Running head: SECONDARY GENERALIZATION IN CATEGORIZATION

Secondary generalization in categorization: an exemplar-based account

Yves Rosseel & Maarten De Schryver

Department of Data Analysis, Ghent University

Correspondence address: Yves Rosseel, Department of Data Analysis, Ghent University,

Henri Dunantlaan 1, B-9000 Gent (Belgium).

Email: Yves.Rosseel@UGent.be

Abstract

The parallel rule activation and rule synthesis (PRAS) model is a computational model for generalization in category learning, proposed by Vandierendonck (1995). An important concept underlying the PRAS model is the distinction between primary and secondary generalization. In Vandierendonck (1995), an empirical study is reported that provides support for the concept of secondary generalization. In this paper, we re-analyze the data reported by Vandierendonck (1995) by fitting three different variants of the Generalized Context Model (GCM) which do not rely on secondary generalization. Although some of the GCM variants outperformed the PRAS model in terms of global fit, they all have difficulty in providing a qualitatively good fit of a specific critical pattern.

Secondary generalization in categorization: an exemplar-based account

Perhaps the most important work of André Vandierendonck in the field of categorization and concept learning is his paper entitled "A parallel rule activation and rule synthesis model for generalization in category learning" (Vandierendonck, 1995). In this paper, a computational model of category learning is proposed that –certainly at its time– was unlike any other model in the categorization field. The parallel rule activation and rule synthesis (PRAS) model is a production model, similar to Anderson's ACT model (Anderson, 1978, 1983), where information is stored in a special if-then format called production rules. However, during the late eighties and early nineties, most of the computational models (both in categorization and elsewhere) were so-called *connectionist* models. Indeed, in this period, several highly influential connectionist models had been published in the categorization literature (Gluck & Bower, 1988; Kruschke, 1992). There was a wide-spread sentiment during that period among many modelers (including the first author of this paper at that time) that connectionist models were the future. Other types of models, including the PRAS model, did not receive so much attention. Perhaps, this is the reason why some important ideas underlying the PRAS model have been somewhat ignored in the categorization literature. One such idea that I will focus on in this paper is the idea of *secondary generalization*. Briefly, secondary generalization is generalization that stems from abstract information, while primary generalization is generalization that stems from exemplar information. In Vandierendonck (1995), an empirical study was presented that supported the idea of secondary generalization. The empirical evidence was fairly convincing, and still poses a challenge for models that only rely on primary generalization.

In this paper, we will re-analyze the data reported by Vandierendonck (1995). Because we will often refer to this paper and its dataset, we will refer to the

Vandierendonck (1995) paper as the ‘PRAS paper’, and the dataset in that paper will be referred to as the ‘PRAS dataset’. The goal of this paper is to give an exemplar-based account of the results in the PRAS paper. We will push the exemplar models to the limit (and perhaps even over the limit) in an attempt to fit the PRAS data without relying (explicitly) on secondary generalization. If we succeed, the exemplar theorists may cry victory once again. If we fail, the empirical study of the PRAS paper will stand as one of a few interesting exceptions where the exemplar theory falls short, and categorization modelers should consider the implications for their models.

The paper is organized as follows. First, we will briefly review exemplar and abstraction based models in categorization, including hybrid models like the PRAS model. Next, we discuss the concept of secondary generalization and describe the empirical study that was reported in the PRAS paper. We then give an overview of the exemplar models that we will fit to this dataset. Finally, we will reflect on the results of our model fitting experiment and their implications for old and new models of categorization.

Exemplars, Abstraction and the PRAS model

Individual members of a category are called exemplars. There is a strong tradition in the categorization literature which assumes that a category is simply represented by a set of exemplars which are known to belong to that category (Hayes-Roth & Hayes-Roth, 1977; Medin & Schaffer, 1978; Nosofsky, 1984; Estes, 1986). Category learning then is merely a matter of storing these exemplars if they present themselves as a member of the category (hence, implicitly assuming that we can only learn from 'labeled' exemplars). Many different flavors of exemplar theory have been proposed, but importantly, at the heart of every model based on exemplar theory is the idea that no abstraction takes place during category learning. Categorizing a new target stimulus is solely based on the set of stored exemplars.

A different perspective is taken by so-called *abstraction models*. In these models, category level information is inferred from the observed exemplars by some sort of mechanism for abstraction. For example, in *prototype models*, a category is represented by a single prototype, a special (possibly unobserved) exemplar that captures the central tendency of the individual exemplars that belong to that category (Homa, 1984; Reed, 1972; Posner & Keele, 1968; Minda & Smith, 2001; Smith & Minda, 2002). After learning, the prototype has replaced the individual exemplars and forms the only basis for categorizing future stimuli.

A second type of abstraction are *rules*, which can often be verbally expressed (Trabasso & Bower, 1968; Bourne, 1982). Rules can be one dimensional, as for example "all red objects belong to category A", but multidimensional rules can be constructed as well. In a rule-based model of categorization, the rules have replaced the exemplars, and the categorization of new stimuli is solely based on these rules. During learning, new rules can be constructed, and existing rules can be adapted, capturing the common features of exemplars belonging to the same category.

Inevitably, hybrid models have been proposed where the representation of a category can consist of both exemplar information and abstract information. Models that combine exemplars and prototypes have been proposed by Medin, Altom, and Murphy (1984) and Busemeyer, Dewey, and Medin (1984). Hybrid models involving both exemplars and rules have been proposed by Erickson and Kruschke (1998) and Nosofsky, Palmeri, and McKinley (1994), among others.

The PRAS model of Vandierendonck (1995) is also a hybrid model. In the PRAS model, both exemplar-based and rule-based information can be used to represent a single category. Both types of information are stored by means of production rules (Anderson, 1983). In the context of categorization, production rules can be considered as if-then statements where the if-part contains a description of an exemplar or exemplar features, and where the then-part implies a category assignment. For example, a production rule for classifying animals as birds or non-birds could be:

IF	the animal has feathers
AND	the animal has wings
THEN	classify it as a bird

The beauty of a production system is that the condition part (the if-part) can contain either a highly specific description of a unique exemplar, or it can describe a set of features that apply to a larger set of exemplars (for example "has wings"). By combining different types of information for the condition part, a production system is an ideal environment for building a hybrid model of categorization, where both exemplar-level and more abstract information can be stored in a similar representational format. The representation of a single category may consist of many (possibly conflicting) production rules, at different levels of abstraction. When a target stimulus must be classified, all these production rules may become activated (hence the name *parallel rule activation*), albeit with different strengths. Each production rule provides evidence for a certain category. When a category assigment must be made, the production system collects the accumulated evidence for each of the competing categories. Finally, a decision rule converts the evidence for the different categories into a category decision, often in a probabilistic manner.

A vital feature of production systems is that they are capable of making inferences based on experience. Borrowing the example used in Vandierendonck (1995): if a production system learns that a specific brown animal is a horse, and it learns that a specific black animal is also a horse, it may infer that a horse can have any color. Combining information of two production rules into a new production rule is called *rule synthesis*. Inferring that a horse can have any color is of course a rather crude (over)generalization. In the PRAS model, a more subtle type of generalization is used. For example, after observing both a black and a brown horse, the PRAS model would typically infer that the color of a horse is somewhere between brown and black. The exact range of the generalization is governed by a free parameter in the model (i.e. the ρ parameter, see page 445 in the PRAS paper). This type of generalization is not confined to a single dimension. If several dimensions are involved, the PRAS model assumes that a rectangular area in the psychological space is constructed in between the exemplars over which the generalization takes place (see Figure 1 in the PRAS paper).

To make things more concrete, suppose that our exemplars vary in only two dimensions (as will be the case in the empirical example below). A typical production rule in the PRAS model has the following form:

$$\begin{array}{ll} \text{IF} & x \in [a_{1.\min}, a_{1.\max}] \\ \text{AND} & x \in [a_{2.\min}, a_{2.\max}] \\ \text{THEN} & \text{classify } x \text{ in category } A \end{array}$$

where x is a new target stimulus, and $a_{1.\min}$ and $a_{1.\max}$ are the lower and upper end of a range in the psychological space along the first dimension. Note that if the values of $a_{1.\min}$

and $a_{1.\max}$ are minus and plus infinity respectively, the condition is always fulfilled. On the other hand, if $a_{m.\min} = a_{m.\max}$ for every dimension m, the range is confined to a single point in the psychological space. This is how exemplars are represented in the PRAS model. The example illustrates an important feature of the PRAS model: unlike other hybrid models that combine exemplar and rule information, there is no separate system or submodel for the exemplar part and the abstracted (rule) part of the system. Instead, the representation of exemplar information and abstracted information forms a continuum. On this continuum, examplars are represented by zero-range condition parts. By widening the ranges over one or more dimensions, more general (and hence more abstract) information is gradually formed, all within the same representational format.

Primary and Secondary generalization and the PRAS dataset

Once the PRAS model has been trained to categorize a set of training exemplars, how does the model proceed to categorize a new target stimulus? Suppose for simplicity that the representation of a category currently consists of a single production rule where the condition part corresponds to a specific examplar:

IF	$x \in [6, 6]$ (first dimension)
AND	$x \in [5, 5]$ (second dimension)
THEN	classify x in category A

where the coordinates $(6, 5)$ correspond with the location of a stored exemplar in a two-dimensional psychological space. What happens if a new stimulus with coordinates, say, $(5, 4)$ is presented to the system? The coordinates do not perfectly match the stored exemplar in the production rule. However, the coordinates are fairly close together, and therefore, the target stimulus and the stored exemplar are perceived to be rather similar. A fundamental observation in the (category) learning literature is that similar stimuli lead to similar responses. This is known as *generalization* and its properties have been studied extensively in the classical conditioning literature (Mostofsky, 1965; Ghirlanda & Enquist, 2003). In our example, generalization would suggest that a stimulus with coordinates $(5, 4)$ might still trigger the response part of the production rule. But how do we define 'similar'? It seems natural that stimuli that are further apart in the psychological space are less similar than stimuli that are closer together. Shepard (1957) suggested that similarity between two exemplars i and j is an exponential decay function of their psychological distance:

$$\eta_{ij} = \exp(-c * d_{ij})$$

where c determines the steepness of the exponential curve. This relationship has been empirically observed in so many different studies that the relationship was coined the *universal law of generalization* (Shepard, 1987). The distance measure in this formula is often defined by the weighted Minkowski distance:

$$d_{ij} = \left(\sum_{m=1}^{M} w_m \left| x_{im} - x_{jm} \right|^r \right)^{1/r}$$

where M is the number of dimensions, x_{im} and x_{jm} are the coordinates in psychological space along the mth dimension, w_m is a nonnegative weight expressing the degree of attention to the mth dimension (using the constraint $\sum_k w_m = 1$), and r defines the metric. If $r = 2$, we obtain the Euclidean distance metric. However, in the PRAS paper and in this paper, a city-block metric is assumed and $r = 1$. In the upper panel of Figure 1, a typical generalization gradient is shown for the first dimension of the stored exemplar in our example production rule.

Insert Figure 1 about here

But suppose now that our production rule contains a non-zero range on one of the two dimensions:

IF $x \in [4, 8]$ (first dimension)
AND $x \in [5, 5]$ (second dimension)
THEN classify x in category A

If the coordinates of a target stimulus, say $(6, 5)$, fall inside the range specified in the condition part of the production rule, the distance between the target stimulus and the stored information is defined to be zero ($d_{ij} = 0$). Therefore, all target stimuli that fall inside this range will trigger the same respons. This type of generalization, which is based on abstracted information, is called *secondary generalization.* In contrast, generalization that is based on a single exemplar is called *primary generalization.* If the coordinates of a target stimulus, say $(3, 5)$, fall outside the range specified in the condition part of the production rule, we need to compute the distance to the nearest boundary of that range. In this case, $d_{ij} > 0$, and similarity is again computed by the univeral law of generalization. Since we only compute the distance from the nearest boundary (and not from the middlepoint) of the range, it is said that in these cases, both primary and secondary generalization operate simultaneously. In the lower panel of Figure 1, both primary (the gradient curves at the left and the right) and secondary generalization (the plateau in the middle) is shown for the first dimension of the abstracted information in our example production rule.

But do we really need a secondary generalization mechanism for learning categories? The PRAS paper describes an empirical study that was specifically designed to answer this question. We will briefly describe the setup and main results of the study. For more details, we refer the reader to the PRAS paper.

Consider the stimulus pattern layout in Figure 2. The figure shows the positions of 9 stimuli in a two-dimensional space. The numbers in squares indicate category P exemplars. The numbers in circles indicate category Q exemplars. The category acquisition phase consisted of five blocks. In each block, all four training exemplars were shown in a random order. During this phase, feedback showing the correct category label was given after each trial. The training phase was followed by a transfer phase where the complete set of nine exemplars was presented five times in a random order. During the transfer phase, no feedback was given.

Insert Figure 2 about here

Three variants of this task were used. In the first variant, the layout as shown in Figure 2 was used. This is called the R0 condition. In a second variant, the stimulus layout was rotated thirty degrees counterclockwise (the R30 condition). In a third variant, the stimulus layout was rotated sixty degrees counterclockwise (the R60 condition). The critical test patterns were pattern 5 and pattern 6. To see why these test patterns are critical, consider pattern 5 in Figure 2. Note that patterns 3 and 4 are far apart. If no abstractions are formed, the critical pattern 5 would be more similar to the top exemplars (pattern 1 and 2) than to the bottom exemplars. Therefore, an exemplar-based account would predict that pattern 5 is classified more often as a P pattern, while an abstraction-based model would predict that pattern 5 is classified more often as a Q pattern. The only way out for an exemplar model is to stretch dimension 2 (and hence shrink dimension 1) to increase the intrasimilarity of the two category Q exemplars. However, in the rotated conditions (R30 and R60), stretching and shrinking the dimensions would not help. Therefore, if in these rotated sets, pattern 5 is indeed classified more often in the Q category, this may provide evidence in favour of an abstraction-based representation.

Sixty-eight first-year psychology students participated in the PRAS study. Twenty-four subjects were assigned to the R0 task, twenty-four subjects were assigned to the R30 task, and twenty subjects were assigned to the R60 task. The results are shown in the first column of Table 2 in the PRAS paper, and again for convenience in the first column of Tables 1, 2 and 3 in the current paper. These observed proportions are based on the total number of category P responses during the transfer phase, pooled over blocks and subjects. Interestingly, the critical pattern 5 was only assigned to category P in about 10.0%, 38.3% and 9% of the cases for the R0, R30 and R60 sets respectively. In the PRAS paper, three variants of the PRAS model and the standard GCM model were fit to the data. The first variant of the PRAS model is a primary generalization (PG) only mode, where no abstractions are made. In the second variant of the PRAS model, secondary generalization (SG) was obligatory whenever possible. In the third variant, secondary generalization was probabilistic (FG), and the probability of making an abstraction depended on a free parameter (the π parameter, see page 448 in the PRAS paper). The fits of the different models are discused at length and summarized in Table 2 in the PRAS paper. Briefly, the conclusions were that in the R0 set, there was little need for a secondary generalization process. The fits of the GCM and PG models were very similar to the fits of the SG and FG models, including the predictions for the critical pattern 5. However, in the R30 condition, the fits of the primary generalization only models (GCM and PG) were clearly inferior when compared to the secondary generalization models (SG and especially FG). Indeed, for the critical pattern 5, only the predictions of the FG model were more or less in line with the observed proportions. Similarly, in the R60 condition, the SG and FG models yielded the best predictions. Overall, in the R30 and R60 conditions, the predictions of the models that allow for secondary generalization were better than the exemplar based models, suggesting that –at least in these types of categorization tasks– there is a need for an abstraction mechanism in order to qualitatively explain subjects' categorization behaviour.

Alternative exemplar models

The goal of this paper is to re-analyze the PRAS data using a larger variety of exemplar models. In what follows, we will describe three variants of the standard GCM that have not been considered in the original PRAS paper. Two of these variants have been described elsewhere (Ashby & Maddox, 1993; De Schryver, Vandist, & Rosseel, 2009), a third variant has been constructed solely for the purpose of this paper. But before we describe the three variants of the GCM, we first describe its standard formulation.

The Standard GCM model

In what follows, we will assume a two-category task with categories P and Q and two-dimensional stimuli. According to the standard formulation of the GCM (Nosofsky, 1984, 1986), the probability that stimulus i is classified in category P is given by

$$P(P|i) = \frac{b_P \sum\limits_{j \in P} \eta_{ij}}{b_P \sum\limits_{j \in P} \eta_{ij} + (1 - b_P) \sum\limits_{j \in Q} \eta_{ij}}, \tag{1}$$

where b_P $(0 \leq b_P \leq 1)$ is the category P response bias and η_{ij} denotes the similarity between target stimulus i and stored exemplar j. The similarity measure is assumed to be related to the psychological distance d_{ij} by,

$$\eta_{ij} = \exp(-c * (d_{ij})^q), \tag{2}$$

where $q = 1$ yields an exponential function and $q = 2$ yields a Gaussian function. The parameter c $(0 \leq c < \infty)$ is interpreted as a sensitivity parameter reflecting overall discriminability in the psychological space. In a two-dimensional space, the psychological distance between stimuli i and j is given by

$$d_{ij} = [w_1 \, |x_{i1} - x_{j1}|^r + \ (1 - w_1) \, |x_{i2} - x_{j2}|^r]^{1/r}, \tag{3}$$

where x_{i1} and x_{i2} are the psychological values of exemplar i on the two dimensions. The parameter w_1 $(0 \leq w_1 \leq 1)$ is the attention weight for dimension 1. The exponent r

defines the distance metric ($r = 1$: city-block-metric; $r = 2$: Euclidean metric). In this paper, we have assumed $q = r = 1$. The standard GCM has three free parameters: the response bias b_P, the attention weight w_1, and the sensitivity parameter c.

Alternative model 1: GCM-γ

An important extension of the GCM was proposed by Ashby and Maddox (1993). By including a response-scaling parameter (γ), the GCM-γ can account for more deterministic response patterns. Technically, the only change in this model is the response rule which is now defined as:

$$P(P|i) = \frac{b_P \sum\limits_{j \in P} (\eta_{ij})^\gamma}{b_P \sum\limits_{j \in P} (\eta_{ij})^\gamma + (1 - b_P) \sum\limits_{j \in Q} (\eta_{ij})^\gamma}. \tag{4}$$

The GCM-γ model has four free parameters: b_P, w_1, c and γ. If $\gamma = 1$, we again obtain the standard formulation of the GCM. In this case, the decision rule assumes that observers respond probabilistically by “matching” to the relative summed similarities of each category. On the other hand, when $\gamma > 1$, observers respond more deterministically with the category that yields the larger summed similarity.

The response-scaling parameter (γ) seems to play a crucial role if exemplar-based models are contrasted with abstraction-based models. For example, in the work of Smith and colleagues (Minda & Smith, 2001; Smith & Minda, 2002), a number of studies were reported that supported the prototype model. In their work, the standard formulation of the GCM was used as a representative model for the exemplar account. However, their results were heavily critized by Nosofsky and Zaki (2002). They showed that if the GCM-γ variant was used instead of the standard GCM, all the experimental results reported by Smith et. al. could be accounted for.

In the PRAS paper, the standard GCM model was unable to qualitatively fit the observed data in the R30 and R60 conditions. This finding was used to support the idea of secondary generalization. But perhaps, by extending the GCM with a response-scaling parameter, an exemplar-based model relying on primary generalization only might be able to fit the data after all.

Alternative model 2: GCM-REX

The core assumption of exemplar models is that a category is represented by a set of exemplars. In practice, most applications of exemplar theory have implicitly or explicitly assumed that *all* exemplars that were encountered during a training phase are stored and become part of the category representation, even if this number is excessively high. For example, in the McKinley and Nosofsky (1995) study, 4000 exemplars were presented during a category learning task, and all exemplars were assumed to be part of the category representations.

However, it may be the case that instead of a full set of exemplars, only a subset of these exemplars is used to represent a category. This hypothesis has led to the development of a family of exemplar models collectively called *Reduced Exemplar* (REX) models (Rosseel, 2002; De Schryver et al., 2009). Here, we will use the most basic variant, called ‘Rex Leopold I’. Rex Leopold I is designed to be identical to the GCM, with the only exception that the full set of exemplars can be replaced by a reduced set of exemplars. The remaining exemplars form a true subset of the full set. Several other variants of the Rex family have been developed (e.g. Rex Leopold II, Rex Albert I), but since we only use one version, we will refer to it as GCM-REX in the remainder of this paper.

Originally, the REX models were designed to reduce the set of exemplars in large datasets. In addition, REX models are usually fit to individual data, since it is assumed that the reduced exemplar sets may differ among individual subjects. Nevertheless, we had some limited success in datasets where the responses are aggregated over individuals, and the number of exemplars is relatively small.

In this dataset, there are only two training exemplars per category. That means there are only 9 possible subsets (including the full set). For each subset, we have fitted the standard GCM (where the full set was replaced by a reduced set), and for each condition, we only retain the subset with the best fit (in terms of loglikelihood). Implicitly, the GCM-REX model adds an extra (binary) free parameter for each exemplar: it is either included in the representation or not. Therefore, the GCM-REX model has seven free parameters: b_P, w_1, c, and 4 binary parameters for the training exemplars. Note that we have not included the γ parameter when fitting the GCM-REX models to the data. The reason is that we do not want to confound the various extensions of the standard GCM. In the first variant, we only change the response rule. In this variant, we only allow for training exemplars to be removed. In the third variant, we allow for exemplars to move around in the psychological space.

Alternative model 3: GCM-MOVE

Just like it is 'standard practice' to include all training exemplars in the representation of a category when the GCM is fitted to empirical data, it is 'standard practice' to assume that the coordinates of these training exemplars in the psychological space remain fixed and do not change over time. Here, we suggest that exemplars may very well move around in the psychological space, after they have been stored. When we store an exemplar for the first time, the features of that exemplar (the exact colour, the exact height, . . . ; in short, all the features of that specific exemplar) may be stored in memory with great precision. But when time passes, and many more exemplars of the same category have been encountered, our recollection of that original exemplar may have become somewhat blurred, or imprecise. There is a large literature in the memory literature describing the loss of accuracy and distortion of memory traces (see Koriat, Goldsmith, and Pansky (2000) for a review). In the categorization literature, it is largely acknowledged that the memory for specific exemplars is not perfect. But only a few computational models of categorization have taken this explicitly into account. In their work on General Recognition Theory (GRT), Ashby and his colleagues have advocated a probabilistic approach: instead of representing an exemplar as a fixed point in the psychological space, GRT-based models assume that they are represented by a (typical Gaussian) probability distribution in the psychological space (Ashby & Townsend, 1986; Ashby & Perrin, 1988; Ashby, 1992; Rosseel, 2002).

Here, we propose an alternative way to reflect the ambiguity of the exact location of stored exemplars in their psychological space. Instead of using a probabilistic representation, we again use a fixed-point representation, as this is the default approach for exemplar models. However, we will assume that the location of an exemplar has been shifted towards the center of the category. The amount by which the exemplar is shifted is governed by a free parameter $0 \leq \delta < 1$. If $\delta = 0$, the coordinates of the exemplars remain fixed at their original locations, as in the standard formulation of the GCM. Note that we do not consider $\delta = 1$, since in this case, all exemplars would move to the center of the category, and we get a prototype model. However, for values of δ between zero and one, the exemplars are somewhere located between their original location and the category center.

Importantly, by moving the exemplars towards the category center, the intrasimilarity of the exemplars belonging to the same category is increased. In the R0 condition of the PRAS study, we observed that the intrasimilarity of the category P and category Q exemplars could be increased by shrinking the first dimension (see Figure 2). However, for the R30 and R60 condition, stretching and shrinking did not help. We speculate that instead of stretching and shrinking the dimensions, we can increase the intrasimilarity of the exemplars by shifting their coordinates towards the category center. The GCM-MOVE variant has four free parameters: the three parameter of the standard

GCM, and the δ parameter.

Fitting the GCM and its variants to the PRAS data

To fit these models to the observed proportions reported in Tables 1, 2 and 3, a computer search was used to find the values of the free parameters (separately for each condition) that maximed the loglikelihood function,

$$L = \sum_i \ln N_i! - \sum_i \sum_k \ln f_{ik}! + \sum_i \sum_k f_{ik} \ln p_{ij}$$

where N_i is the total frequency with which stimulus i was presented and f_{ik} and p_{ik} are, respectively, the observed frequency and predicted probability with which stimulus i is classified in category k (Nosofsky, Clark, & Shin, 1989). Based on the information in the PRAS paper, we have used $N_i = 120$, $N_i = 120$ and $N_i = 100$ for the R0, R30 and R60 conditions respectively. The observed frequencies were computed by multiplying the observed proportions as reported in the PRAS paper by these total frequencies. For each model, we report two measures of global fit. The loglikelihood, multiplied by minus two for convenience ($-2L$), and the AIC score, which is defined by

$$\text{AIC} = -2L + 2p$$

where p is the number of free parameters in the model. The AIC score penalizes models with more free parameters. Lower values for $-2L$ and AIC are better. In addition, following the PRAS paper, we also report the root mean squared deviations (RMSDs) for the training patterns (1–4), the critical patterns (5 and 6) and the remaining transfer patterns (7–9).

Results

The predicted proportions category P responses for each model are reported in Table 1 for the R0 condition, Table 2 for the R30 condition, and Table 3 for the R60 condition. The RMSD values are reported in Table 4. Finally, the estimated parameter values for the GCM model and its three variants are reported in Table 5.

Insert Table 1 about here

Insert Table 2 about here

Insert Table 3 about here

Insert Table 4 about here

Insert Table 5 about here

GCM

When the standard GCM is fitted to the PRAS data, the estimated values for the free parameter are highly similar to the values reported in Table 5 of the PRAS paper (cf. the rows labeled ‘GCM-9’). Similarly, the predicted proportions category P responses are very close to the values reported in Table 2 of the PRAS paper (cf. the column labeled ‘GCM-9’). In the R0 condition, the global fit of the standard GCM is (much) better than the global fit of the PRAS models. However, for the R30 and R60 condition, the PRAS models fit the data equally well, and sometimes slightly better (i.e. PRAS-SG in the R60 condition). Nevertheless, and this is the important point made by Vandierendonck (1995), the GCM is not able to qualitatively fit the critical pattern 5 in the R30 and R60 conditions. This is best illustrated in Table 4 which contains the root mean squared deviations for the critical patterns. Especially the PRAS-FG variant does a much better job predicting the subjects' responses of the critical pattern 5 than the standard GCM model.

GCM-γ

Overall, the fits of the GCM-γ variant are much better than the fits of the standard GCM. Moreover, the GCM-γ model did a much better job in predicting the responses of the critical pattern 5 in the R30 and R60 conditions. For the R30 condition, the predictions are even closer to the observed proportions than the best of the two PRAS models. For the R60 condition, the predictions are much closer to the observed data than the standard GCM model, but the PRAS predictions are qualitatively still better for the critical patterns.

At first sight, this seems to confirm the findings of Nosofsky and Zaki (2002) that adding a response-scaling parameter γ to the standard GCM can make a huge difference. Indeed, the good overall fits and the excellent fit of the GCM-γ model for the critical pattern 5 in the R30 condition may cast doubt over the claim in the PRAS paper that secondary generalization is an essential mechanism if we wish to explain the PRAS data.

However, there are two caveats. First, the PRAS models did still fit the critical pattern 5 better in the R60 condition. Second, the estimated parameter values of the GCM-γ model are rather extreme, in all three conditions. As can be seen in Table 5, either the value of the c parameter or the value of the γ parameter is extremely high. Moreover, the higher the value of c, the lower the value of γ and vice versa. In fact, it turns out that for this particular dataset, the c and γ parameters are not identified. Only their ratio c/γ is identified. There is a whole range of values for c and γ that yield the same fit if the ratio c/γ is kept constant. It is not clear at this point if this problem is due to the small number of training exemplars, the specific layout of the stimulus patterns in the psychological space, or the specific values for the observed proportions, or any combination of these.

Thus, although both the global and qualitative fits of the GCM-γ model are very good, the fact that two parameters of the model were not identified in this datset, we hesitate to interpret these results in favour of the GCM-γ model.

GCM-REX

In the R0 condition, the best fitting model is a reduced model using a subset of three training exemplars (pattern 1 was removed). The overall fit of this reduced GCM is very good, and certainly better than the standard GCM. However, the prediction for the critical pattern 5 is not very good. In fact, all other models did better in this condition. Thus although removing an exemplar in this condition may have improved the global fit, it did not help for predicting the subjects' proportions of the critical pattern 5.

For the two other conditions (R30 and R60), the best fitting model is the one were no exemplars are removed. In other words, the best fitting model is simply the standard GCM model where all exemplars are retained in the category representations.

GCM-MOVE

In the R0 condition, the estimated value for δ is zero. This suggests that there is no need to move the exemplar coordinates to the center in order to improve the model fit. As a result, the parameter values and predictions of GCM-MOVE in this condition are identical to the standard GCM.

In the R30 condition, however, the estimated value for δ is 0.12. Not only is the global fit better than the standard GCM, the fit of the critical pattern 5 is better too, although it is still not as good as the PRAS-FG fit of that pattern.

In the R60 condition, $\delta = 0.28$, and the global fit is again better than the standard GCM. Also, the fit of the critical pattern 5 has much improved in comparison to the standard GCM, although again, the fit of the PRAS-FG model is still better.

Discussion of the fitting results

Of the three variants of the GCM that we have proposed in an attempt to fit the PRAS data, only the GCM-MOVE model is able to provide both a good global fit, and a decent fit of the critical pattern 5 in the R30 and R60 conditions. Although the global fits of the GCM-γ model are even better, the c and γ parameter were not identified in this model. The GCM-REX variant is only able to provide a good global fit of the data in the R0 condition, but can not improve on the standard GCM in other conditions. As for the critical pattern 5, the best predictions are still made by the PRAS variants.

What have we learned from our model fitting experiment? First, and this is not a new finding, the GCM-γ model can behave strangely in certain circumstances. In our case, the problem was that two parameters (c and γ) were not identified. In other studies, the parameters of the GCM-γ are identified, but the estimated values are difficult to interpret in terms of an exemplar model. Indeed, the GCM-γ model has been critized by several authors because it is believed that the γ parameter may do more than merely making the predictions of the model more deterministic. For example, Smith and Minda (2002) have claimed that adding the γ parameter to the GCM "enlarged prototype-enhancement effects" and that "adding gamma can be tantamount to adding a prototype". Their claims were partially confirmed by Navarro (2007). In this case, the GCM-γ model should not be used as an example of an exemplar model that only relies on primary generalization.

Second, the GCM-REX model did not do any better than the standard GCM. Of course, the small number of training exemplars is a disadvantage for the GCM-REX model. However, we believe that even with a larger number of training exemplars, the GCM-REX model would still not be able to qualitatively fit the critical pattern 5 in this dataset. The deeper reason is that the GCM-REX has difficulty fitting aggregated data. Ideally, we should be able to fit the GCM-REX model to the individual datasets. Only in this scenario would we be able to judge if the GCM-REX is capable or not to qualitatively fit the critical pattern 5.

Third, the GCM-MOVE model did a excellent job in fitting the data. It may not have done as well as the PRAS models for the critical pattern 5, but it fitted the data much better than the standard GCM. In fact, we are pleasantly surprised by the good performance of the GCM-MOVE variant, and we will investigate its capabilities in future studies.

Conclusion

In this paper, we have tried to extend the standard GCM model in an attempt to fit the data in the PRAS paper. Three variants were considered. In the first variant (GCM-γ), we followed Nosofsky and Zaki (2002) and extended the GCM with a response-scaling parameter. In a second variant, we followed De Schryver et al. (2009) and allowed for the exemplar representation to exclude some exemplars. In the third and last variant, we extended the GCM by allowing the exemplars to move towards the category center. Although several variants were able to obtain a better global fit in some or all of the conditions, no variant was able to convincingly do better for the critical pattern 5, at least not for all three conditions.

We conclude that the PRAS dataset is still a challenging dataset for modelers of category learning and generalization. We therefore hope that the fitting experiment that we have reported in this paper will stimulate other people in the categorization field to take a (second?) look at both the PRAS model and the PRAS dataset reported in Vandierendonck (1995).

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*, 249-277.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Bourne, L. E., Jr. (1982). Typicality effects in logically defined categories. *Memory & Cognition*, *10*, 3–9.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 638–648.

De Schryver, M., Vandist, K., & Rosseel, Y. (2009). How many exemplars are used? explorations with the rex leopold i model. *Psychonomic Bulletin & Review*, *16*, 337–343.

Erickson, M. A., & Kruschke, J. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

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

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*, 15–36.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning & Verbal Behaviour*, *16*, 321–338.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation.* (Vol. 18, pp. 49–94). New York: Academic Press.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accurarcy. *Annual Review of Psychology*, *51*, 481-537.

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

McKinley, S. C., & 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.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 333–352.

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

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.

Mostofsky, D. I. (Ed.). (1965). *Stimulus generalization.* Standford: Stanford University Press.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, *51*, 85 – 98.

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., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 282–304.

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

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.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

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

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science.

Science, *237*, 1317-1323.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research.* New York: Wiley.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, *2*, 442–459.

Author Note

Correspondence concerning this paper should be addressed to Yves Rosseel, Department of Data Analysis, Ghent University, Henri Dunantlaan 1, B-9000 Ghent (Belgium). Email: Yves.Rosseel@UGent.be

Table 1

Observed and predicted proportions of category P responses in the R0 condition

Pattern	Obs	GCM	PRAS-SG	PRAS-FG	GCM-γ	GCM-REX	GCM-MOVE
				Set R0			
1	.900	.959	.967	.906	.954	.937	.959
2	.967	.964	.960	.907	.965	.970	.964
3	.033	.088	.041	.106	.076	.060	.088
4	.075	.038	.034	.110	.039	.033	.038
5	.100	.116	.026	.117	.126	.136	.116
6	.967	.927	.953	.891	.930	.929	.927
7	.767	.751	.722	.652	.726	.735	.751
8	.750	.657	.574	.545	.676	.698	.657
9	.683	.741	.797	.735	.749	.743	.741
-2L		64.73	94.92	94.60	60.66	55.01	64.73
npar		3	4	5	4	3 + 4	4
AIC		70.73	102.92	104.60	68.66	69.01	72.73

Table 2

Observed and predicted proportions of category P responses in the R30 condition

Pattern	Obs	GCM	PRAS-SG	PRAS-FG	GCM-γ	GCM-REX	GCM-MOVE
				Set R30			
1	.817	.805	.861	.842	.895	.805	.847
2	.800	.803	.735	.756	.789	.803	.814
3	.242	.167	.154	.127	.242	.167	.217
4	.017	.084	.092	.088	.051	.084	.085
5	.383	.528	.124	.454	.440	.528	.499
6	.467	.560	.433	.518	.478	.560	.546
7	.717	.518	.563	.548	.625	.518	.500
8	.750	.687	.370	.633	.706	.687	.676
9	.250	.290	.249	.294	.215	.290	.257
-2L		94.45	201.43	93.88	61.20	94.45	90.24
npar		3	4	5	4	3 + 4	4
AIC		100.45	209.43	103.88	69.20	108.45	98.24

Table 3

Observed and predicted proportions of category P responses in the R60 condition

Pattern	Obs	GCM	PRAS-SG	PRAS-FG	GCM-γ	GCM-REX	GCM-MOVE
				Set R60			
1	.940	.934	.920	.915	.930	.934	.930
2	.810	.806	.816	.796	.810	.806	.810
3	.310	.326	.346	.342	.367	.326	.373
4	.080	.028	.045	.040	.040	.028	.042
5	.090	.205	.085	.103	.141	.205	.143
6	.310	.323	.362	.357	.342	.323	.321
7	.670	.557	.609	.958	.584	.557	.588
8	.390	.402	.420	.434	.440	.402	.445
9	.220	.239	.110	.111	.167	.239	.167
-2L		62.85	57.10	144.98	54.06	62.85	53.76
npar		3	4	5	4	3 + 4	4
AIC		68.85	65.10	154.98	62.06	76.85	61.77

Table 4

Root mean squared deviations between model predictions and observed data for the three task conditions

Pattern	GCM	PRAS-SG	PRAS-FG	GCM-γ	GCM-REX	GCM-MOVE
			Set R0			
1-4	.038	.040	.051	.033	.027	.038
5	.016	.057	.017	.026	.036	.016
6	.040	.014	.076	.037	.038	.040
7-9	.055	.124	.139	.060	.048	.055
			Set R30			
1-4	.039	.070	.072	.031	.039	.034
5	.145	.259	.071	.057	.145	.116
6	.093	.034	.051	.011	.093	.079
7-9	.101	.237	.121	.057	.101	.099
			Set R60			
1-4	.019	.027	.029	.027	.019	.028
5	.115	.005	.013	.051	.115	.053
6	.013	.052	.047	.032	.013	.011
7-9	.048	.075	.080	.063	.048	.063

Table 5

Estimated parameter values of the standard GCM and its three variants

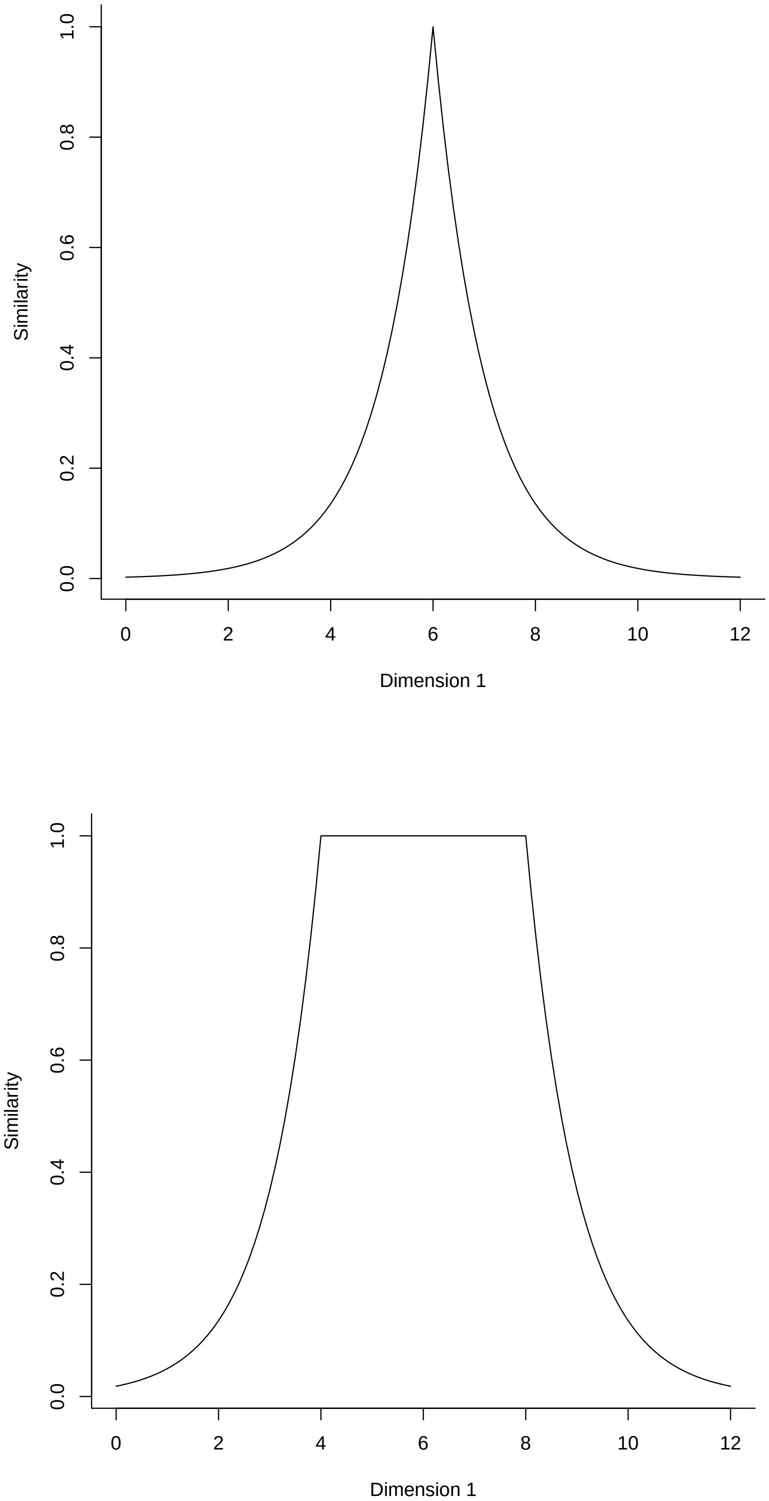
Par	GCM	GCM-γ	GCM-REX	GCM-MOVE
		Set R0		
c	1.94	14.03	2.06	1.93
b_P	0.53	0.56	0.66	0.53
w_1	0.27	0.23	0.31	0.26
γ	–	0.13	–	–
δ	–	–	–	0.00
		Set R30		
c	1.61	0.29	1.61	1.81
b_P	0.43	0.33	0.43	0.46
w_1	0.62	0.62	0.62	0.63
γ	–	8.75	–	–
δ	–	–	–	0.12
		Set R60		
c	1.53	0.07	1.53	1.40
b_P	0.54	0.44	0.54	0.51
w_1	1.00	1.00	1.00	1.00
γ	–	20.04	–	–
δ	–	–	–	0.28

Figure Captions

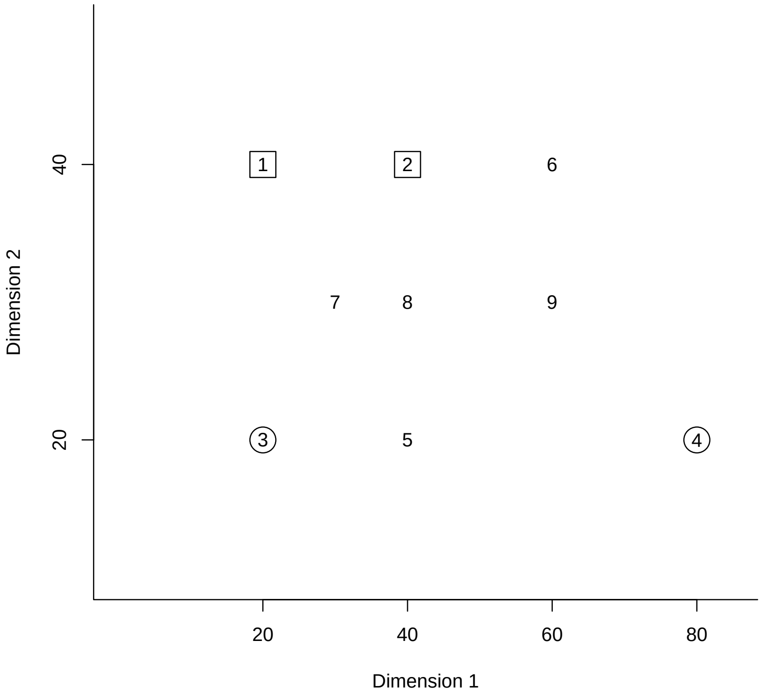
Figure 1. Similarity gradient for different values of a target stimulus along the first dimension. The upper panel illustrates a primary generalization gradient. The lower panel illustrates both secondary generalization (between the values 4 and 8) and primary generalization. In both panels, the value for the steepness parameter of the exponential decay function is $c = 1$.

Figure 2. Stimulus pattern layout in Condition R0 of the categorization task described in the PRAS paper. The numbers in squares indicate category P exemplars. The numbers in circles indicate category Q exemplars. The other patterns are only presented during the transfer phase.

Secondary generalization in categorization, Figure 1



Secondary generalization in categorization, Figure 2


Dimension 2
Dimension 1