# Representational shifts during category learning

**Wolf Vanpaemel (wolf.vanpaemel@psy.kuleuven.be)**
Department of Psychology, K.U. Leuven, Leuven, B-3000, Belgium

**Daniel J. Navarro (daniel.navarro@adelaide.edu.au)**
School of Psychology, University of Adelaide, Adelaide SA 5005, Australia

**Abstract**

Prototype and exemplar models form two extremes in a class of mixture model accounts of human category learning. This class of models allows flexible representations that can interpolate from simple prototypes to highly differentiated exemplar accounts. We apply one such framework to data that afford an insight into the nature of representational changes during category learning. While generally supporting the notion of a prototype-to-exemplar shift during learning, the detailed analysis suggests that the nature of the changes is considerably more complex than previous work suggests.

## Introduction

Classification tasks present people with stimuli and their accompanying category labels, and require label prediction for novel stimuli. Starting with seminal work in the 1970s (Rosch, 1978), the psychology of categorization has been assumed to be best thought of in terms of a kind of "family resemblance". For example, prototype theories (Reed, 1972) represent a category using a single prototypical stimulus, which need not necessarily correspond to a real object. Similarity to a category is defined as similarity to the prototype. In contrast, exemplar theories (Medin & Schaffer, 1978; Nosofsky, 1986) represent a category as the set of all of its previously observed members (its exemplars), and the category similarity as the aggregated similarity to the exemplars. More recently, it has been argued (Love, Medin, & Gureckis, 2004; Anderson, 1991; Rosseel, 2002; Vanpaemel, Storms, & Ons, 2005) that exemplar representations and prototype representations constitute the two extremes of a spectrum of models. Although different authors have adopted slightly different formalisms to advance their viewpoint, they share the common view that human conceptual structure is sufficiently flexible to adopt simple, highly abstracted "prototype-like" representations at times, but can also accommodate highly differentiated "exemplar-like" representations at others.

An elegant experimental test of this idea was conducted by Smith and Minda (1998), involving a series of categorization experiments. In each experiment, prototype and exemplar accounts were contrasted at different stages of the learning process, to see which model provided a superior account. Although the overall pattern of performance across experiments is complex, the general finding was that exemplar models tend to be favored late in learning, with the possibility that prototype models are favored early in learning (but see, e.g., Nosofsky & Zaki, 2002 for a contrasting view). However, one drawback to the study is that only prototype and exemplar models were evaluated, leaving the vast majority of potential category representations unexplored. Since these data provide a natural testing ground for exploring the flexibility of category representations, the current paper undertakes precisely this analysis. The plan of the paper is as follows. We first introduce the formal modeling framework, and then discuss the data provided by Smith and Minda (1998). We then analyze two key data sets from this paper, looking first to extract an explicit model for the individual differences in performance (Webb & Lee, 2004) before analyzing the data set using the Varying Abstraction Model (VAM) introduced by Vanpaemel et al. (2005) which provides a much richer set of potential category representations.

## Treating Categories as Mixtures

In most theories of human concepts (e.g., Nosofsky, 1986; Love et al., 2004), people are assumed to have some internal representation of a category $C$ that provides a probability distribution $p(\cdot \mid C)$ over possible objects in the world. When translated into formal models, it is typical to assume that this distribution is a mixture of several component densities, generally on the implicit assumption that each "element" of the category representation constitutes a psychologically distinct component of the category.

### The General Approach

We begin by introducing the general approach, in which the psychological representation of a category is modeled as a mixture of simpler components. If the internal representation contains $q$ components, the probability assigned by category $C$ to the $i$th stimulus $x_i$ is given:

$$p(x_i \mid C) = \sum_{j=1}^{q} w_j \; p(x_i \mid j), \tag{1}$$

where $p(x_i \mid j)$ is the density assigned by component $j$ to the $i$th stimulus $x_i$, and $w_j$ weights the contribution made by each component. The general mixture formulation in Equation 1 is often translated into a specific statistical model by applying the exponential law for generalization developed by Shepard (1987). In view of this law, it is natural to treat each of the component distributions as an exponential density,

$$p(x_i \mid j) \propto \; e^{-\lambda d(x_i, \mu_j)}, \tag{2}$$

where $\mu_j$ denotes the internal representation of the $j$th component to the category, $\lambda$ is a scaling parameter that governs the specificity of the generalization away from that representation, and $d(\cdot, \cdot)$ describes a *psychological distance* function. When applying such models, it is typical to assume

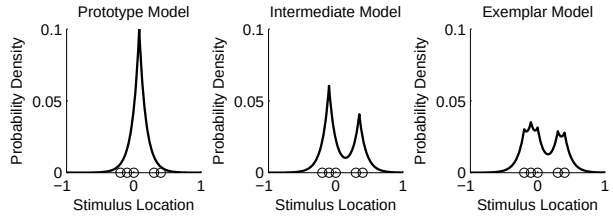


Figure 1: Category densities for the a one-dimensional category consisting of items located at $\mathbf{x} = (-0.2, -0.1, 0, 0.3, 0.4)$ and $\lambda = 10$. The density on the left is produced by the prototype model, and the one on the right by the exemplar model. The middle density belongs to a model that groups the three stimuli on the left and the two stimuli on the right.

that $\lambda$ has the same value for all components. While there are many possibilities for a psychological distance function, a common choice is to use one of the attention-weighted Minkowski $r$ metrics,

$$d(x_i, \mu_j) = \left( \sum_{k=1}^{m} a_k \left| x_{ik} - \mu_{jk} \right|^r \right)^{\frac{1}{r}}, \qquad (3)$$

where $a_k$ represents the proportion of attention applied to the $k$th stimulus dimension. To provide a concrete illustration of the approach, Figure 1 shows three different mixture representations of the same category, corresponding to prototypes, exemplars and an intermediate case. In order to describe human behavior in a two-alternative forced-choice task between categories $A$ and $B$, it is typical to apply a standard choice rule,

$$P(x_i \in A \,|\, x_i) = \frac{p(x_i \,|\, A)}{p(x_i \,|\, A) + p(x_i \,|\, B)}. \qquad (4)$$

## A Simplified Framework

The general mixture model formulation is broad enough to cover a range of approaches. However, the simplest proposal is perhaps the one introduced by Vanpaemel et al. (2005). Unlike some approaches (e.g., Love et al., 2004; Anderson, 1991) it makes no particular assumptions about how human learning takes place, and it is much more constrained than the mixture model formulation adopted by Rosseel (2002) in terms of how the weights $w_j$ and probabilities $p(x_i \,|\, j)$ are assigned. While the simplicity of this arrangement does not necessarily make it a superior cognitive model, it provides a very clean framework in which to ask questions about representational structure without introducing any additional psychological principles that could confound the analysis.

In Vanpaemel et al.'s (2005) Varying Abstraction Model (VAM) the mixture components $\mu_j$ and $w_j$ that might otherwise be treated as free parameters are fully determined by a specific partition of category members into a set of clusters. Each cluster implies a specific psychological representation $\mu_j$, that can be viewed as a kind of sub-prototype. Thus, a category of $n$ exemplars represented in terms of $q$ clusters can be described using the vector $\mathbf{c} = (c_1, \ldots, c_n)$, where

Table 1: Stimulus representations for the non-linearly separable categories used by Smith and Minda (1998), experiments 2 (panel a) and 3 (panel b). In both panels each column corresponds to a feature (i.e., letter), and each row to a stimulus.

(a)

| | | | | | | |
|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 0 | 0 |
| | 0 | 1 | 0 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 0 | 0 | 1 | 0 |
| | 0 | 0 | 0 | 0 | 0 | 1 |
| | 1 | 1 | 1 | 1 | 0 | 1 |
| B | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 0 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 0 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 0 | 1 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 0 |
| | 0 | 0 | 0 | 1 | 0 | 0 |

(b)

| | | | | |
|---|---|---|---|---|
| A | 0 | 0 | 0 | 1 |
| | 0 | 1 | 0 | 0 |
| | 1 | 0 | 1 | 1 |
| | 0 | 0 | 0 | 0 |
| B | 1 | 0 | 0 | 0 |
| | 1 | 0 | 1 | 0 |
| | 1 | 1 | 1 | 1 |
| | 0 | 1 | 1 | 1 |

$c_i \in 1, 2, \ldots, q$ indexes the representational cluster to which the $i$th stimulus belongs. As such $q$ can be interpreted as the *level of abstraction* of the category representation. In the constrained framework, the representation of the $j$th cluster is taken to be the average of the representations of its constituent stimuli. Thus, $\mu_{jk} = (1/n_j) \sum_{i|c_i=j} x_{ik}$ where $n_j$ denotes the number of stimuli that belong to cluster $j$. Applying the same logic, the mixture weights are constrained to reflect the proportion $w_j = n_j/n$ of category members that fall in the cluster.

In this framework, the partitions $\mathbf{c}_A$ and $\mathbf{c}_B$ of categories A and B define a particular model for these categories, with an overall level of abstraction $q_A + q_B$. The model's free parameters are the attention weights $a_k$ and the specificity $\lambda$ (the metric $r$ is taken to be a property of the stimulus space itself, and is held fixed). This formulation ensures that all models have the same number of free parameters. The standard prototype and exemplar models are special cases of the VAM: a category represented using a single-cluster partition $\mathbf{c} = (1, 1, \ldots, 1)$ has a prototype representation, and a category represented using a cluster for every stimulus $\mathbf{c} = (1, 2, \ldots, n)$ has an exemplar representation. In between these two extremes, however, lie a wealth of infrequently-explored representational possibilities.

## Looking for Representational Shifts

It has been argued (Smith & Minda, 1998) that much of the categorization literature is overly-reliant on experiments that provide participants with a great deal of training before attempting to measure the structure of their conceptual representation. With that in mind, Smith and Minda (1998) conducted a series of experiments aimed to show that the category representation changes as learning progresses. In this paper, we focus on their experiments 2 and 3, involving non-linearly separable categories. For both experiments, the stimuli took the form of pronounceable nonsense words (e.g., *gafuzi*, *daki*), on the assumption that each letter corresponds to a feature. In experiment 2, both categories consisted of seven exemplars possessing six features each, and were designed to be well-differentiated category structures even despite the

Table 2: The series of representations used to build models for experiment 2. Each column corresponds to a partition, either of the category A exemplars or the category B exemplars.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\mathbf{c}_A$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| | 1 | 1 | 1 | 1 | 1 | 2 | 3 |
| | 1 | 1 | 1 | 1 | 2 | 3 | 4 |
| | 1 | 1 | 1 | 2 | 3 | 4 | 5 |
| | 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $\mathbf{c}_B$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| | 1 | 1 | 1 | 1 | 1 | 2 | 3 |
| | 1 | 1 | 1 | 1 | 2 | 3 | 4 |
| | 1 | 1 | 1 | 2 | 3 | 4 | 5 |
| | 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

fact that both categories contain obvious exception items: the logical structure of the categories is shown in Table 1(a). In contrast, experiment 3 used the smaller, less differentiated category structures shown in Table 1(b). Both experiments involved 16 participants who were presented with 560 trials, divided into 10 segments of 56 trials each. On each trial, one of the stimuli was presented and the participant was asked to classify it as a member of category $A$ or category $B$. Feedback was provided after each trial. Smith and Minda (1998) analyzed the data by fitting exemplar and prototype model to each segment, in order to find evidence for representational transitions during learning. They concluded that a shift had occurred during experiment 2, but not during experiment 3.

Although the idea of a representational change is in agreement with the spirit of the VAM, Smith and Minda (1998) only considered prototype and exemplar models, which makes it difficult to trace out these changes in any detail. To address this, in the remainder of the paper, we reanalyze the data from these experiments using the VAM, exploring the full class of potential category representations. For the smaller category, we fit all $15 \times 15 = 225$ category models, but for the larger category the $877 \times 877 = 769129$ models are too many to work with, particularly since model fitting is required for 10 different trial segments. Accordingly, in this case we used only a smaller set of $7 \times 7 = 49$ models, with one model at each level of abstraction (i.e., number of clusters) for each category. These 49 models were found by applying a simple average-link clustering procedure to the stimulus representations in Table 1(a), and are shown in Table 2. Each model was fit to the observed classification accuracies using maximum likelihood estimation, which means that the values of the free parameters maximizing the likelihood of observing the data were determined. The free parameters were the scaling parameter $\lambda$ and five or three attention weights $a_k$ in experiment 2 or 3 respectively.

## A Varying Abstraction Analysis

Our analysis involves three stages. Firstly, we analyze the individual differences in the data, in order to make sure that we can draw sensible conclusions about what particular participants were doing. Secondly, we reproduce Smith and Minda's (1998) prototype-to-exemplar result within the context of the VAM. Finally, we use this framework to develop a more detailed picture of the nature of the representational changes.

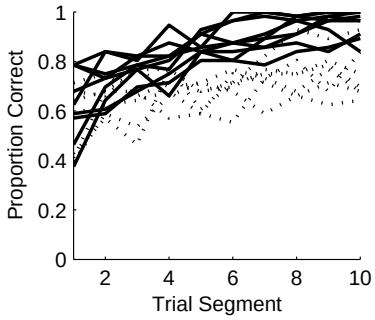


Figure 2: Empirical learning curves for all 16 participants in the experiment 2. The data segregate naturally into two groups.

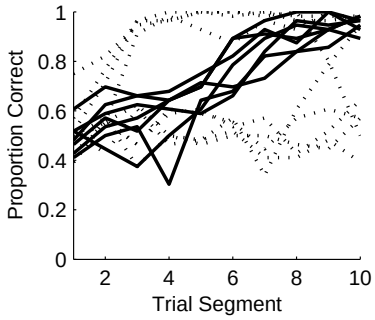


Figure 3: Empirical learning curves for all 16 participants in the experiment 3. The data segregate naturally into three groups.

### Part 1: Individual Differences

Recent work (Webb & Lee, 2004) has emphasized the fact that category learning tasks show strong individual differences, and highlighted the fact that averaging across people may lead to substantial distortions. In Smith and Minda (1998), this problem was solved by fitting models to each participant independently. However, in addition to inflating the risks of overfitting, this approach is unwieldy and time-consuming. A faster and more robust approach emphasizes both the similarities and differences between people, and seeks to find groups of participants with similar patterns of performance.

To apply this idea, we took the learning curves for each participant, and partitioned them into meaningful groups. To do so, we applied the Minimum Description Length (MDL) clustering technique pioneered by Kontkanen, Myllymäki, Buntine, Rissanen, and Tirri (2005) and extended to learning curves by Navarro and Lee (2005). This method, which is based on information theoretic ideas, assigns two observations to the same group only if this allows a better compression of the overall data set. Although the technical details are complicated (see Grünwald, 1998, for details on MDL), what matters for the current purposes is that the approach allows us to find a statistically-optimal method of grouping

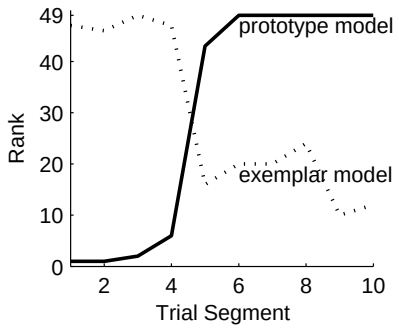


Figure 4: The rank of the prototype and exemplar models at each segment for experiment 2.

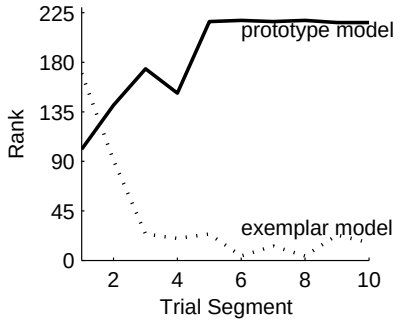


Figure 5: The rank of the prototype and exemplar models at each segment for experiment 3.

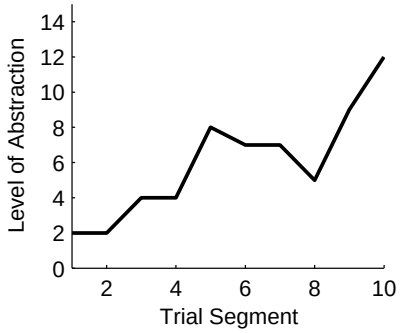


Figure 6: The level of abstraction of the best model at different stages of learning during experiment 2.

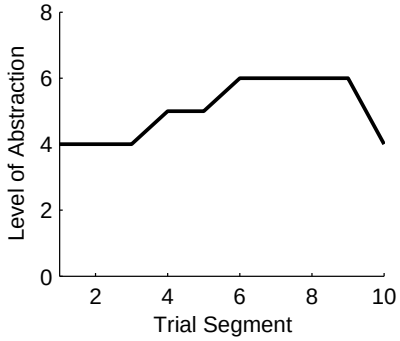


Figure 7: The level of abstraction of the best model at different stages of learning during experiment 3.

people's data. Applying this method, we were able to extract three strikingly different types of performance for the experiment 3 data (Figure 3) and two probably-distinct groups for the experiment 2 data (Figure 2). Due to space constraints, we restrict the analyses in this paper to the largest groups, indicated by the solid lines in both figures.

## Part 2: Comparing Prototypes to Exemplars

Our approach to analyzing the prototype-to-exemplar shift differs from Smith and Minda's in several ways. Firstly, we fit a much broader range of models to the data (225 for experiment 3, and 49 for experiment 2), and used a maximum likelihood estimation rather than the least squares approach adopted in the original paper. Secondly, we fit data that were aggregated in an optimal fashion, as discussed in the previous section. Finally, unlike Smith and Minda (1998), we did not include a "guessing parameter".

Despite the very substantial differences in representational possibilities considered, the choice of loss function, individual differences, and guessing behavior, the basic pattern of exemplar and prototype performance remains intact. This is most naturally shown by looking at how well the two models fared at different stages of learning, when compared to all (225 or 49) models under consideration, as illustrated in Figures 4 and 5. In one key respect, this pattern is far more compelling in the current analysis than in the original: in Smith and Minda (1998), the exemplar and prototype models are evaluated without consideration of the other representational possibilities. Happily, when a broad spectrum of representational possibilities are included to provide an appropriate context, the substantive finding remains unchanged. In experiment 2, there is an early advantage for the prototype model, and a late advantage for the exemplar model, with the changeover point located between segments four and five. In experiment 3, only during the first segment does the prototype model outrank the exemplar model, and the extent of the exemplar advantage grows throughout the experiment.

The dramatic shift in the relative fortunes of the prototype and exemplar models illustrated in Figure 4 and 5 suggests that some kind of representational shift has taken place during the learning process, particularly with respect to the larger category structure used in experiment 2. This was essentially the conclusion in Smith and Minda (1998), but our application of the VAM allows us to gain further insight in the nature of the representational shift, a topic which we turn to in the next section.

## Part 3: A Richer View of Representational Change

The analyses reported by Smith and Minda (1998) and shown in our Figures 4 and 5 imply that the representational shift involves a jump from a prototype representation to an exemplar representation. However, this is somewhat misleading, in the sense that the shift appears to be considerably more complex. To illustrate this, we classified every model in terms of its overall level of abstraction (i.e., $q_A + q_B$).

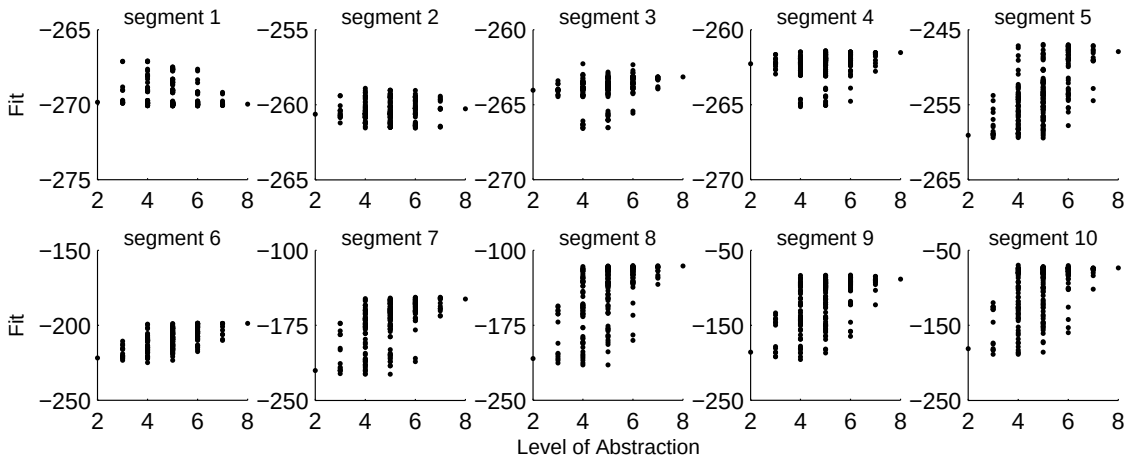


Figure 8: Scatterplot of the fit of all the 225 models (y-axes) versus the level of abstraction $q_A + q_B$ (x-axes) at each segment for experiment 3. Over the first three segments, the best models tend to be more prototype-like (though the prototype model itself performs poorly). As learning progresses, the profile of good models shifts, and from segments 5-10 it is clear that the best models all have a differentiated, exemplar-like structure.

Figures 6 and 7 display, at all 10 segments, the level of abstraction of the model that best accounts for the behaviorial data. For both experiments, the level of abstraction of the best model changes systematically across trials. Experiment 2 in particular shows a steady progression from prototypes to exemplars. In the first two trial segments, the very best model is the prototype model, and in the last segment the best model is very nearly an exemplar model (having an level of abstraction of 12, out of a maximum possible of 14). However, the transition here is steady, very nearly linear. In other words, while the exemplar model improves so dramatically against the prototype model that the shift looks discrete (as in Figure 4), the inclusion of a broader class of models suggests that the change is somewhat more gradual across the best fitting models (as in Figure 6). Moreover, examination of Figure 7 suggests that a small shift takes place in experiment 3, which was not evident in the original analysis.

The analysis presented above suggests a representational shift in which conceptual structures smoothly move from prototypes to exemplars via a range of intermediate models. To get a more detailed picture of the pattern of changing model fits during the course experiment 3, instead of looking at the best fitting models only, we can look at all 225 possible representational models. This is illustrated in Figure 8 which shows 10 scatterplots, one for each trial segment. Each plot displays the level of abstraction and the data fit for each of all the 225 possible representational models. In the first four segments, highly abstract representations are able to account for the data relatively well (though notably the prototype model itself fits poorly), while very detailed exemplar-like representations perform poorly. In contrast, from segment five onwards the profile reverses: in order to provide a good account of human performance, the category representation requires at least four clusters across the two categories. Although not shown, a similar change occurs for the 49 models analyzed in experiment 2, also at segment five.

Another detailed picture of the shift is shown in Figure 9, this time looking at all 49 models used in experiment 2. In this figure, the overall level of abstraction $q_A + q_B$ is split in its two constituent parts. Each plot displays the number of clusters required to represent category A, the number required to represent category B, and the data fit for each of the 49 models considered. The prototype model sits at the $(1, 1)$ level of abstraction, while the exemplar model is at $(7, 7)$. Figure 9 shows the maximized log-likelihood for each model at each level of abstraction per category, at each stage of experiment 2. It shows that, like in experiment 2, an abrupt shift takes place after segment four. Until that segment, models with a highly abstract representation are clearly the best, but from segment four onwards the models with a more detailed representation are dominant.

## Discussion

Neither prototypes nor exemplars appear to provide a sufficient account of human category learning, at least for large categories (Smith & Minda, 1998), since neither model accounts for changes in representational structure, and there are very clear signs that these changes occur in empirical data. On top of this, when we consider the fact that prototypes and exemplars are two extremes in the class of mixture representations (Rosseel, 2002), a prototype-to-exemplar shift should be expected to pass through the kind of intermediate models that are encompassed by VAM introduced by Vanpaemel et al. (2005). To demonstrate that this does in fact occur, we reanalyzed the data from two experiments in Smith and Minda (1998), replicating their results (Figures 4 and 5). Our analysis indicated that such changes are somewhat more complex than previously suggested. Firstly, unlike Smith and Minda (1998), we are able to find evidence of a small shift in experiment 3, as well as the large changes in experiment 2. Also,

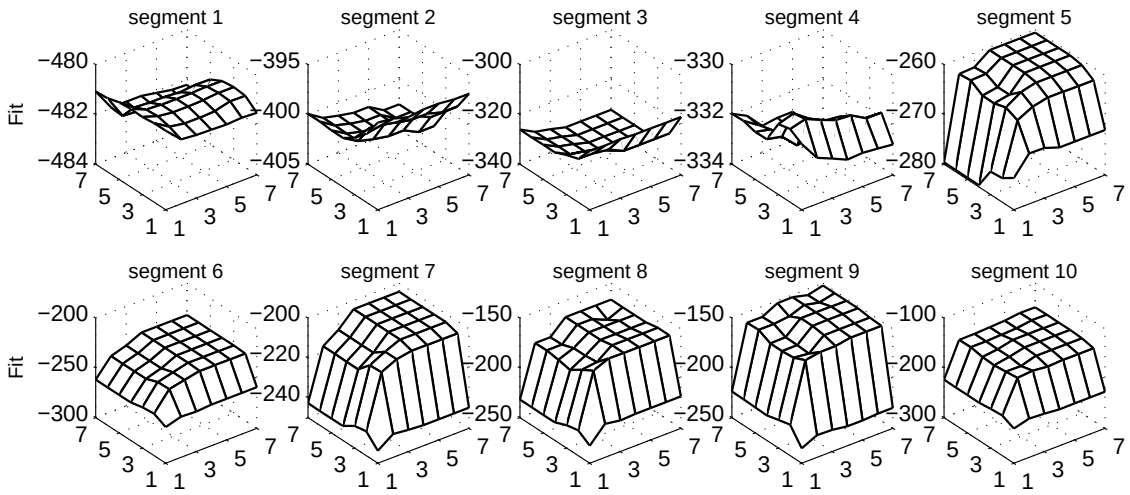


Figure 9: Fit of all the 49 models at each level of abstraction for both categories at each segment in experiment 2. In each subplot, the dependent variable is the model fit, measured in terms of the log-likelihood. The independent variables are the levels of abstraction for each category, $q_A$ and $q_B$. As is clear from inspection of the plot, there is a fairly sharp change in the profile of models at around segment 4-5.

although the overall level of abstraction of the best-fitting models moves smoothly from abstract, prototype-like models towards differentiated, exemplar-like models (Figures 6 and 7), when we look at the performance of all possible models (Figures 8 and 9) there appear to be some much sharper transitions, as the performance of previously good models can deteriorate rapidly. In light of these findings, it appears that the current trend toward developing and applying mixture models for categorization can provide useful insights, by allowing us to trace changes in representation in more detail than previously possible.

## Acknowledgments

WV was supported by the Research Council of the University of Leuven (IDO/02/004), and DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). We would like to thank John Paul Minda for providing us with the data, and the reviewers for detailed and helpful comments.

## References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Grünwald, P. (1998). *The minimum description length principle and reasoning under uncertainty*. Unpublished doctoral dissertation, University of Amsterdam.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J., & Tirri, H. (2005). An MDL framework for data clustering. In *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

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

Navarro, D. J., & Lee, M. D. (2005). An application of minimum description length clustering to partitioning learning curves. In B. Petrov & B. Csaki (Eds.), *2005 ieee international symposium on information theory* (p. 587-591). Piscataway, NJ: IEEE.

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

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*, 392–407.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178–210.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 2277–2282). Mahwah, NJ: Lawrence Erlbaum.

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* (p. 1440-1445). Mahwah, NJ: Erlbaum.