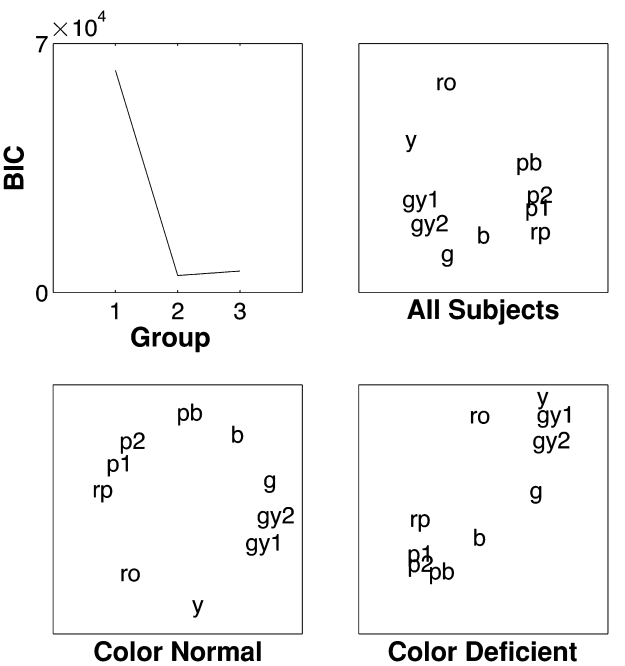

Figure 9. The results of applying the group approach to Helm's (1959) color data. The top left panel shows the BIC for different numbers of subject groups. The top right panel shows the best one-group multidimensional scaling representation. The bottom left and bottom right panels show the best two-group multidimensional scaling representations. The color stimuli have the abbreviated labels "rp" = reddish purple, "ro" = reddish orange, "y" = yellow, "gy1" = greenish yellow 1, "gy2" = greenish yellow 2, "g" = green, "b" = blue, "pb" = purple-blue, "p1" = purple 1, and "p2" = purple 2.

the top right panel shows the best multidimensional scaling representation found under the one-group assumption. This representation resembles the color "circle" or "horseshoe" that has repeatedly been found in multidimensional scaling analyses of color stimuli (see, e.g., Shepard, 1980) but appears to be degraded.

The bottom left and bottom right panels of Figure 9 show the representations for the subject groups in the best two-group model. This best partitioning corresponds exactly to the known distinction between color-normal and color-deficient subjects. The color-normal representation shows a color circle without the degradation evident in the one-group representation. The color-deficient representation also shows a color circle, but with significant distortion corresponding to a reduction in the red–green axis, consistent with deuteranopy.

Discussion

In light of the two-group representations, the source of the degradation in the best multidimensional scaling representation of the aggregated data is clear: There are two groups of subjects, with large and meaningful individual differences warranting separate stimulus representations. The group approach is able to identify these differences in an automated way, using complexity-sensitive model selection to determine the appropriate number of groups, optimize the assignment of individuals to these groups, and find the best multidimensional scaling representations for each group.

It is worth emphasizing that these capabilities extend well beyond those of alternative "individual differences" extensions of multidimensional scaling, such as INDSCAL (see, e.g., Carroll & Chang, 1970) and INDCLUS (e.g., Carroll & Arabie, 1983). These models accommodate individual variation by allowing each subject to weight the axes of an underlying representational space in different ways. They do not model groups of subjects, although it is possible to use the individual weights in a heuristic way to address this possibility. More fundamentally, they do not allow for the possibility that different groups of subjects might use fundamentally different representations. The present approach allows for different groups to use representations that are not related to one another by simple weighting transformations, and it is able to accommodate the case of different groups using spaces with different dimensionalities. If, for example, the color-deficient subjects in Helm's (1959) study had been severely enough impaired to eliminate the red–green distinction, our approach would have been able to find a one-dimensional representation for that group of subjects, while retaining the two-dimensional representation for the color-normal subjects.

On the other hand, the comparison with INDSCAL-type methods highlights a general limitation of the group approach as it currently stands. In the individual differences modeling presented here, every parameter of a cognitive model is learned for each group of subjects. When the differences in behavior between groups of subjects require each parameter to be changed, this flexibility is necessary. It seems quite possible, however, that there will be situations in which the differences between groups of subjects relate only to a subset of the parameters. For example, if groups of subjects have spatial representations that differ only through simple weighted transformations, an INDSCAL model will be more parsimonious, because it will not need to respecify coordinate locations for every group. Of course, there is nothing in our conceptual framework preventing the consideration of individual differences models in which different groups vary only in subsets of their parameters. Reliance on sophisticated model selection methods could ensure that the relative simplicity of these more constrained accounts would be detected and would be preferred when appropriate. The only extension of the present approach that would be required would be the ability to consider a richer class of model families that would allow for variation in subsets of parameters across groups. This promises to be a fruitful area for future research.

GENERAL DISCUSSION

The group-based modeling approach we have presented here is designed to account for individual differences in cognition. There are alternative approaches that accommodate individual differences by specifying distributions of basic model parameters and then learning the "hyperparameters" of these distributions from data (see, e.g., Peruggia, Van Zandt, & Chen, 2002; Rouder & Lu, 2005; Rouder, Sun, Speckman, Lu, & Zhou, 2003). These models are often developed within a hierarchical Bayesian modeling framework. This has proved to be an informative and useful approach and is likely to be strengthened and extended by current research activity. We view our approach as complementary. Rather than modeling individual differences as smooth (typically unimodal) variations in basic parameters, we are interested in cases in which different groups of subjects use fundamentally different basic parameter values, so it makes sense to partition the parameter space. It could be argued that these sorts of individual differences are more basic or important than those that involve minor parametric variation. In any case, we believe that the category learning and color perception analyses presented here show the usefulness of our approach.

Ultimately, of course, both approaches could be reconciled by using sufficiently flexible distributional forms in a hierarchical approach. At a conceptual level, we demonstrate how this could be done in Figure 10. The bottom-most panel shows some hypothetical data measuring the decrease in some aspect of cognitive performance over time. Each data curve corresponds to a single subject, and there are many subjects with similar performance whose performance deteriorates slowly over time. Suppose we consider individual differences

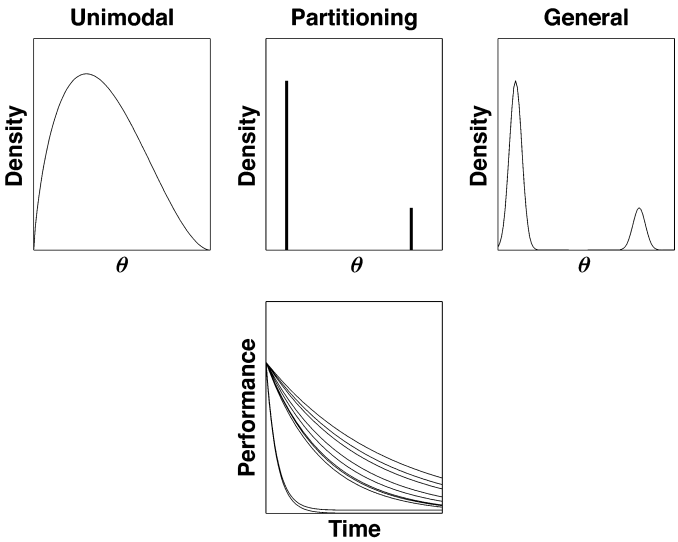

Figure 10. Three classes of distributions over model parameters for modeling individual differences in a hierarchical Bayesian framework.

by applying a simple cognitive model with a single decay rate parameter θ, so that larger values of θ correspond to more rapid deterioration in performance.

The top left panel of Figure 10 indicates the results of assuming a unimodal distribution over this parameter. Most of the density is allocated to the small parameter values, because most subjects show slow deterioration. The distribution also extends to larger values of θ, however, because some subjects show much more rapid deterioration. The problem with this representation is that it fails to capture the obvious between-groups individual difference in human performance, because it cannot use multimodal distributions to describe the variation in the parameter across individuals. The top center panel shows the results of partitioning subjects into two groups, in the way considered here. The two partitions correspond to a sharply peaked multimodal distribution for the parameters. This approach does capture the between-groups difference, which seems to be the most important regularity in the data. It fails, however, to capture the second-order effects of individual variation within the groups. The top right panel characterizes the general approach that combines the best features of the unimodal and partitioning approaches. It allows for multimodal distributions, capturing the "major" between-groups variation as well as the "minor" within-groups variation. Developing this general approach within a hierarchical Bayesian framework is a priority for future research.

One of the weaknesses of the category learning and color perception analyses presented here is the reliance on the BIC to compare different competing individual differences models. Although 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 when 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), and these models have 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 (Rissanen, 1996; see also Myung, Balasubramanian, & Pitt, 2000) and normalized maximum likelihood (Rissanen, 2001). For cognitive models that resist the formal analysis needed to derive these measures, an alternative is to use numerical methods, such as Markov chain Monte Carlo (see, e.g., Gilks, Richardson, & Spiegelhalter, 1996) to approximate the Bayesian quantities that compare model families.

A final intriguing possibility for future research, and a natural extension of the approach presented here, involves using fundamentally different models to capture between-subjects variation rather than relying solely on different parameters within the same basic model. In category learning, for example, it may make sense to model some subject groups using ALCOVE or its descendants but to apply a very different category learning model to others, such as the fast and frugal account provided by categorization by elimination (Berretty et al., 1999). For stimulus representation, some groups of subjects could be modeled using a featural representation and others with a dimensional representation. Other cognitive modeling areas, not considered here, offer similar possibilities. In memory retention, for example, one group of

subjects could be modeled using a power function, whereas another group could be modeled using an exponential decay function.

If tackled successfully, these sorts of extensions to the modeling approach presented here would lead to a very powerful approach for modeling individual differences in cognition. Our group approach to cognitive modeling is a more general one than approaches that average or aggregate data, and thus assume that there are no individual differences. Ours is a more succinct approach than those that use subject-by-subject analysis, and it offers advantages in terms of the key modeling goals of explanation and prediction. Our evaluation of the group approach on the simulated binary experiment provides clear evidence of its ability to explain important cognitive parameters and predict cognitive performance. Our practical demonstrations, using multiple ALCOVE models to capture differences in category learning and multiple MDS representations to capture individual differences in color perception, both provide concrete examples of its usefulness. They show 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.

REFERENCES

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

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**, 144-151.

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

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination: Using few clues to choose. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 235-254). New York: Oxford University Press.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Carroll, J. D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika*, **48**, 157-169.

Carroll, J. D., & Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart–Young" decomposition. *Psychometrika*, **35**, 283-319.

Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman & Hall.

Erickson, M. A. (1999). Rules and exemplars in category learning (Doctoral dissertation, Indiana University, Bloomington, 1999). *Dissertation Abstracts International*, **60**, 2377B.

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

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

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculation. *Journal of the Royal Statistical Society: Series B*, **56**, 501-514.

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

Grünwald, P. D. (1998). *The minimum description length principle and reasoning under uncertainty*. Amsterdam: University of Amsterdam, Institute for Logic, Language & Computation.

Helm, C. E. (1959). *A multidimensional ratio scaling analysis of color relations*. Princeton, NJ: Princeton University & Educational Testing Service.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. New York: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 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, & Cognition*, **25**, 1083-1119.

Laming, D. (1992). Analysis of short-term retention: Models for Brown–Peterson experiments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1342-1365.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, **45**, 149-166.

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, **47**, 32-46.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1666-1684.

More, J. J. (1977). The Levenberg–Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical analysis: Proceedings of the biennial conference held at Dundee, June 28–July 21* (Lecture Notes in Mathematics, Vol. 630, pp. 105-116). Berlin: Springer.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, **97**, 11170-11175.

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**, 832-840.

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

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

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

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, **7**, 375-402.

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

Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In C. Gatsonis, R. E. Kass, A. Carriquiry, A. Gelman, D. Higdon, D. K. Pauler, & I. Verdinelli (Eds.), *Case studies in Bayesian statistics* (Vol. 6, pp. 319-334). New York: Springer.

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

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 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**, 1712-1717.

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.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, **68**, 589-606.

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

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

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

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

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. R. Pomerantz & G. L. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53-71). Washington, DC: American Psychological Association.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 3-9). Cambridge, MA: MIT Press.

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* (Vol. 11, pp. 59-65). Cambridge, MA: MIT Press.

Treat, T. A., McFall, R., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, **13**, 549-565.

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

Webb, M. R., & Lee, M. D. (2004). Modeling individual differences in category learning. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1440-1445). Mahwah, NJ: Erlbaum.

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

Yang, L., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 663-679.

NOTES

1. We thank Robert Nosofsky for suggesting this framework for demonstrating our theoretical ideas.

2. One of the attractions of Kruschke's (1993) study is that the same parameterization was used to account for all four tasks, demonstrating that differences in the category structures accounted for much of the variation in human performance. We chose to consider each task separately because this approach provides a clearer demonstration of the individual differences we are attempting to model.

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

4. The number of free parameters in a multidimensional scaling representation actually depends on the nature of the metric space and its invariances. For the metric space possibilities we are considering, however, exact specifications of the number of parameters would only make negligible differences in comparison with the inaccuracy inherent in the BIC approximation itself. The choice $n(S_g - 1) + 1$ is a "worst case" value that gives an upper limit on the number of free parameters, consistent with the conservatism of the BIC.

5. This approach is a little different from the one used in the simulation study and the category learning example, where data from subjects in the same group were aggregated and modeled using the same parameterization. It would be possible, in these earlier cases, to consider an alternative approach in which the data for each subject in a group were modeled separately using the same parameterization.

(Manuscript received July 24, 2003;
revision accepted for publication February 8, 2005.)