Journal of Mathematical Psychology 59 (2014) 132–150

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning☆

Annelies Bartlema[a,*,1], Michael Lee[b], Ruud Wetzels[c,d], Wolf Vanpaemel[a]

[a] KU Leuven, University of Leuven, Belgium
[b] University of California, Irvine, United States
[c] Informatics Institute, University of Amsterdam, Netherlands
[d] Spinoza Centre for Neuroimaging, Amsterdam, Netherlands

HIGHLIGHTS

- Bayesian hierarchical mixture methods are used to model individual differences.
- We demonstrate this method in two example applications in the domain of category learning.
- These analyses lead to different conclusions than analyses on grouped data.
- Different groups of people and variation within each group when learning a category.
- Hierarchical mixture models suitable for parameter estimation and model selection.

ARTICLE INFO

Article history:
Available online 17 January 2014

Keywords:
Individual differences
Bayesian method
Hierarchical mixture model
Model selection
Parameter estimation
Category learning

ABSTRACT

We demonstrate the potential of using a Bayesian hierarchical mixture approach to model individual differences in cognition. Mixture components can be used to identify latent groups of subjects who use different cognitive processes, while hierarchical distributions can be used to capture more minor variation within each group. We apply Bayesian hierarchical mixture methods in two illustrative applications involving category learning. One focuses on a problem that is typically conceived of as a problem of parameter estimation, while the other focuses on a problem that is traditionally tackled from a model selection perspective. Using both previously published and newly collected data, we demonstrate the flexibility and wide applicability of the hierarchical mixture approach to modeling individual differences.
© 2013 Elsevier Inc. All rights reserved.

1. Introduction

William K. Estes was a pioneer in using formal modeling approaches to build an understanding of the fundamental properties of human cognition. He made major theoretical contributions to understanding many cognitive capabilities, including the cornerstones of memory, learning, categorization and decision-making. Pioneering a formal mathematical and statistical approach to understanding cognition raises methodological as well as theoretical challenges. Here, Estes (1956) provided one of the first and clearest warnings on the dangers of modeling aggregated human behavioral data, without respect for possible individual differences:

> Just as any mean score for a group of organisms could have arisen from sampling any of an infinite variety of populations of scores, so also could any given mean curve have arisen from any of an infinite variety of populations of individual curves. Therefore no 'inductive' inference from mean curve to individual curve is possible, and the uncritical use of mean curves even for such purposes as determining the effect of an experimental treatment upon rate of learning or rate of extinction is attended by considerable risk … we can no longer expect averaged data to yield any direct answer to the question, 'What is the form of the individual function?' (Estes, 1956, pp. 134–135).

Estes explains that, except under some exceptional circumstances, the form of the individual functions or the distribution of individual parameter values cannot be inferred from aggregated data. He

☆ Data and code are available on the following URL http://faculty.sites.uci.edu/mdlee/BartlemaEtAl2014.zip.

* Correspondence to: Annelies Bartlema, Faculty of Psychology and Educational Sciences, Tiensestraat 102, bus 3713, 3000 Leuven, Belgium.
E-mail address: annelies.bartlema@ppw.kuleuven.be.

[1] Annelies Bartlema is a Doctoral Research Fellow with the Research Foundation-Flanders (FWO).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.12.002

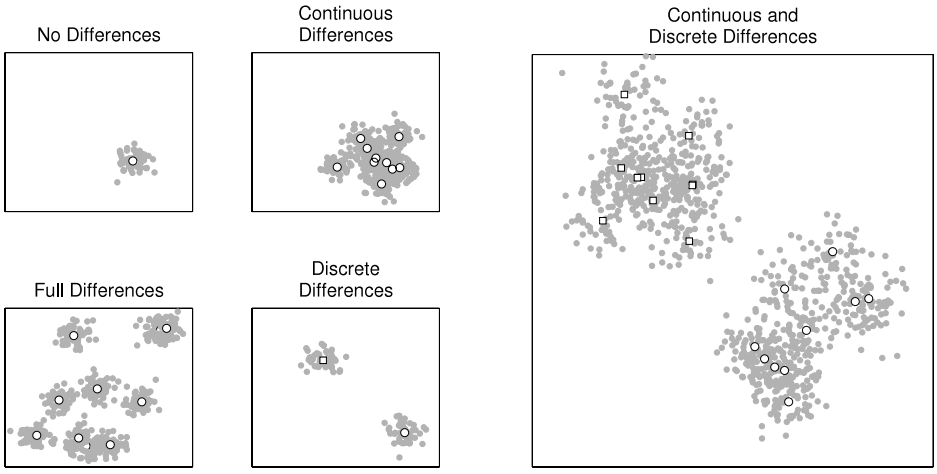


Fig. 1. Five different modeling assumptions about individual differences.

returned to this theme towards the end of his career (Estes & Maddox, 2005), reminding the field that the search for invariants in cognition can only succeed when meaningful variations like individual differences are acknowledged.

Thanks to Estes' clear warnings, modelers of human cognition have used a number of approaches that address the issue of individual differences. Sometimes this is done by focusing on individual data (Cohen, Sanborn, & Shiffrin, 2008), sometimes clustering approaches are used to identify groups of subjects (Lee & Webb, 2005), and sometimes hierarchical random-effect models have been used to allow continuous variation between subjects (Rouder & Lu, 2005; Shiffrin, Lee, Wagenmakers, & Kim, 2008). In this paper, our goal is to demonstrate the usefulness of a hierarchical mixture approach as a more general framework for dealing with individual differences.

The structure of the paper is as follows. We first provide a general framework for the representation of individual differences. In particular, we propose to model both of what we call discrete and continuous individual differences – corresponding to major differences in cognitive processes, as well as more minor parametric variation within a process – using hierarchical latent mixture models. Next, we summarize the graphical model approach we take throughout. In the following two sections, we illustrate in two example applications how the Bayesian hierarchical mixture approach we advocate can be used to understand whether people are different, and how these differences can be modeled. The first example application tackles an issue that is typically seen as a problem of parameter estimation, whereas the second example application is concerned with model selection. In both example applications, a hierarchical mixture approach reveals clear individual differences, and thus provides a more accurate and complete understanding than can be gleaned from an analysis based on aggregate data. Overall, our two applications demonstrate how hierarchical latent mixture models, when coupled with the generality and power of Bayesian inference, provide a useful means of addressing Estes' warnings about the perils of aggregated data.

Both of our applications involve human category learning, a field in which Estes has made several important contributions (e.g., Estes, 1986a,b). In particular, the first application deals with individual differences in selective attention, whereby people learn to focus on those aspects of stimuli most useful for learning categories. The second application deals with the nature of abstraction in the representation of categories. As our goal is to demonstrate a general approach for tackling individual differences using hierarchical latent mixture models, the example applications are not intended to constitute major advances to the existing knowledge of individual differences in category learning – which are well established (Johansen & Palmeri, 2002; Lee & Webb, 2005; Navarro, Griffiths, Steyvers, & Lee, 2006; Shin & Nosofsky, 1995; Smith & Minda, 2002) – nor to make strong theoretical claims about selective attention and representational abstraction.

2. Modeling individual differences

Fig. 1 shows five schematic parameter spaces, corresponding to different basic assumptions that can be made about individual differences. The top-left panel corresponds to the assumption there are no individual differences. There is a single true point, represented by the circular marker, corresponding to the one parameterization of the cognitive process that is common to all people. The gray region show what inferences can be made about that point from the finite behavioral data that might actually be observed in an experiment. This assumption of no individual differences is made, whether intentionally or not when relying upon averaged or aggregated data.

The bottom-left panel corresponds to the assumption that every individual is different, and that they are all independent of one another. There is now a single true point for each person, and no structure in the relationship between these points. This assumption of full individual differences corresponds to the case where a model is fit separately to each subject in an experiment.

The top-center panel corresponds to the assumption that there are individual differences between people in the parameterization of the cognitive process they use. These individual differences are continuous, in the sense that they vary smoothly around some central tendency. 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.

The bottom-center panel also corresponds to the assumption that there are individual differences between people, but these differences are more fundamental. There are now two discrete types of true points, shown by square and circular markers, corresponding to two different parameterizations. These types may be very different parameterization of the same cognitive process, or represent two qualitatively different cognitive processes that different people use. These sorts of individual differences can be accommodated by mixture models, in which the different mixture components correspond to the different types of processes or strategies people are assumed to use.

Finally, the large right-hand panel of Fig. 1 combines the continuous and discrete individual differences, to provide the most general approach. There are now both different discrete types

of parameterizations, and constrained individual variation within each type. This account of individual differences can be accommodated using a combination of hierarchical and mixture modeling. The mixture component identifies the fundamentally different cognitive processes, as indicated by the squares and circles, and the hierarchical component captures the smooth variation within each process. Our applications consider this most general approach to individual differences.

3. Graphical model

Throughout our demonstration of how hierarchical mixture models allow for both continuous and discrete individual differences, we use the Bayesian framework for implementing models and making statistical inferences. The Bayesian approach not only provides principled, complete and coherent statistical inference (Gelman, Carlin, Stern, & Rubin, 1995), but also facilitates the use of flexible and general cognitive models (Lee, 2008, 2011a).

Our Bayesian implementation of the models used in the two applications is achieved using the formalism provided by graphical models.[2] A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. The nodes represent variables of interest, and the graph structure indicates dependencies between the variables, with children depending on their parents. In drawing graphical models, we use the following notational conventions: Unobserved variables are represented without shading and observed variables with shading; Continuous variables are represented with circular nodes and discrete variables with square nodes; Stochastic variables are represented using single borders, and deterministic unobserved variables with double borders. Plates enclose with square boundaries those subsets of the graph that are independently replicated in the model.

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 (Lunn, Thomas, Best, & Spiegelhalter, 2000), which is easy-to-learn software for implementing and analyzing graphical models (see Kruschke, 2011; Lee & Wagenmakers, 2013). Details and tutorials aimed at cognitive scientists are provided by Lee (2008) and Shiffrin et al. (2008).

4. First application: Individual differences in selective attention

The first example application of how the Bayesian hierarchical mixture approach can be used to account for individual differences focuses on a problem that is typically conceived of as a problem of parameter estimation. In particular, we focus on the selective attention parameter in the seminal Generalized Context Model (GCM; Nosofsky, 1986), a model that explains how people learn categories based on feedback.

Selective attention is one of the most compelling theoretical ideas in the study of category learning. When learning a category structure, people tend to attend selectively to those dimensions of the stimuli that are relevant to distinguishing the categories. Nosofsky (1984) showed that, for stimuli represented in terms of underlying continuous dimensions, selective attention could help explain previously puzzling empirical regularities in how easily people learn different category structures (Shepard, Hovland, & Jenkins, 1961). The idea that people selectively attend stimulus dimensions when learning categories is a key assumption of the GCM, which has proven very successful in accounting for human category learning behavior.[3]

The goal of the application is to demonstrate the Bayesian hierarchical mixture approach for modeling individual differences in selective attention.[4] To explore this possibility, we re-analyze human performance on a task conducted by Kruschke (1993), using the GCM in two different ways. In a first analysis, we assume there are no individual differences and consider the aggregate data, while in the second analysis, we use a hierarchical mixture extension of the GCM to allow for individual differences.

4.1. Data

The data we use in the first application come from Kruschke (1993), who studied the ability of ALCOVE (Kruschke, 1992) 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 into two groups of four stimuli in four different ways, as shown in Fig. 2. Kruschke (1993) collected data from a total of 160 subjects, in a between-subject design, with 40 attempting to learn each of the four category structures. The task for each subject 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.

The category structure that is our main focus is the so-called "Condensation B" structure, which is shown in the bottom-right panel of Fig. 2.[5] The eight stimuli are arranged by their heights and positions, and the four below and to the right of the dividing line belong to category A. The stimuli are numbered 1–8 in the figure, for ease of reference later when modeling results are presented.

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 y_{ik}, the number of times the ith stimulus was categorized as belonging to category A by the kth subject, out of the $t = 8$ trials on which it was presented.

4.2. Analysis assuming no individual differences

In an analysis that does not consider individual differences, the behavioral data can be further summarized as $y_i = \sum_k y_{ik}$, the total number of times all subjects classified the ith stimulus into category A, out of $t = 40 \times 8$ total presentations.

4.2.1. The GCM

The GCM assumes that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. The psychological coordinates for the stimuli are taken from Kruschke (1993), who derived these coordinates from pairwise similarity ratings. Since there are only

[2] Note that this does *not* mean we are proposing "Bayesian" or "rational" versions of the models considered in the applications. We are simply using Bayesian statistics, rather than traditional model-fitting methods and frequentist statistical approaches, to make inferences from data. That is, we are using Bayesian inference as statisticians do, and as psychologists should do, to relate models to data (Kruschke, 2010; Lee, 2011b).

[3] The GCM is very closely related to one of the models Estes proposed in his family of array models (Estes, 1986a). We focus on the GCM rather than the array model because, unlike the GCM, the array model does not include selective attention parameters (or, put differently, it assumes all of the features of an exemplar receive equal attention).

[4] An earlier version of this application, using a different model, but adopting a similar approach to partly the same data, was presented in Lee and Wetzels (2010).

[5] We will come back to the remaining three category structures in the discussion.

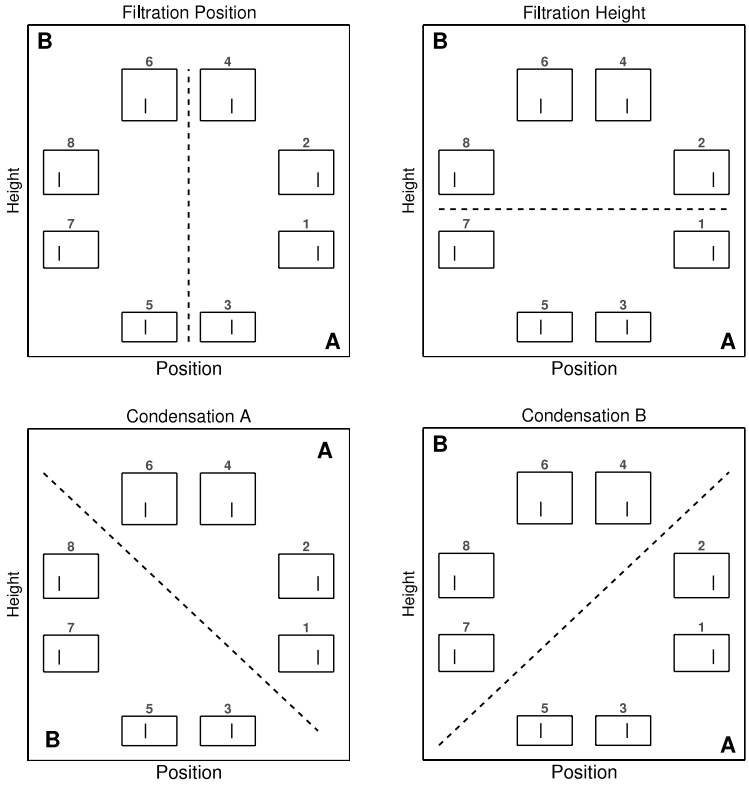


Fig. 2. The four category structures from Kruschke (1993).

two dimensions, the ith stimulus is represented by the point $\mathbf{p}_i = (p_{i1}, p_{i2})$. The GCM further assumes classification decisions are based on similarity comparisons with the stored exemplars, with similarity determined as a nonlinearly decreasing function of distance in the attention-weighted psychological space. The first dimension has an attention weight, w with $0 \leq w \leq 1$, and the second dimension then has an attention weight $1 - w$. These weights act to 'stretch' attended dimensions, and 'shrink' unattended ones. Formally, the similarity between the ith and jth stimuli is $s_{ij} = \exp\{-c(w d_{ij}^{[1]} + (1-w) d_{ij}^{[2]})\}$, where c is a generalization parameter and $d_{ij}^{[m]} = |p_{im} - p_{jm}|$ is the distance between the ith and jth stimuli on the mth dimension.

According to the GCM, categories are represented by collections of 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 s_{iA} is $\sum_{j \in A} s_{ij}$, where the summation is based on assignment variables, a_j, that take values 0 and 1 to indicate whether or not the jth stimulus belongs to category A or to category B. Finally, categorization response decisions are based on the exponentiated Luce Choice rule (Luce, 1959), as applied to the overall similarities. The probability that the ith stimulus will be classified as belonging to category A, rather than category B, is modeled as $r_i = b s_{iA}^{\gamma} / \left(b s_{iA}^{\gamma} + (1-b)\, s_{iB}^{\gamma}\right)$, where b is the response bias towards category A, and γ is the response-scaling parameter that reflects the amount of determinism in responding (Ashby & Maddox, 1993; Navarro, 2007). Throughout the paper we focus on the basic version of the GCM, meaning that b and γ are not treated as free parameters, but fixed to $1/2$ and 1 respectively. It would be straightforward, however, in an extended analysis, to place prior distributions on these model parameters, and make inferences about them from data using the models we develop. The observed decision data themselves are then modeled as $y_i \sim \text{Binomial}(r_i, t)$, meaning that each of the t presentations of the ith stimulus has a probability r_i of being categorized as belonging to category A.

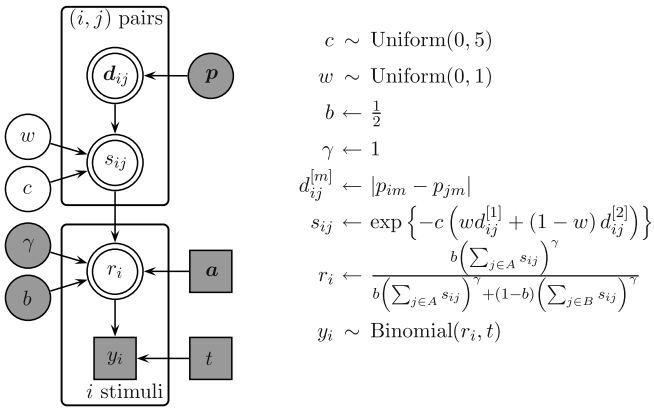


Fig. 3. Graphical model implementation of the GCM analysis without individual differences.

In our application, we will rely on the basic version of the GCM, meaning that b and γ are not treated as free parameters, but fixed to $1/2$ and 1 respectively. For the two remaining parameters c and w we assume vague priors to reflect the absence of strong theoretical expectations about attention allocation and generalization in this design (but see Vanpaemel & Lee, 2012).

A graphical model implementation of the GCM is shown in Fig. 3 (Lee, 2008; Lee & Wagenmakers, 2013; Vanpaemel, 2009). The stimulus coordinates $\mathbf{p}$ generate the pairwise distances on each dimension $d_{ij}^{[m]}$. The attention w and generalization c parameters combine with these distances to generate the pairwise similarities s_{ij}. These similarities, combined with the indicator variable $\mathbf{a}$, the bias parameter and the response-determinism parameter, in turn, lead to response probabilities r_i which generate the observed data y_i.

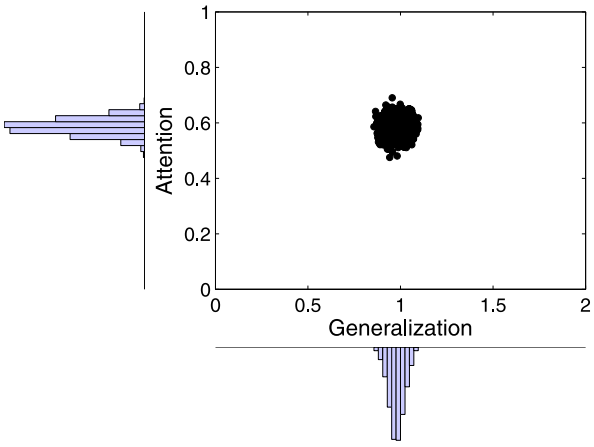


Fig. 4. Joint and marginal posterior distributions over attention w and generalization c parameters of the GCM, when applied to the aggregated data from the "Condensation B" condition.

4.2.2. Results

Our results are based on 3 chains of 10,000 samples each, with a burn-in of 1000 samples, whose convergence was checked using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997). The key result is shown in Fig. 4, which plots the joint posterior distribution of the generalization and attention parameters, as a cloud of dots, as well as their marginal posterior distributions, as histograms. The marginal posterior for the attention parameter w – which gives the weight for the position dimension – lies between about 0.53 and 0.64. This result can be interpreted as showing that people give significant attention to both dimensions, although they are probably focusing a little more on line position than on rectangle height. For this condensation category structure, both stimulus dimensions are relevant to determining how stimuli belong to categories, and so the shared attention result makes sense. Overall, analyzing grouped data produces what could be regarded as a psychologically reasonable inference about selective attention, consistent with previous theorizing.

4.3. Analysis assuming individual differences

While the analysis of aggregate data produces an intuitively reasonable result, it does assume that all of the subjects used exactly the same parameterization of the GCM to guide their category learning. That is, it makes the "No Differences" assumption schematically displayed in the top-left panel of Fig. 1. An inspection of the raw behavioral data, however, suggest that this might be an unrealistic assumption. Fig. 5 shows the proportion of times each subject made A or B classification decisions for each stimulus. These summary plots of individual subject categorization decisions suggest a large degree of variation between subjects, and raises the possibility that there are psychologically meaningful individual differences.

In particular, Fig. 5 suggests there are at least three different types of categorization behavior. The first type, exemplified by for example subject 1, corresponds to subjects who categorize each stimulus roughly equally often in each category. This sort of behavior is consistent with guessing behavior, or some other decision-making process that does not learn the categories effectively. These subjects may be viewed as contaminants who did not learn the category structure during the experiment.

A distinct pattern is observed for subject 9. This subject makes almost all his mis-classifications on stimuli 4 and 5. This is consistent with a focus on the position dimension: if subjects attend selectively to the position of the interior line, their decision boundary is parallel to the height dimension and they will incorrectly categorize those stimuli 4 and 5. This corresponds to large values of the attention parameter w. Another pattern is visible for other subjects, such as subject 33. This subject categorizes stimuli 2 and 7 poorly, suggesting a decision boundary parallel to the position dimension. This is consistent with a focus on the height dimension, corresponding to low values of the attention parameter w.

Based on these sorts of observations, our model with structured individual differences assumes three latent groups of subjects with different categorization behavior. A simple way to account for the behavior of the subjects who are guessing, or otherwise not learning, is not to use the GCM, but to use a contaminant model (Zeigenfuse & Lee, 2010) that sets their response probabilities to be $r_{ik} = 0.5$ for all stimuli. We account for the two other groups by allowing for two different parameterizations of the GCM, with

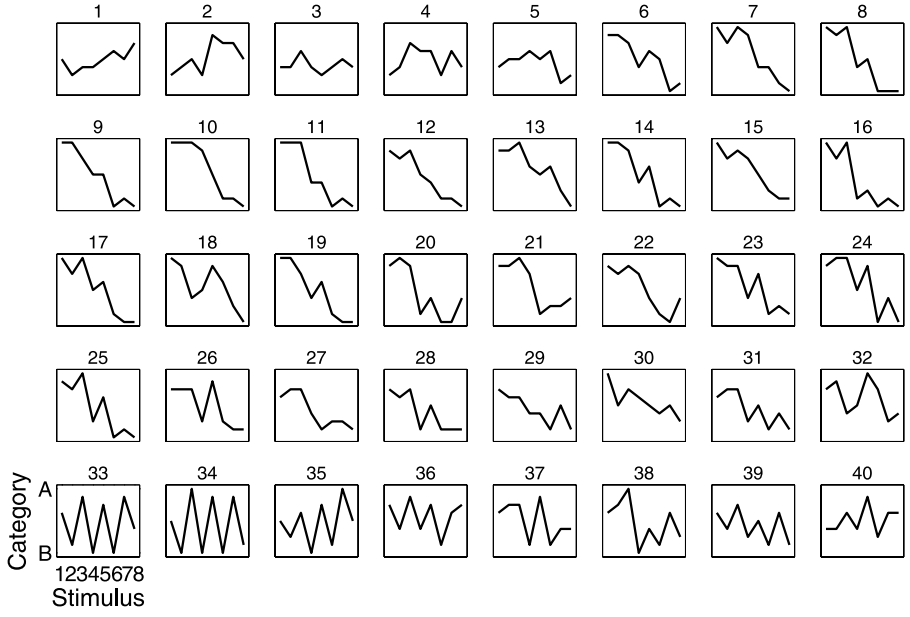
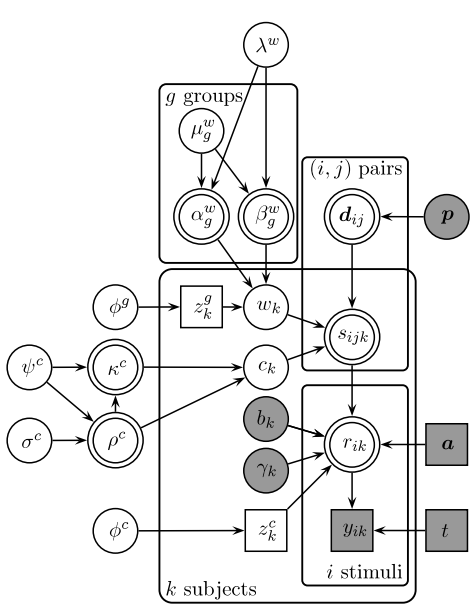


Fig. 5. Observed category decisions for the individual subjects, showing the proportion of times each stimulus is categorized as belonging to category A.



$$
\begin{aligned}
(\mu_1^w, \mu_2^w) &\sim \text{Uniform}(0,1)^2, \quad \mu_1^w > \mu_2^w \\
\lambda^w &\sim \text{Gamma}(1.1051, 0.01051) \\
\alpha_g^w &\leftarrow \mu_g^w \lambda^w \\
\beta_g^w &\leftarrow (1-\mu_g^w)\lambda^w \\
\psi^c &\sim \text{Uniform}(0.01, 10) \\
\sigma^c &\sim \text{Uniform}(0.01, 10) \\
\kappa^c &\leftarrow 1 + \psi^c \rho^c \\
\rho^c &\leftarrow \frac{\psi^c + \sqrt{(\psi^c)^2 + 4(\sigma^c)^2}}{2(\sigma^c)^2} \\
\phi^c, \phi^g &\sim \text{Uniform}(0,1) \\
z_k^c &\sim \text{Bernoulli}(\phi^c) \\
z_k^g &\sim \text{Bernoulli}(\phi^g) \\
c_k &\sim \text{Gamma}(\kappa^c, \rho^c) \\
w_k &\sim \begin{cases} \text{Beta}(\alpha_1^w, \beta_2^w) \text{ if } z_k^c = 0,\ z_k^w = 0 \\ \text{Beta}(\alpha_2^w, \beta_2^w) \text{ if } z_k^c = 0,\ z_k^w = 1 \end{cases} \\
b_k &\leftarrow \tfrac{1}{2} \\
\gamma_k &\leftarrow 1 \\
d_{ij}^{[m]} &\leftarrow |p_{im} - p_{jm}| \\
s_{ijk} &\leftarrow \exp\left\{-c_k\left(w_k d_{ij}^{[1]} + (1-w_k)\, d_{ij}^{[2]}\right)\right\} \\
r_{ik} &\leftarrow \frac{b_k\left(\sum_{j\in A} s_{ijk}\right)^{\gamma_k}}{b_k\left(\sum_{j\in A} s_{ijk}\right)^{\gamma_k} + (1-b_k)\left(\sum_{j\in B} s_{ijk}\right)^{\gamma_k}} \\
y_{ik} &\sim \begin{cases} \text{Binomial}(r_{ik}, t) \text{ if } z_k^c = 0 \\ \text{Binomial}(\tfrac{1}{2}, t) \quad \text{if } z_k^c = 1 \end{cases}
\end{aligned}
$$

Fig. 6. Graphical model implementation of the GCM analysis with individual differences, allowing for three latent groups of subjects.

one group attending less to the position dimension than the other group.

The discrete individual differences are captured by the mixture component of the model, which assigns subjects to one of the three groups. Within each group, people are expected to be similar, but not identical. We allow for these additional continuous differences within groups by adding an hierarchical component to the model, meaning that the individual subject parameters are drawn from group distributions. Overall, the resulting hierarchical mixture model allows for discrete and continuous individual differences, as per the right-most sub-panel of Fig. 1.

4.3.1. A hierarchical mixture extension of the GCM

Fig. 6 shows the graphical model that extends the GCM to allow for the three types of individual differences just described. Two indicator variables, z_k^c and z_k^g identify the group membership for the kth subject. The first variable z_k^c indicates whether the subject's categorization behavior is guided by the GCM process, or if the subject belongs to the contaminant group. The second variable, z_k^g indicates, when the GCM is the appropriate model, to which of the two parameterizations the subject belongs. Both variables are Bernoulli distributed, with base-rates of ϕ^c and ϕ^g, respectively.

There is a plate for the subjects, so the kth subject has attention w_k and generalization c_k parameters. These parameters are hierarchically drawn from hyper-distributions corresponding psychologically to groups of subjects. The generalization parameters c_k is hierarchically sampled from a single gamma distribution, parameterized by its mode ψ^c and standard deviation σ^c (Kruschke, 2012), that do not differ over groups. The attention parameters w_k, which are constrained to lie between 0 and 1, are drawn from Beta distributions, parameterized by their mean μ_g^w for the gth group and precision λ^w.

The parameters ψ^c, σ^c, λ^w, ϕ^c and ϕ^g are all given vague priors. To formalize the theoretical assumption that the mean for the attend position group will be larger than that of the attend size group, an order restricted prior is placed on the μ_g^w parameters in a way that generates a uniform distribution over the joint parameter space and gives equal density to the valid region in which $\mu_1^w > \mu_2^w$ (Lee & Wagenmakers, 2013).[6]

The mixture components dictated by z_k^g and z_k^c, allow people to belong to qualitatively different groups, as per the "Discrete Differences" sub-panel of Fig. 1. The hierarchical extensions allow people within the same group to have similar but different attention and generalization parameters, as these are sampled from a continuous distribution corresponding to their group, as per the "Continuous Differences" sub-panel of Fig. 1.

4.3.2. Results

The results are again based on 3 chains of 10,000 samples each, with a burn-in of 1000 samples, whose convergence was checked. The upper panel of Fig. 7 shows the posterior mean of the z indicator variables corresponding to the inferred allocation of the 40 subjects into the three groups. The first 5 subjects are inferred to belong to the contaminant group, the next 27 subjects to the attend position group, and the final 8 to the attend height group. Only for a couple of subjects the inference is somewhat uncertain. For the vast majority of the individual subjects, the inference regarding their group membership is reasonably certain, which is a first indication that the three proposed groups provide a useful perspective on the nature of the differences between the individuals.

The joint and marginal posterior distributions of μ_g^w and ψ^c are shown in the two next panels. The first group has an attention weight around 0.8, indicating that most attention is given to the position of the interior line of the stimuli. The second group has an attention weight close to 0. This group attends almost exclusively to the height of the stimuli.

[6] A difficulty that can arise with applying Bayesian mixture models is the identifiability of the groups. This problem is known as label switching (Jasra, Holmes, & Stephens, 2005), occurring because the posterior distribution is invariant to permutations in the labeling of parameters, and causing the different marginal posterior distributions to be identical for each mixture component. A general solution to deal with this problem is to provide additional information to the model to tell the different groups apart. In this application, label switching was not an issue because of the inequality constraints put on the parameter space.

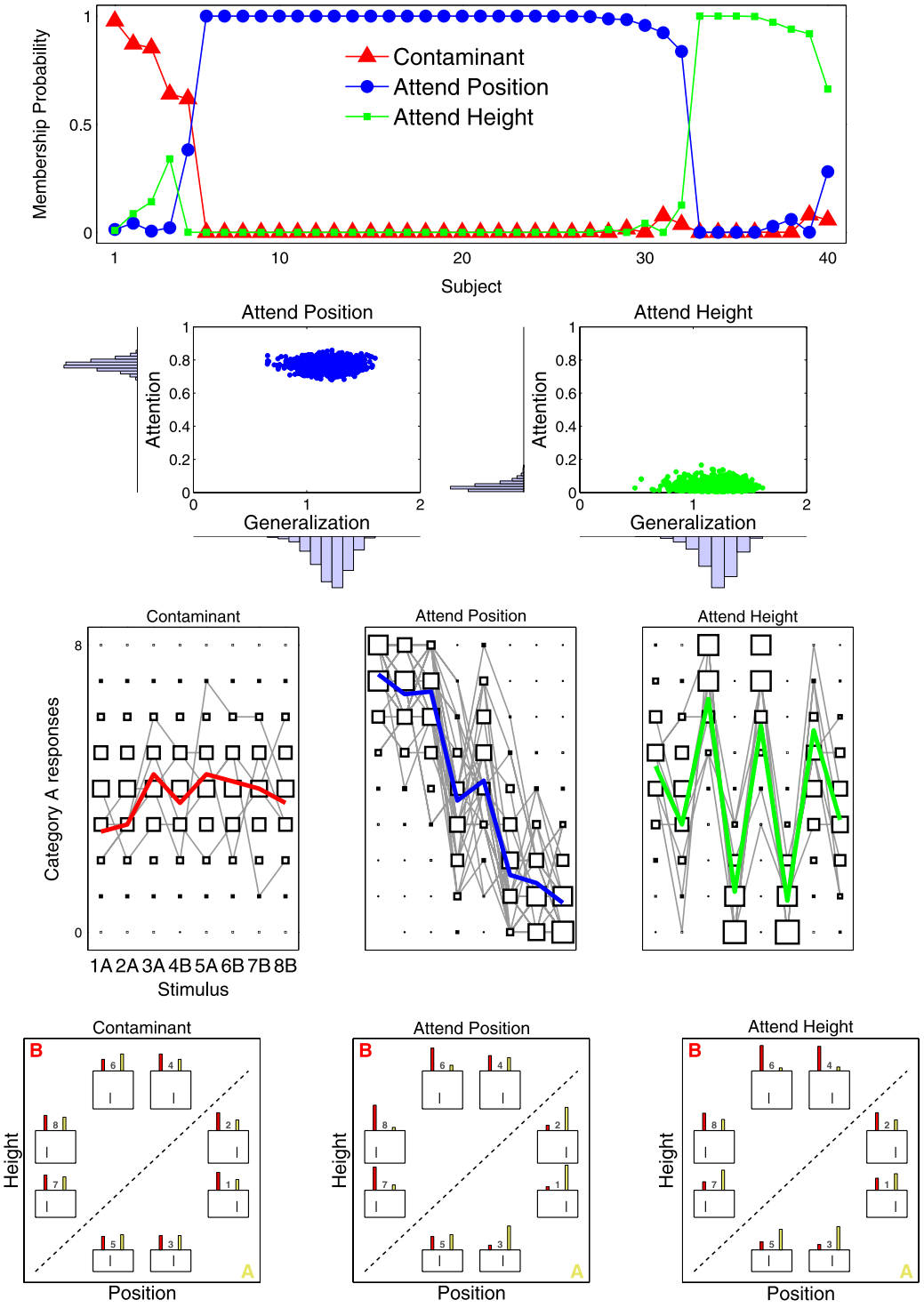


Fig. 7. Results from the GCM analysis with individual differences, assuming three latent groups of subjects, showing the allocation of subjects to groups, the posterior and posterior predictive distributions for the groups, and the interpretation of the different groups in terms of the stimuli and category structure itself (see text for details).

The next row shows a posterior predictive check of the model to the behavioral data, overlaying the models predictions and the empirical data. A posterior predictive check is a basic Bayesian method for assessing the fit of a model to data, based on integrating the data it generates across the joint posterior distribution (Lee & Wagenmakers, 2013; Shiffrin et al., 2008). 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 the square being proportional to posterior predictive mass. The many thin lines show the individual subject empirical behavior for that group, with subjects assigned to their most likely group, and the single thick line averages these thin lines. It is clear that the three groups have subjects showing qualitatively different patterns of categorizing the stimuli. The close correspondence between the data (lines) and the posterior model predictions (squares) indicate the model is able to capture the data sufficiently well.

The bottom panels of Fig. 7 interpret the different category learning of the groups. The original stimulus space and category structure is shown, with bars above the stimuli showing the

average number of times each stimulus is placed in category A (right bar) and category B (left bar) by the members of that group. These panels show that subjects in the contaminant group categorize all stimuli roughly equally in both categories. For the attend position group, 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 the attend height group, stimuli 2 and 7 are categorized poorly, which is consistent with a focus on the height dimension.

4.4. Discussion

Our hierarchical mixture analysis of Kruschke's (1993) Condensation B data, using a GCM with the ability to model continuous and discrete individual differences, tells a potentially interesting story. It suggests that besides contaminant subjects not learning effectively, there are two groups of subjects, each of whom focus most of their attention on just one stimulus dimension while learning the category structure. The result of the analysis on the grouped data, showing attention being distributed roughly evenly across both dimensions, may be a by-product of failing to consider individual differences in modeling, just as Estes warned.

In our hierarchical mixture extension of the GCM, the number of groups was not inferred but rather assumed a priori. It might thus seem that our analysis reached a foregone conclusion, in the sense that we postulated and found three groups, which would have little explanatory power. However, imposing a model with three groups does not automatically result in finding three groups. If the GCM subjects were not usefully modeled as using different attention allocations, this would be clear from the results. The model only creates a possibility that there are different groups, but this assumption can be overruled by the data if it is not plausible. To make this general point concretely, we applied the same model to the other conditions in the Kruschke (1993) experiment. Details are reported in Appendix A, but we summarize the key findings to explain how data can evaluate the appropriateness of modeling assumptions about individual differences.

One straightforward way in which the data can overrule the assumption about the number of groups is when groups are inferred to be empty. This turns out to be the case for the "Filtration Position" condition of Kruschke (1993), for which only the position dimension is relevant for correct categorization, as is clear from the top-left panel of Fig. 2. When we applied the same model to these data, only one group was inferred by the model and no subjects were assigned to either the contaminant and second attention group.

A second situation in which we start by assuming a three group model but do not conclude that there are three different groups of subjects is when the model turns out to be not appropriate for the data. This turns out to be the case for the "Filtration Height" condition, shown in the top-right panel of Fig. 2, for which only the height dimension is relevant. When the model is applied to these data, most of the 40 subjects are inferred to belong to the attend height group, with a attention weight close to zero, while six subjects are assigned to the attend position group, with attention weights above 0.5. Crucially, a posterior predictive check reveals that the model does not meet a basic requirement of descriptive adequacy. In particular, the GCM is not an adequate model for those subjects in the attend position group, as the model predictions, shown by squares, do not seem to come close to the data, shown by lines. This basic deficiency of the model neuters the interpretation of there being two different attention groups. Again, although we assumed three groups, we again do not conclude there are three groups.

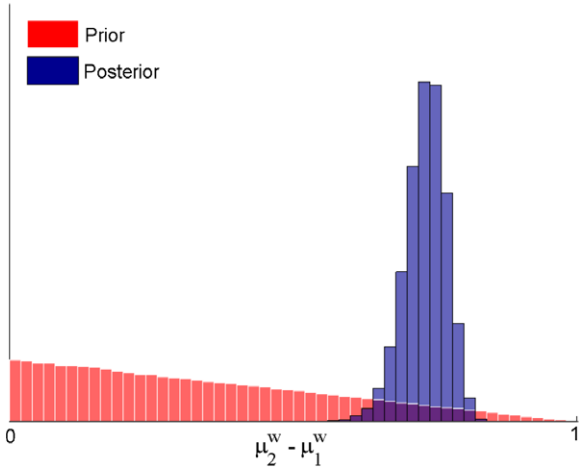


Fig. 8. Prior and posterior density for the difference in μ^w for the attend position and attend height group.

Even finding non-empty groups, however, using a model that is descriptively adequate is not sufficient to have confidence in the number of groups posited by the model. A third way in which imposing a model with three groups does not automatically result in three different groups of subjects is illustrated by applying the model to the "Condensation A" condition, shown in the bottom-left panel of Fig. 2. Applying the three-group model to these data, we find the model assigning subjects to each group, and having sufficient descriptive adequacy. However, inspection of the marginal posterior distributions for the two attention groups reveals that these distributions heavily overlap. So although we find two attention groups, it turns out that the two groups are not very different in their attentional allocation, with several subjects having an attention weight close to 0.5. Therefore, it is meaningless to conclude there are two different groups of subjects allocating attention differently. The indistinguishability of both groups is also reflected in the assignment of subjects to the groups, which is more uncertain than in the analyses of the other conditions of Kruschke (1993). Thus, once again, although we assumed three groups, we do not find clearly distinguished groups and thus do not conclude there are two groups.

As these counter-examples make clear, it is only for the "Condensation B" task that we assumed three groups, observed that the model described the data sufficiently well, that all groups were assigned subjects, and that the differences between groups assumed in the model were observed in the posterior. Based on this constellation of results, we gained confidence that it is reasonable to infer that there are three groups of subjects.

The reasonableness of assuming three groups, however, does not preclude the reasonableness of a different number of groups, just as finding evidence for any model of a cognitive process does not preclude the possibility of the data also being consistent with a different model. Accordingly, we also considered both a two-group model, assuming only one attention group and a contaminant group, and a four-group model, assuming an additional group of subjects dividing attention more equally across both dimensions. We compared the two- and three-group model by calculating a Bayes Factor, using the Savage–Dickey method, a method that applies to nested models only (see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010, for a tutorial). The ratio of the prior and posterior density for the difference between μ_1^w and μ_2^w at zero approximates the Bayes Factor, assuming all other parameters in the model are nuisance parameters. The absence of posterior samples near zero prohibited a reliable quantitative estimation of the Bayes factor, but visual comparison of the prior and posterior densities at zero, as shown in Fig. 8, indicated a massive preference for the three-group model over the two-group model.

The results of the application of the four-group model to the "Condensation B" data are presented in Appendix B. No subjects are clearly inferred to belong to the attend both group, and the group attention weights for the attend both and attend position group overlap, suggesting little useful additional psychological insight is provided by including the fourth group. Following the same strategy for comparing the two- and three-group model, we looked at the prior and posterior difference between the attend both and attend position group. Comparison of the prior and posterior densities for the difference between μ_2^w and μ_3^w at zero indicates no preference for the model with the extra group. Thus, overall, there seems to be no strong evidence for a group of subjects paying equal attention to both dimensions for this condition.

While we believe, on the basis of these analyses, the three-group model is a useful and appropriate one, we do not claim that any strong theoretical conclusions regarding the usefulness of selection attention as a psychological variable immediately follow. First, different conclusions might be reached using different models, or different version of the same model. Further, making firm theoretical claims about selective attention will only be possible when data collected across a wide variety of conditions are considered. For example, selective attention has been hypothesized to be controlled by factors such as, the integral or separable nature of the stimulus dimensions, the perceptual salience of the stimulus dimensions, degree of learning, the exact learning conditions, and the relative diagnosticity of the stimulus dimensions (Nosofsky, 1986; Vanpaemel & Lee, 2012). Accordingly, we certainly do not claim our example application has immediate strong implications for the existing large and coherent body of work examining selective attention mechanisms in category learning.

What the application does show, however, is how the hierarchical mixture approach, coupled with Bayesian analysis, provides a powerful and general approach for taking individual differences into account. Very flexible assumptions about the nature of individual differences for basic category learning processes can be implemented and evaluated against data, and there is the possibility of different theoretical conclusions being drawn because individual differences are acknowledged.

5. Second application: individual differences in category representation

Our second illustration of how the Bayesian hierarchical mixture approach can be used to reveal and deal with individual differences involves a problem that is traditionally tackled from a model selection perspective. Staying in the domain of category learning, we now consider the empirical comparison of the GCM with the Multiplicative Prototype Model (MPM), which Estes helped to introduce (Estes, 1986a, see also Minda & Smith, 2001; Nosofsky, 1987, 1992; Reed, 1972).

The crucial difference between the GCM and the MPM is the way categories are assumed to be represented. As explained in the first application, the GCM proposes that people store individually all of the exemplars of a category in memory. When a categorization decision is made, the presented stimulus is compared to all of the stored exemplars to decide to which category it belongs. The MPM, on the other hand, assumes that an abstract summary of the stimuli belonging to a category is stored in memory. When a categorization decision is made, the presented stimulus is compared only to these summaries, referred to as prototypes, to make a categorization decision. Apart from the category representation assumed, the MPM and GCM are formally identical.[7] Comparing models that differ in only a single assumption is a powerful approach to model evaluation advocated by Estes, who realized the importance and appeal of well-controlled comparisons.

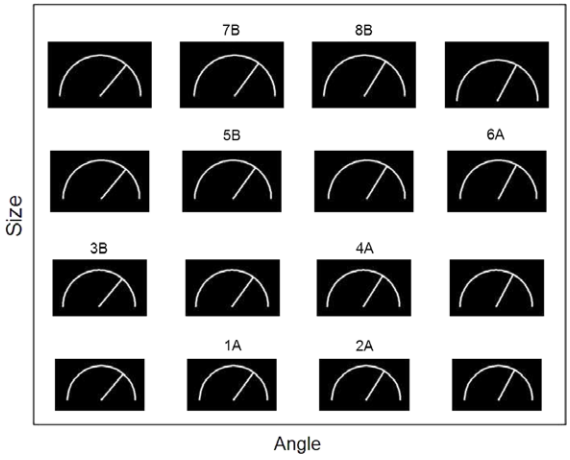


Fig. 9. Diagonal category structure, based on Nosofsky (1989).

In this application, we demonstrate how the Bayesian hierarchical mixture approach can be used to investigate the possibility that different people might rely on different representations when learning the same category structure. We again present two analyses, this time using newly collected data. In the first, we assume there are no individual differences and consider the grouped data, while in the second, we use the hierarchical mixture approach to allow for structured continuous and discrete individual differences.

5.1. Data

5.1.1. Subjects

Thirty-one subjects completed the experiment. They were aged from 17 to 29 years, with a mean age of 20.3 years. There were 9 males and 22 females.

5.1.2. Stimuli

The stimuli were 16 so-called 'Shepard circles' (Nosofsky, 1989), which are semi-circles with radial lines, varying on two continuous dimensions. They varied orthogonally in size (four levels of radius length: 0.904, 1.016, 1.128, 1.24 cm) and angle of orientation of the radial line (four levels: 46°, 54°, 62°, 70°). The categories to be learned correspond to a diagonal structure, which was taken from Nosofsky (1989), and is shown in Fig. 9. Four training stimuli were assigned to category A and four to category B. The remaining eight transfer stimuli were left unassigned. For ease of reference, the training stimuli are numbered 1–8, together with their category assignment A or B.

5.1.3. Procedure

Subjects were presented with 40 training blocks in which each of the training stimuli was presented once per block in random order. During these trails, subjects received corrective feedback and their total percentage correct was shown. Stimuli were presented until an answer was given by pressing one of two buttons. Following Johansen and Palmeri (2002), after 4, 8, 12, 16, 24, 32 and 40 blocks of training, a single transfer block was presented during which all 16 stimuli were presented once, without feedback. The experiment consisted in total of 432 trials, of which 320 were training trials and 112 were transfer trials. The experiment took about 30 min to finish.

Our analysis focuses on the data from the last 16 training and 3 transfer blocks. The mean percentage correct over these last

[7] As will become clear below, the MPM does not have a response determinism parameter in its response rule. Thus, the MPM and GCM our only identical – except for the assumed representation – when this parameter is assumed to be equal to 1, as was the case in the original version of the GCM, and in the current applications.

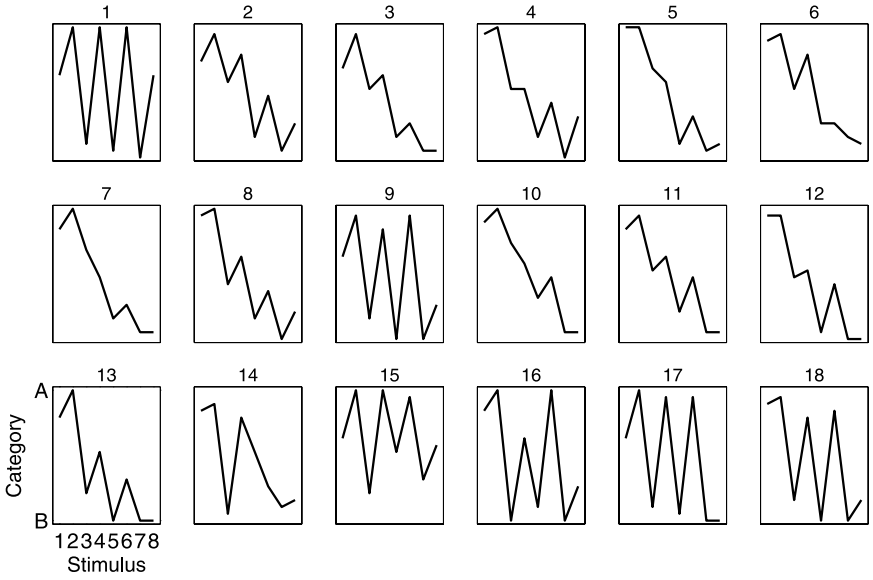


Fig. 10. Observed category decisions for the individual subjects, showing the proportion of times each stimulus is categorized as belonging to category A.

16 training blocks was 69%. Following Nosofsky (1998), we only considered data from learners who achieved 70% accuracy or more, leading to the exclusion of 13 subjects. The remaining 18 subjects had an average percentage correct of 78% over the last 128 trials. Again, the relevant data in this application are y_{ik}, the number of times the ith stimulus was categorized as belonging to category A by the kth subject, out of the $t = 19$ (for the assigned stimuli) or $t = 3$ (for the unassigned stimuli) trials on which it was presented.

5.2. Analysis assuming no individual differences

As in the first application, the grouped data are the total number of times all subjects classified the ith stimulus into category A, out of $t = 18 \times 19$ or $t = 18 \times 3$ total presentations: $y_i = \sum_k y_{ik}$.

5.2.1. The MPM

One crucial difference between the MPM and the GCM is that in the MPM determining the overall similarity between a stimulus and a category involves considering the similarity between that stimulus and a summary item—the prototype. The most dominant implementation of a prototype is the average of all category members, so the coordinates of a prototype used by the MPM are the average of the coordinates of the exemplars of that category. The coordinates come from the multi-dimensional scaling solution from Nosofsky (1989), who used the same stimuli. The similarity between the stimuli and the prototypes is calculated much as for the GCM. Thus, formally, the similarity between the ith stimulus and category A is $s_{iA} = \exp\left\{-c\left(w d_{iA}^{[1]} + (1-w)\, d_{iA}^{[2]}\right)\right\}$, with $d_{iA}^{[m]}$ the distance between the i the stimulus and the category A prototype on dimension m. The response probability for the ith stimulus is then given by $r_i = b s_{iA} / (b s_{iA} + (1-b)\, s_{iB})$. As noted by Ashby and Maddox (1993), the response determinism parameter γ is not identifiable in the MPM, so cannot be incorporated in the model.

5.2.2. Results

To compute the Bayes Factor for comparing the MPM and the GCM, we evaluated each marginal likelihood separately using standard numerical grid sampling, using steps of 0.001 for w and 0.005 for $0 < c < 5$. The Bayes Factor indicates that the grouped data strongly favor the GCM over the MPM, with the log Bayes Factor being approximately 18 in favor of the GCM. This finding that people rely on all exemplars when learning the category structure rather than on a single prototype is consistent earlier studies, using similar learning conditions, similar stimuli, and categories of similar size, complexity and structure (Nosofsky, 1992; Nosofsky & Zaki, 2002; Vanpaemel & Storms, 2010).

5.3. Analysis assuming individual differences

Our analysis based on grouped data, suggesting people rely on an exemplar representation, produces results that are consistent with much of the literature, and makes psychological sense given the limited need for cognitive economy the category structure imposes. However, once again, inspection of the individual data, as shown in Fig. 10, suggests that there are considerable individual differences, ignored by the analysis of aggregate data.[8] Some subjects – such as subjects 17 and 18 – make almost no mistakes in categorizing the stimuli. Other subjects – such as subjects 2 and 3 – incorrectly categorize stimuli 3 and 6, which can be thought of as exception stimuli, since these stimuli lie close to the opposing category prototypes. It seems plausible that these individual differences might be related to different representational strategies.

Our analysis with structured individual differences again takes the form of a hierarchical mixture model, assuming two latent groups with different categorization behavior: a group of MPM learners and a second group of GCM learners.[9]

The discrete individual differences correspond to different category representations, following either the GCM or the MPM. These are captured by the mixture component of the model, assigning

[8] While all analyses are done using both the training and transfer data, all figures show the training data only, as these provide most information.

[9] Unlike in the first application, we do not report analyses based on a model including a contaminant group in this application. An analysis not reported here using a model including such a group classified two subjects as contaminants only, while inspection of the percentages correct reveals that there are many more subjects that did not learn the category structure. Clearly, the basic contaminant approach in which it is assumed that all subjects that do not follow the MPM or the GCM are simply guessing is not successful in removing subjects who did not learn the structure. It is possible to include a more sophisticated contaminant process than simply guessing, as other processes besides guessing might be underlying the behavior of the 13 subjects who did not learn the category structure. Additional creative modeling work is needed, however, to determine what these subjects were doing (see Zeigenfuse & Lee, 2010).

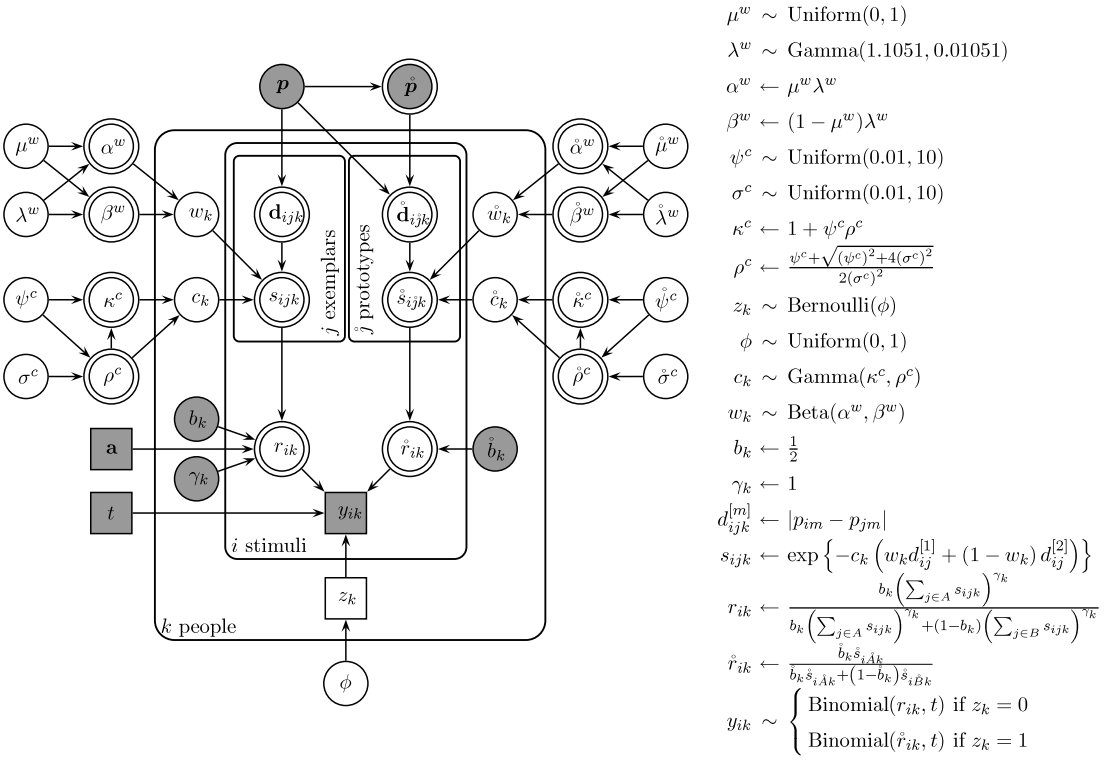


Fig. 11. Graphical model implementation of the hierarchical mixture of the GCM and MPM, allowing for individual differences.

each subject to one of these two latent groups. Within each group, the hierarchical component allows for continuous individual differences, corresponding to different parameterizations of the category learning processes that act on the exemplar or prototype representations. In sum, we assume both discrete and continuous individual differences, as shown in the main panel of Fig. 1.

5.3.1. A hierarchical mixture model for comparing the GCM and the MPM

Fig. 11 shows the graphical model for the hierarchical mixture model of the GCM and MPM category learning accounts. The coordinates in **p** give the representations of both the individual training and transfer stimuli, and so the training stimuli give the exemplar representation used by the GCM. The prototype representation for the categories used in the MPM follows from averaging these training coordinates, and is denoted by $\dot{\mathbf{p}}$. The left side of the graphical model relates to the GCM, and the right side corresponds to the MPM. Only the definitions for the exemplar models are given next to the model, as the definitions and prior distributions for the prototype model are exactly the same, provided that instead of the exemplars coordinates the prototype coordinates are used. The only exception is for the response rule $\dot{r}_{ik}$, since the response determinism parameter is not identifiable in the MPM (Ashby & Maddox, 1993).

The indicator variable z_k controls which representational model, and parameterization of the processes acting on that model, are used by the kth subject. Within each representational possibility, individual subject parameters c_k and w_k are drawn from gamma and beta group distributions, respectively. As in the first application, these distribution are parameterized using location and dispersion parameters, which are assigned vague priors. Since we are now considering two different models, we relax the assumption made in the first application that some parameters do not change over groups.[10] These indicator variables are assumed to follow a base-rate ϕ, so that $z_k \sim \text{Bernoulli}(\phi)$, and the base-rate is given a uniform prior $\phi \sim \text{Uniform}(0, 1)$.

Again, the indicator variable z accounts for discrete individual differences, displayed in the "Discrete Difference" panel of Fig. 1, while the hierarchical distributions over the generalization and attention parameters account for the continuous individual differences, displayed in the "Continuous Difference" panel of Fig. 1.

5.3.2. Results

As in the previous analyses, the results are based on 3 chains of 10,000 samples each, with a burn-in of 1000 samples, whose convergence was checked. The upper panel of Fig. 12 shows the posterior mean of z, indicating the latent assignment of the subjects to the GCM and MPM groups. The latent assignment for the subjects is somewhat less certain than those in the first application. One reasonable summary is that the first 7 subjects are identified as prototype learners, the last 9 subjects are classified as exemplar learners, but there is considerable uncertainty for the remaining 2 subjects. It is clear that there are individual differences, and that both the MPM and GCM are useful in describing the behavior of the subjects as a whole.

While the main goal of the current analysis to make inferences from these patterns of assignment to the prototype and exemplar representational models, an attractive property of doing fully Bayesian inference over graphical models is that joint and marginal posterior distributions are automatically available for all of the other parameters. This includes group as well as individual subjects distributions for the attention w and generalization c parameters.

[10] Model identifiability, addressed through theoretical meaningful priors in the first application, is addressed in this application by the use of meaningfully different data-generating models—the GCM and the MPM.

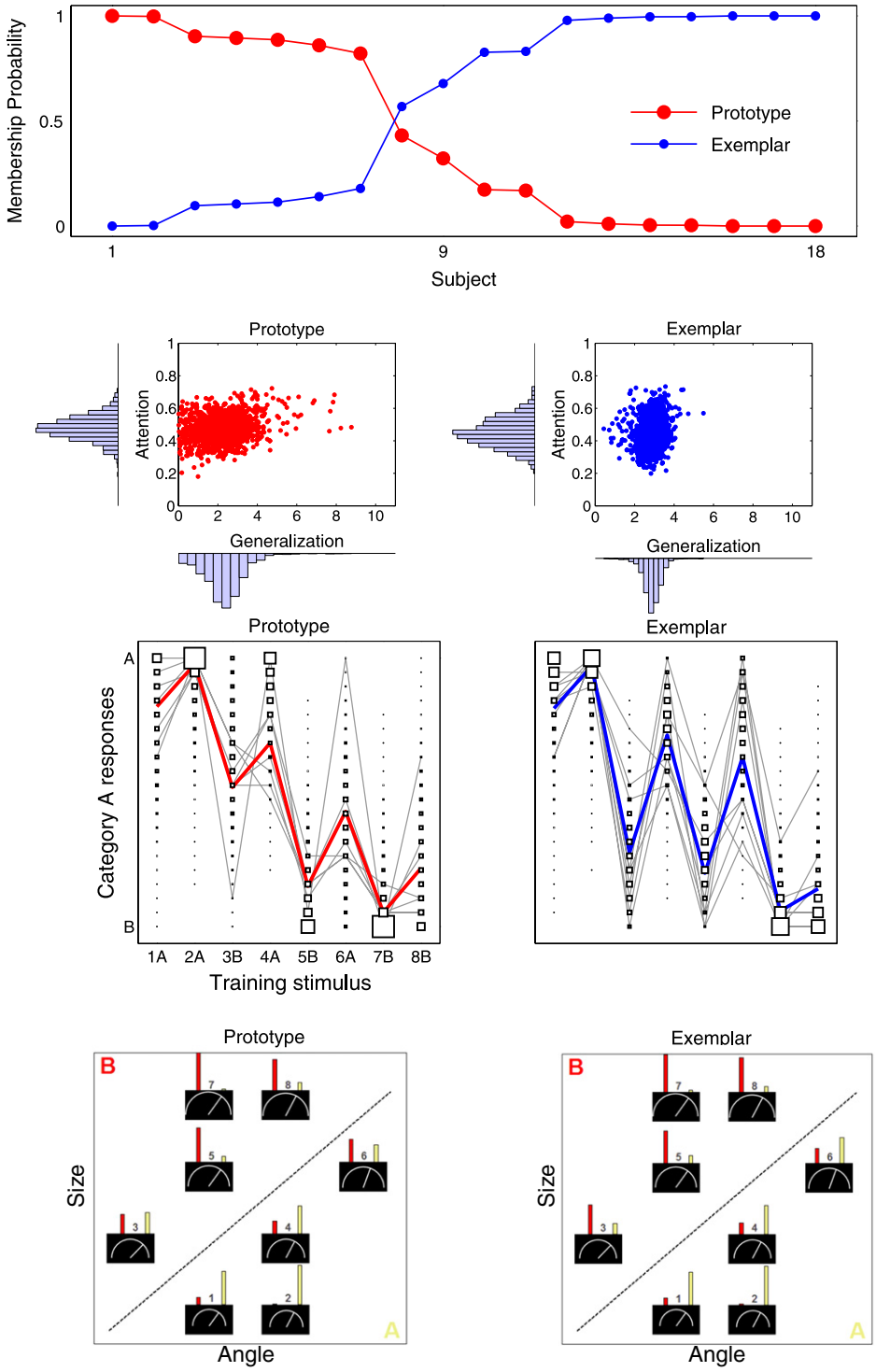


Fig. 12. Results from the analysis using a hierarchical mixture of the GCM and MPM, showing the allocation of subjects to groups, the posterior and posterior predictive distributions for the groups, and the interpretation of the different groups in terms of the stimuli and category structure.

Accordingly, the second row of panels in Fig. 12 shows the joint and marginal posterior distributions for the group parameters μ_g^w and ψ_g^c. There is a similar range for both parameters across both groups, although there is somewhat more uncertainty in the estimation of the group level generalization parameter for the subjects using the prototype representation. This sort of information is not directly available when exemplar and prototype debate is seen from an exclusively model selection perspective and when for example, a Bayes Factor is computed.

The next row in Fig. 12 shows the posterior predictive distribution for both groups, using squares. It also shows, as thin lines, the individual subject data for the subjects, with subjects assigned to their most likely group, as well as their average using a thick line. It is clear that two qualitatively different patterns emerge. The hierarchical mixture model seems to be able to describe the data reasonably well, as indicated by the agreement between the data (lines) and the posterior model predictions (squares).

The results are interpreted in terms of concrete behavior in the bottom panels in Fig. 12. The right and left bars plotted above the

stimuli show the average number of times each stimulus is placed in category A and category B, respectively, by the members of that group. The prototype subjects have difficulty classifying stimuli 3 and 6. They classify stimulus 3 slightly more often as belonging to category A, although it belongs to category B. They also incorrectly classify stimulus 6 more often as belonging to category B than category A. Subjects inferred to rely on an exemplar representation, instead tend to overall classify all training stimuli correctly.

5.4. Discussion

In line with Estes' early warnings, the results of our hierarchical mixture analysis indicate again that the inferences based on the grouped data are, at best, incomplete and, at worst, misleading. Although there was decisive evidence for the GCM based on the aggregated group data, the hierarchical mixture analysis allowing for individual differences suggests, instead, that there is a group of prototype learners and a group of exemplar learners.

As in the first application, the number of groups was chosen a priori rather than inferred. Fixing the number of groups might seem much less problematic here than in the first application, as the basic research question of contrasting the GCM and MPM determined the appropriate models. There is no deep reason, however, the modeling approach we have used cannot consider additional models, such as the those provided by Varying Abstraction Model (Lee & Vanpaemel, 2008; Vanpaemel & Storms, 2008). This sort of expanded consideration is exactly analogous to the consideration of different numbers of groups we pursued in the first application. In general, any set of candidate accounts of how people learn categories (or any other cognitive capability) can be formalized and evaluated using the hierarchical models and Bayesian inference methods we have demonstrated. Developing the models themselves, and choosing which set to evaluate, remain creative modeling acts that are not formalized by our approach. But, once specified, their specification as generative graphical models, and evaluation using Bayesian inference, can proceed in the ways our applications have demonstrated.

As for the first application, our main purpose was to demonstrate the usefulness of a hierarchical mixture approach to reveal and account for individual differences in cognition rather than making firm theoretical claims about representational abstraction. Indeed, there are several important reasons that our analyses are of limited use for making such claims. First, in order to draw strong theoretical conclusions, analyses might need to test more elaborate versions of the GCM and MPM than the basic version we used, relying on a more elaborate response rule including a bias parameter and a response determinism parameter. Appendix C reports the results of an analysis comparing the MPM to a version of the GCM where the response determinism parameter is not fixed but rather given a gamma prior. It turns out that all but one subjects were assigned to the GCM, showing that, firstly, theoretical conclusions might crucially depend on the inclusion of a response determinism parameter (Nosofsky & Zaki, 2002) and, secondly, that including the response determinism parameter is of limited value for our purpose of demonstrating the hierarchical mixture approach to reveal individual differences. Another reason we do not draw strong theoretical conclusions is that claims about representational abstraction requires systematic investigation of category learning across many tasks, as there are many factors that have been shown to influence the level of abstraction, such as the time point in learning, category size, category complexity (Feldman, 2000), and stimulus complexity (Johansen & Palmeri, 2002; Minda & Smith, 2001). Obviously, the data we have considered correspond to only one of many possible experimental conditions.

What this application does demonstrate is the usefulness of the hierarchical mixture approach, in allowing modeling to include assumptions about individual differences. In addition, by showing how the same hierarchical mixture approach to individual differences as used in the first application can also be applied to problems that are typically treated from a model selection perspective, this second application highlights the flexibility and broad applicability of the approach.

6. General discussion

In this paper, we demonstrated the usefulness of hierarchical latent mixture models for modeling and understanding individual differences in cognition. Focusing on category learning models, we demonstrated in two applications how individual differences can be treated using a hierarchical mixture framework. In both of the applications, a hierarchical mixture approach reveals clear individual differences, and thus paints a fundamentally different picture than an analysis based on grouped data. We showed how looking at the group assignment, model fit and interpretability of the inferences provide guidelines for choosing a number of groups, and highlighted that choosing the number of groups is akin the choosing the number of models under consideration when the goal is model selection.

A feature of the Bayesian hierarchical mixture approach for dealing with individual differences used in our two applications that deserves highlighting is the fact that it is broadly applicable. It can be used to address problems that are traditionally conceived of as estimation problems, such as in the first application where the goal was inferring attention, and to problems for which answers are traditionally sought using model selection methods, such as in the second application focusing on inferring representational abstraction. In the selective attention application, we considered individual differences relating to two different parameterizations of the same GCM process model. In the representational abstraction application, we considered two different models of category representation, the GCM and MPM. In this sense, the first case study focused on parameter estimation, while the second case study focused on model selection. In both cases, however, a similar hierarchical mixture approach could naturally be applied, with problems of model selection being treated as a special type of parameter estimation. It speaks to the wide applicability of the hierarchical mixture approach that two problems that are traditionally seen as corresponding to different aspects of inference (e.g., Stephan et al., 2010) can be cast under the same umbrella. By treating estimation and selection, the hierarchical mixture blurs the boundaries between model selection and parameter estimation (see also Lee & Vanpaemel, 2008).

The hierarchical mixture approach is much more flexible than we could illustrate in a single paper or in two illustrative applications. All sort of extensions of existing models are made possible by using hierarchical mixture models and Bayesian inference, and we think they will be needed to arrive at a richer account of cognitive processes. For example, as noted earlier, a useful extension would be to include a more sophisticated contaminant process. Another useful extension in the second application would be to, as in the first application, assume different groups of subjects within each model by allowing for different parameterization within both the GCM and MPM.

Overall, we believe the major strengths of the Bayesian hierarchical mixture approach for modeling individual differences is its broad applicability, generality and flexibility. All sort of extensions of existing models are made possible by using hierarchical mixture models and Bayesian inference, and we think they will be needed to arrive at a richer account of cognitive processes. The focus is on building a rigorous and model-based understanding of the core human cognitive capabilities, and doing so in a way that respects the way people are different, as well as what makes them the same. We think and hope Estes would have found promise in this approach.

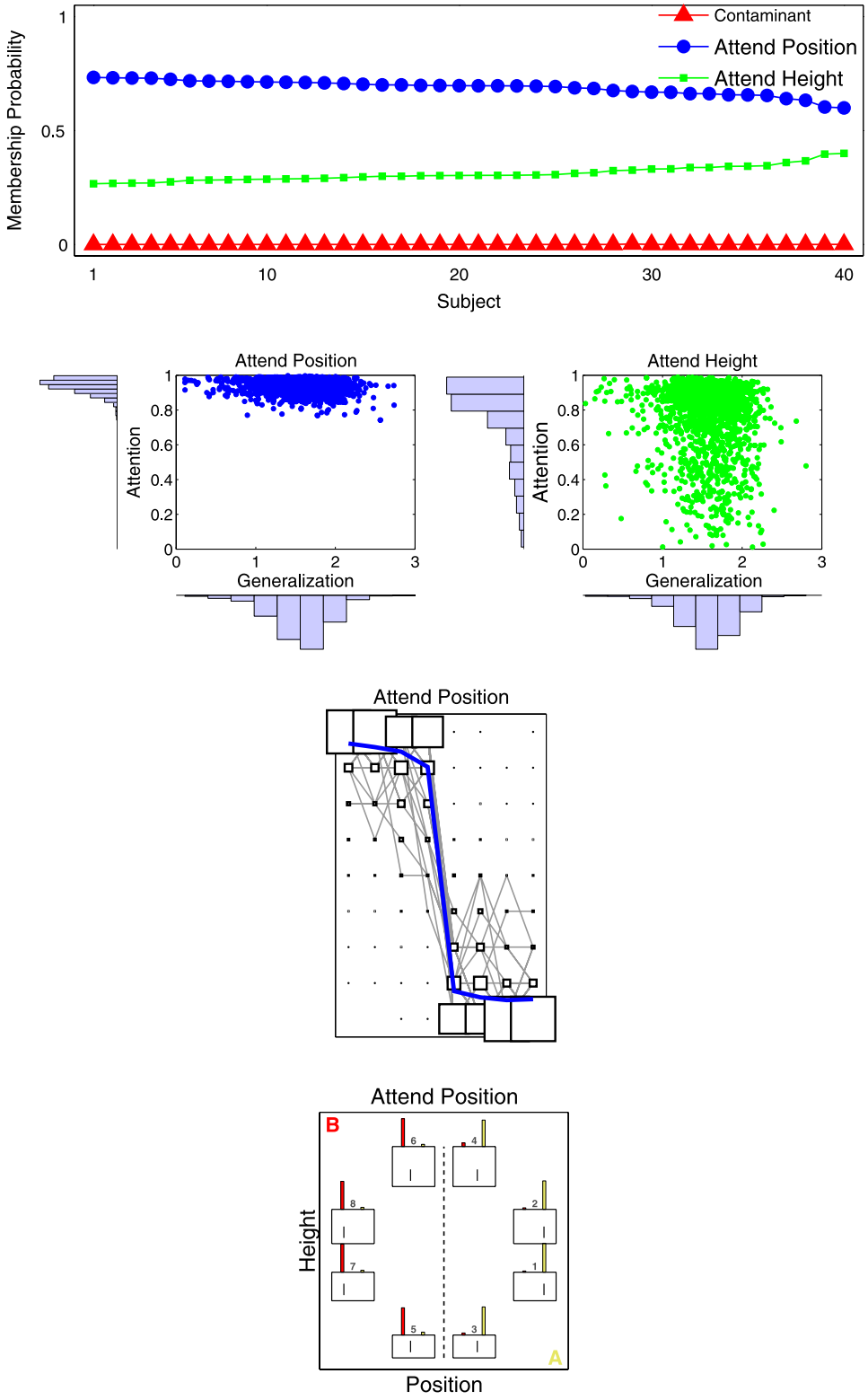


Fig. 13. Results from the selective attention analysis with individual differences on the "Filtration Position" condition, assuming three latent groups of subjects. The results show the latent assignment of subjects to groups, the posterior and posterior predictive distribution for the groups, and the interpretation of the groups in terms of the stimuli and category structure itself.

Appendix A. Applying the three-group model to the other conditions of Kruschke's (1993) experiment

In this section, we provide some details on the application of the three-group model to the other three conditions described by Kruschke (1993), the category structures of which are shown in Fig. 2. The results are shown in Figs. 13–15. As in Figs. 7 and 12, the top panel shows the latent assignment of subjects to one of the three groups (attend position, attend height or contaminant). The next row shows the joint and marginal posterior distributions for

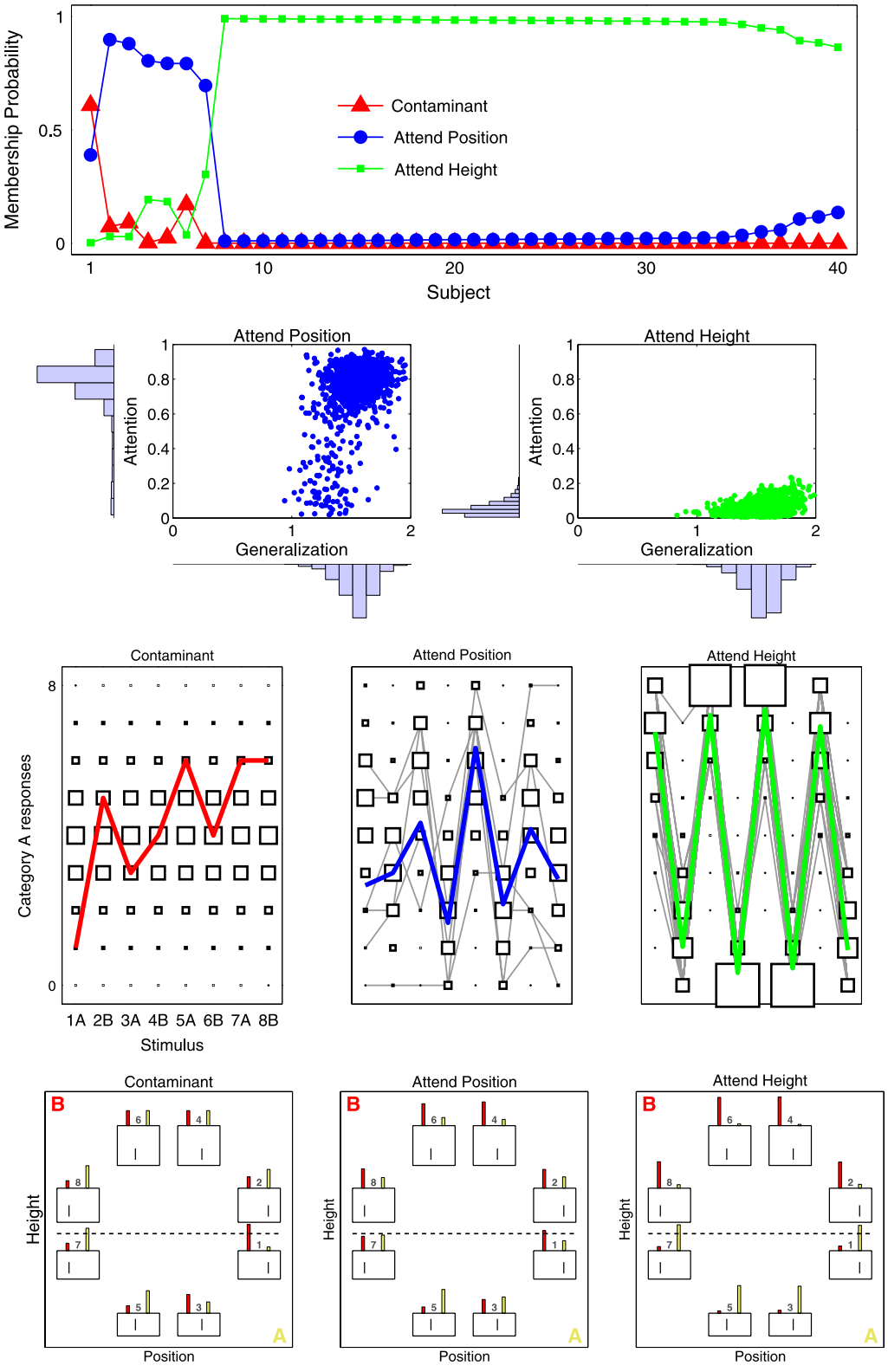


Fig. 14. Results from the selective attention analysis with individual differences on the "Filtration Height" condition, assuming three latent groups of subjects. The results show the latent assignment of subjects to groups, the posterior and posterior predictive distribution for the groups, and the interpretation of the groups in terms of the stimuli and category structure itself.

the hierarchical group attention parameter μ_g^w and generalization parameter ψ^c. A posterior predictive check of the model to the behavioral data is shown in the third row. The bottom row interprets the different category learning behavior of the groups. The original stimulus space and category structure is shown, with bars above the stimuli showing the average number of times each stimulus is placed in category A (right bar) and category B (left bar) by the members of that group.

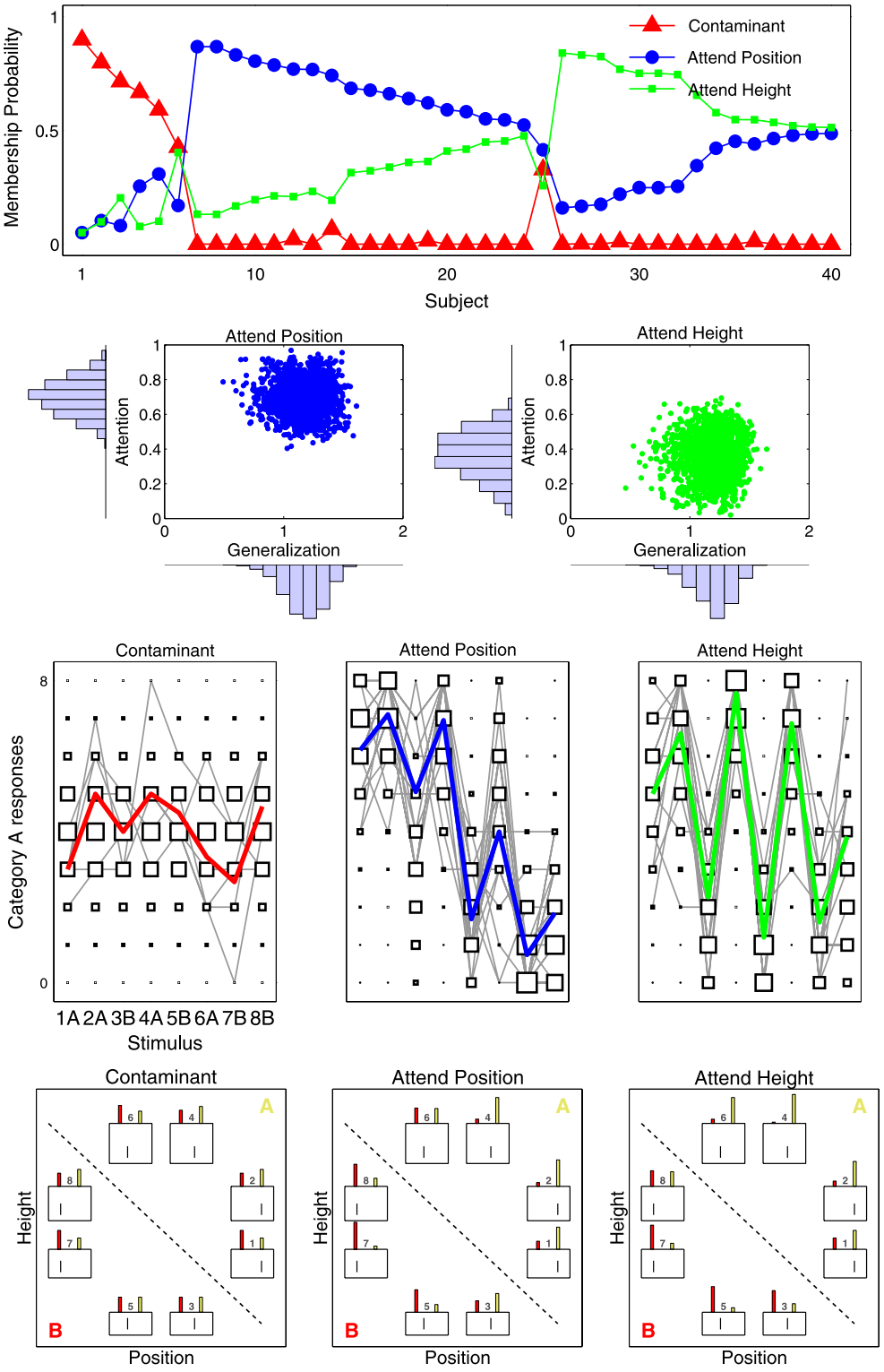


Fig. 15. Results from the selective attention analysis with individual differences on the "Condensation A" condition, assuming three latent groups of subjects. The results show the latent assignment of subjects to groups, the posterior and posterior predictive distribution for the groups, and the interpretation of the groups in terms of the stimuli and category structure itself.

Appendix B. Applying a four-group model to Kruschke's (1993) "Condensation B" condition

In this section, we provide some details on the application of a four-group model on the data from the "Condensation B" condition, allowing for the possibility of a group of subjects paying more equal attentions to both dimensions. We extended the three-group model with an additional attention-group, with the restriction $\mu_1^w > \mu_2^w > \mu_3^w$. This additional attention group has the same prior distribution as the existing groups, and all other parameters and priors are the same as in the three-group model. The results of this analysis are shown in Fig. 16.

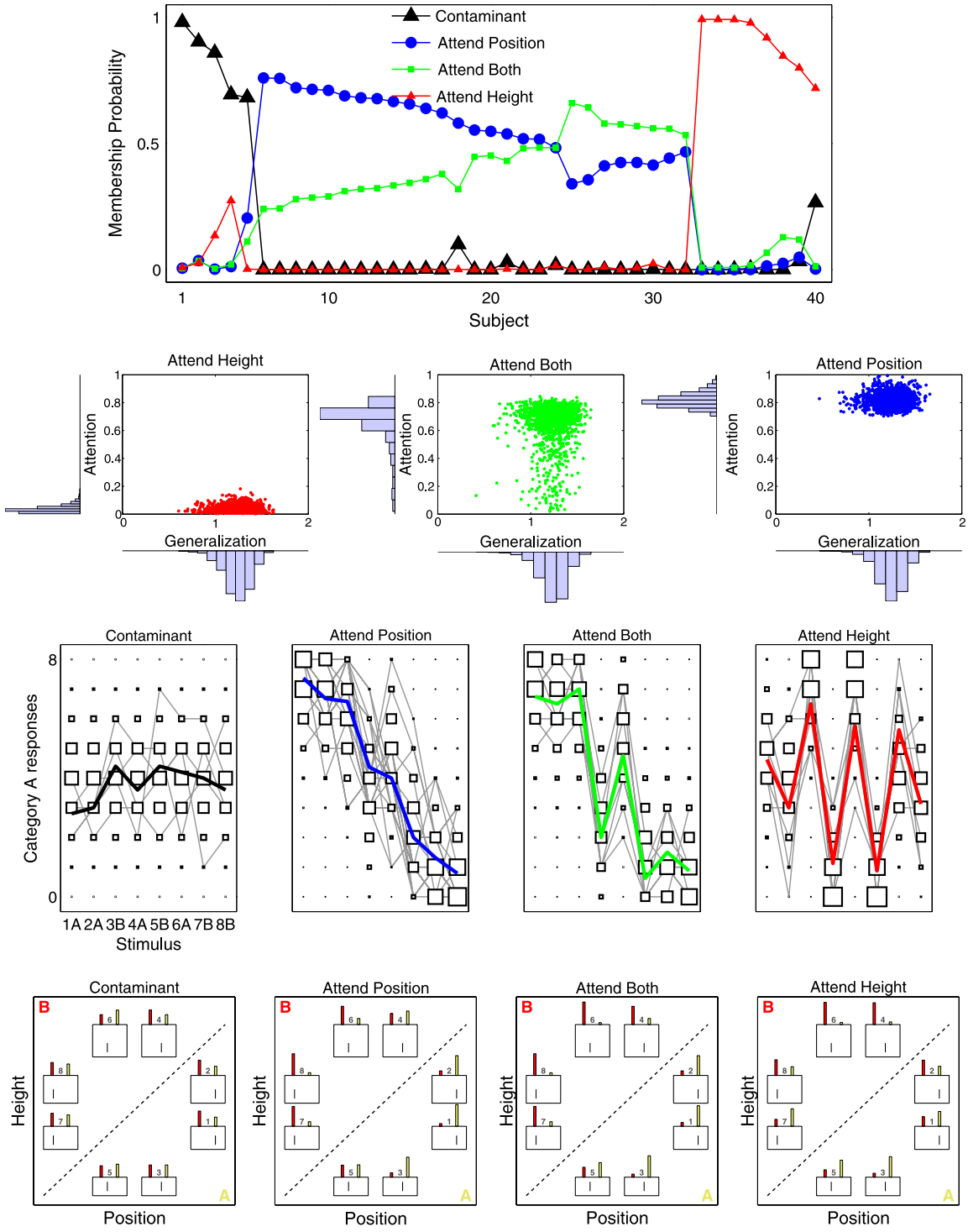


Fig. 16. Results from the selective attention analysis with individual differences on the Condensation B condition, assuming four latent groups of subjects. The results show the latent assignment of subjects to groups, the posterior and posterior predictive distribution for the groups, and the interpretation of the groups in terms of the stimuli and category structure itself.

Fig. 17 shows the prior and posterior distributions for the difference in μ^w for the attend both and attend position group. Comparing the prior and posterior densities in zero suggests the wisdom of including both μ_1^w and μ_2^w, but not μ_3^w. We did not use the Savage–Dickey ratio to formalize these observations into an estimate of a Bayes Factor, as it is not clear all relevant assumptions—especially technical conditions requiring nesting so that simpler models are identical to more complicated ones at particular parameter setting—for the Savage–Dickey approximation to be valid were met.

Appendix C. Comparing the MPM and the GCM with response determinism

In this section, we provide some details on the application of the model used in the second applications, where rather than being fixed to one, the response determinism parameter γ_k is estimated

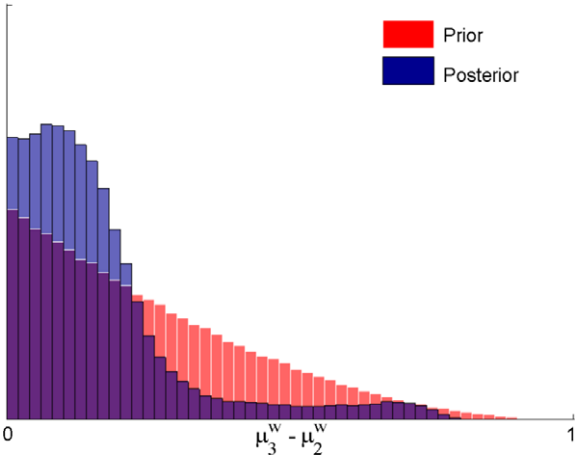


Fig. 17. Prior and posterior density for the difference in μ^w for the attend both and attend position group, for comparing the three- and four-group model.

for each subject. In particular, it is assigned the same distribution as the generalization parameter c (a gamma distribution, parameterized by its mode ψ^γ and standard deviation σ^γ). Results are shown in Fig. 18.

References

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

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

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*, 692–712.

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

Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, *18*, 500–549.

Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *115*, 155–174.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553.

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. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

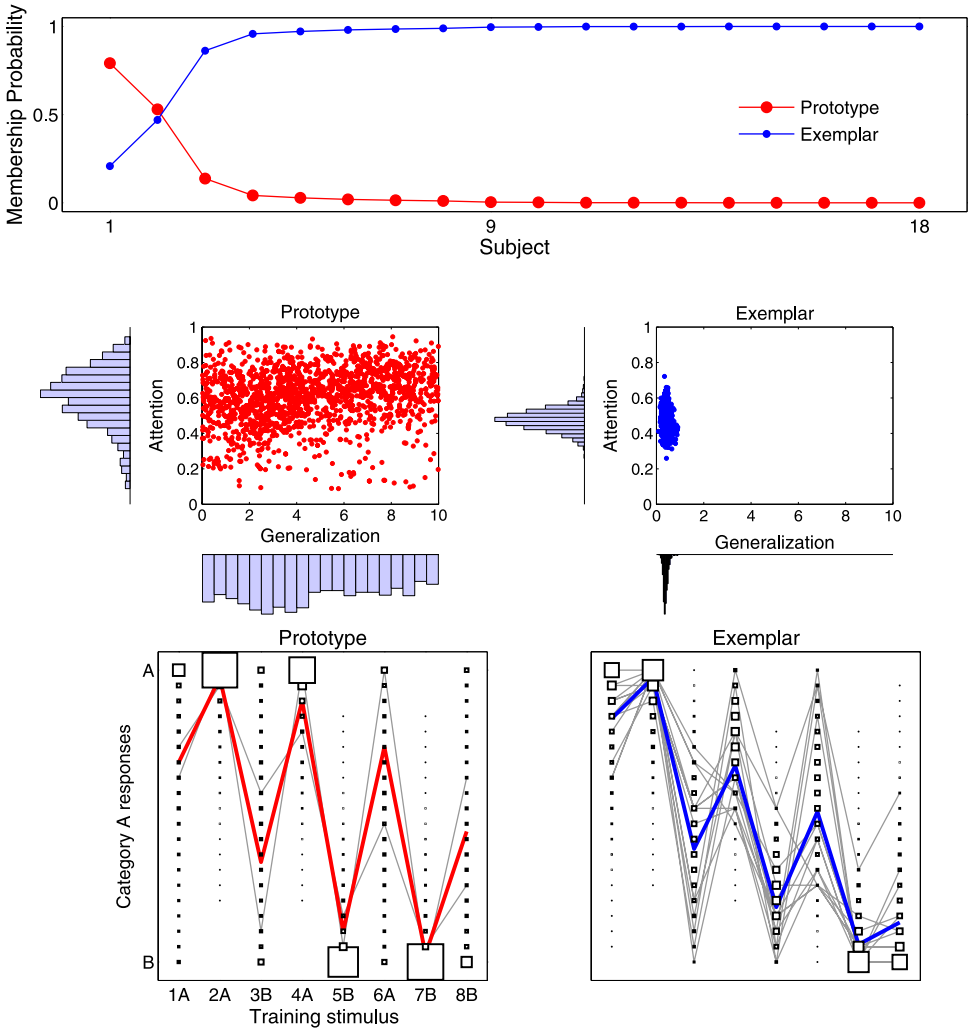


Fig. 18. Results from the analysis using a hierarchical mixture of the GCM with response determinism and MPM, showing the allocation of subjects to groups and the posterior and posterior predictive distributions for the groups.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press, Elsevier.
Kruschke, J. K. (2012). Gamma likelihood parameterized by MODE and SD http://doingbayesiandataanalysis.blogspot.be/2012/08/gamma-likelihood-parameterized-by-mode.html [blog post].
Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.
Lee, M. D. (2011a). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.
Lee, M. D. (2011b). In praise of ecumenical Bayes. *Behavioral and Brain Sciences*, *34*, 206–207.
Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities, and rules in category representation: an example of hierarchical Bayesian analysis. *Cognitive Science*, *32*, 1403–1424.
Lee, M., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press (Chapter 9).
Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605–621.
Lee, M. D., & Wetzels, R. (2010). Individual differences in attention during category learning. In R. Catrambone, & S. Ohlsson (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 387–392). Austin, TX: Cognitive Science Society.
Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.
Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.
Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 775–799.
Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, *51*, 85–98.
Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.
Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 104–114.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.
Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*, 279–290.
Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, *1*, 149–167.
Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 322–339.
Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940.
Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.
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*, *75*.
Shiffrin, R. M., Lee, M. D., Wagenmakers, E. J., & Kim, W. J. (2008). A survey of model evaluation approaches with a focus on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.
Shin, H. J., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148.
Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811.
Stephan, K., Penny, W., Moran, R., Den Ouden, H., Daunizeau, J., & Friston, K. (2010). Ten simple rules for dynamic causal modeling. *Neuroimage*, *49*, 3099–3109.
Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, *41*(4), 1111–1120.
Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056.
Vanpaemel, W., & Storms, G. (2008). Insearch of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749.
Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, *42*, 421–437.
Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.
Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*, 352–362.