On the Interaction Between Exemplar-Based Concepts and a Response Scaling Process

Daniel J. Navarro
School of Psychology
University of Adelaide

Abstract

An analysis of the "response scaling" parameter in the Generalized Context Model is presented. In light of the existing debate over the behavior of the model when this parameter is included, three different interpretations are discussed, in order to illustrate the effect of the parameter at the decision level, the category similarity level, and the representational structure level.

1 Introduction

Classification tasks present people with stimuli and their accompanying category labels, and require label prediction for novel stimuli. Early theories postulated that people use classification rules to solve the labeling problem (e.g., Bruner, Goodnow & Austin, 1956). Based on favorable empirical work (e.g., Rosch & Mervis, 1975) and philosophical dissatisfaction with rule-based accounts of human concepts (e.g., Wittgenstein, 1953), later works assume that category members share a kind of "family resemblance" (e.g., Rosch, 1978). Resemblance theories adopt formal measures of similarity, with theories varying in the manner that the category is represented (see Komatsu, 1992 and Murphy, 2002 for overviews). For example, prototype theories (e.g., Reed, 1972; Smith & Minda, 1998) represent a category using a single prototypical or canonical 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 (e.g., Medin & Schaffer, 1978; Nosofsky 1984, 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.

This paper analyzes the effect of the "response scaling" parameter in the Generalized Context Model (GCM: Nosofsky, 1986), a classic exemplar model. It reviews and extends existing interpretations used by researchers in the field. The intention is not to decide between different views, though a cursory overview of some empirical data is provided. To

provide a suitable background to the discussion, I begin with an overview of the structure and psychological assumptions of the GCM.

2 The Generalized Context Model

The exemplar principle for concepts was introduced by Medin and Schaffer (1978). The central idea is that people represent a concept solely in terms of the collection of previously-seen instances, and do not form any kind of abstract summary representations. The GCM is one of the canonical exemplar models, and is built by making assumptions at several distinct levels.

Representational Structure. As an exemplar theory, the GCM assumes that a concept is equivalent to the set of stimuli that it encompasses. Accordingly, the model retains an internal representation of every previously observed object. It is frequently assumed that stimuli are represented as points in an m–dimensional space (e.g., Shepard, 1957), with similar objects located close to one another. In other words, similarity is some function of distance in the psychological space. Distance d_{ij} in a psychological space is measured using an attention-weighted Minkowski metric (Minkowski, 1891), often called an r-metric or ℓ_p-norm. So, if $x_i = (x_{i1} \dots x_{im})$ is a vector that describes the location of the ith stimulus in this space, we measure psychological distance as follows:

$$d_{ij} = \left[\sum_{k=1}^{m} \left(w_k \left| x_{ik} - x_{jk} \right| \right)^r \right]^{\frac{1}{r}}, \tag{1}$$

where w_k denotes the proportion of attention applied to the kth dimension. Although not particularly important for this paper, these attention weights are crucial for modeling human learning using exemplar models (e.g., Kruschke, 1992). The set of stored object locations constitutes the representational structure for the GCM, typically extracted from empirical similarities using multidimensional scaling (MDS: e.g., Cox & Cox, 1994). The metric r is considered to be a property of this representation, since it describes a fundamental characteristic of the stimuli that is either fixed on theoretical grounds, or extracted with the help of MDS.

Stimulus Similarities. Since the GCM is a resemblance theory as well as an exemplar theory, it assumes that judgments are based on some notion of similarity, where similarity is generally assumed to be an exponentially decaying function of psychological distance (Shepard, 1987). Thus, the stimulus similarity function s_{ij} is given by,

$$s_{ij} = \exp\left(-\lambda d_{ij}\right), \tag{2}$$

where λ denotes the steepness of the exponential decay, called the *generalization gradient* or *typicality gradient*. Shepard bases this exponential law both on a range of empirical

data, and on a rational analysis of the structure of psychological spaces. Shepard argues that evolutionary processes have shaped psychological spaces in such a way that objects that share the same real-world consequences tend to be located near one another. Correspondingly, a set of stimuli with the same consequences form a "consequential region" that matters to the organism. He shows that with little knowledge about the size of the consequential region that contains j, the probability that i falls inside the same region is given by an exponential function of distance. It is worth noting that Shepard's ideas can be extended somewhat (e.g., Tenenbaum & Griffiths 2001; Navarro 2006), producing a range of different generalization gradients that will be discussed later.

Category Similarities. Since an exemplar model treats the category as the sum of its parts, the natural way to produce category similarities is to sum the individual similarities. Therefore, the similarity between novel item i and category c is given,

$$s_{ic} = \sum_{j \in c} s_{ij}. \tag{3}$$

So the result of the assumptions made at the representational and similarity levels is a simple "sum-similarity rule". The intuition underlying this summation rule is that each exemplar is treated as a unique, genuinely independent source of evidence for the corresponding category. As a result, complex rules for combining exemplar similarities are avoided, in favor of simple rules that do not allow any complicated interactions between items.

Decision-Making. Given a set of category similarities, the GCM needs to make a prediction about the likelihood with which someone would assign stimulus i to category c. Formally, if we let θ_{ic} describe this probability, then the decision rule used by the GCM is given by the classic Luce choice rule (Luce, 1963; Shepard, 1957):

$$\theta_{ic} = \frac{s_{ic}}{\sum_d s_{id}}. \tag{4}$$

This probabilistic decision rule provides a method of comparing model predictions to behavioral data, and is closely related to the choice rules employed in logistic regression (e.g. Hosmer & Lemeshow, 2000), and by the random utility models (e.g., McFadden, 1974; Manski, 1977) that are popular in classical economics. It assumes that people assign stimulus i to category c with probability equal to the judged likelihood that it really belongs to category c. If people adopt this "probability matching" strategy, then Equation 4 is a reasonable choice. However, people do not always probability match, so later versions of the model (e.g., Ashby & Maddox, 1993; Nosofsky & Zaki, 2002) altered the decision model by raising the category similarities to some power γ. So if θ^*_{ic} denotes the response probabilities under the revised "GCM-γ" model, we write

$$\theta^*_{ic} = \frac{{s_{ic}}^\gamma}{\sum_d {s_{id}}^\gamma} \tag{5}$$

Throughout the paper, asterisks indicate a quantity that has been affected by γ. Notice that γ heightens the advantage of categories with larger s_{ic}. As $\gamma \to \infty$, the advantage becomes so extreme that the category with largest s_{ic} is always chosen. Since γ can be viewed as a way of translating subjective probabilities into observed decisions, it is traditionally called a *response scaling* parameter.

3 Response Scaling

Adding parameters to formal models is a complex enterprise, as the two special issues of the *Journal of Mathematical Psychology* devoted to model selection can attest. Not surprisingly then, there has been some concern expressed in the literature about the shift from the basic GCM to the more elaborate GCM-γ. The worry is that γ may affect the GCM in a more fundamental manner than is implied by the response scaling effect. Smith and Minda (2002, p. 809) state this concern most concisely, arguing that γ "acts systematically to steepen typicality gradients and to enlarge prototype-enhancement effects". This is more than just a concern about the statistical complexity of the new model (though this is certainly an issue worthy of research). Ultimately, the question is about the relationship between the formal model and underlying psychological theory.

While γ was originally assumed to operate at the *decision-making level*, the concerns with "steepened typicality" and "enlarged prototype-enhancement effects" can be interpreted as suggesting effects on the *category similarity function*. It may even be the case that "adding gamma can be tantamount to adding a prototype" (Smith & Minda 2002, p. 809), suggesting that the parameter could be viewed as introducing changes at the *representational level* as well. With this in mind, the remainder of this paper will examine the effect of the γ parameter under three different interpretations:

- The parameter shapes the decision process only.
- The parameter affects the way that category similarities are constructed.
- The parameter alters the set of represented entities and their generalization gradients.

As indicated, the intention is not to decide between these three views, especially since they are all extreme cases. Nor is it the intention to determine whether the γ parameter allows the GCM to mimic a prototype model: although this is the concern expressed by Smith and Minda, I focus on the somewhat broader question of what the γ parameter actually does. Thus, the aim is to discuss the different interpretations and the consequences that attach to each, in the hope that this may usefully inform subsequent research using the model.

4 Sequential Decision Processes

Traditionally, γ is interpreted as controlling the decision process, and is often justified using limiting arguments, since

$$\lim_{\gamma \to \infty} \theta^*_{ic} = \begin{cases} 1 & \text{if } \arg\max_d \left(\sum_{j \in d} s_{ij} \right) = c \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

Unfortunately, while convenient, this limit does not provide a natural interpretion for $\gamma = 3$, for instance. Fortunately, a number of authors have provided more powerful justifications. For instance, when Ashby and Maddox (1993) introduced the parameter, they argued that the choice rules in Equations 4 and 5 should be interpreted as reflecting deterministic decisions based on limited data, rather than probabilistic decisions based on perfect knowledge, similar to the approach taken by random utility models. In other words, the original usage of γ in fact relied on the notion that people always make deterministic decisions based on the evidence available to them. The probabilistic aspect to the decision process lies only in the noise associated in the evidence accumulation stage. Nosofsky and Palmeri (1997) expanded on this idea, by pointing out that the GCM-γ decision rule is equivalent to the deterministic approach used by a simple *exemplar-based random walk* (EBRW) model. This equivalence is extremely useful, so I briefly reproduce and extend Nosofsky and Palmeri's remarks on this topic.

Consider an idealized "exemplar learner" making decisions about the category membership of some novel stimulus i. On the basis of the learner's previous exposure to category members, the probability that the stimulus belongs to category c is given by the GCM category membership probability θ_{ic}. Accordingly, in a two-category decision task, the learner who wishes to maximize the probability of making the correct response should choose category c if and only if $\theta_{ic} > \frac{1}{2}$, always choosing the most likely category. Notice, however, for the learner to adopt this strategy, the *probability* θ_{ic} must be known to her. This seems extremely unlikely, and does not in fact follow from the exemplar principle: under the exemplar principle we are committed to the assumption that the learner stores *instances*, not probabilities. As a consequence, the learner must rely on stored instances to make a decision about an unknown probability, preferably as quickly as possible. In other words, the stored exemplars must be retrieved, and then used to draw inferences about the value of the membership probability θ_{ic} and thus make the decision as to which category the stimulus belongs in.

How might such a process unfold? If we disregard the (implausible) possibility that the observer can access all her knowledge simultaneously, this becomes a *sequential analysis* problem (e.g., Wald, 1947). Sequential analysis is a branch of statistics concerned with testing statistical hypotheses using data that arrive over time. In psychological terms, we assume that the learner draws a series of observations (either from the environment or

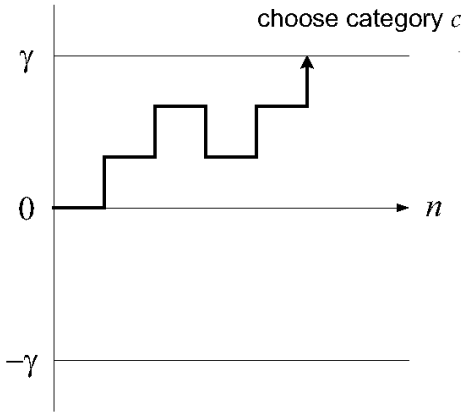


Fig. 1. An example of the EBRW-SPRT decision procedure. Stimuli sampled from memory provide constant amount of evidence in favor of the category to which they belong. Thus, in the EBWR, the random walk takes a step up with probability θ_{ic}, and a step down with probability $1 - \theta_{ic}$. Sampling terminates once one response has γ samples more than the other.

from memory) that provide information about the value of θ_{ic}. The learner terminates this search as soon as "sufficient" information has been gathered to inform a decision as to whether θ_{ic} is greater than or less than $\frac{1}{2}$. Although the tools of sequential analysis are purely statistical methods, they have been very successful in describing psychological decision processes, and are now standard techniques in the field (e.g., Ratcliff 1978; Vickers, 1979; Luce, 1986).

The EBRW model applies these ideas by retaining the exemplar representation used in the GCM, and makes decisions about the unknown value of θ_{ic} by sampling repeatedly (and with replacement) from the set of stored exemplars. Each sample provides a constant amount of evidence favoring the category to which it belongs, but items similar to the cue are more likely to be recalled. In other words, in response to novel stimulus i, the probability that a sampled item is exemplar j is proportional to the similarity s_{ij} between the two. As a result, the probability that a sample provides evidence in favor of category c is θ_{ic}, and the amount of evidence provided is constant. For a two-category decision, this can be conceptualized as the random walk process illustrated in Figure 1. At the start of the process, the walk starts in state 0, with no evidence in favor of either category. On any given sample the walk takes a step up (adds 1 to the tally) with probability θ_{ic}, and a step down (subtracts 1 from the tally) with probability $1 - \theta_{ic}$. The process terminates once the walk hits $+\gamma$ or $-\gamma$, with a walk terminating at the positive boundary producing a category c response[1]. The analytic properties of this process were studied by Feller (1968, ch. 14) under the name of the *gambler's ruin problem*. He demonstrates that if p is the probability of taking a step up and $q = 1 - p$ is the probability of taking a step down,

[1] Among other things, the full EBRW model allows asymmetric decision thresholds.

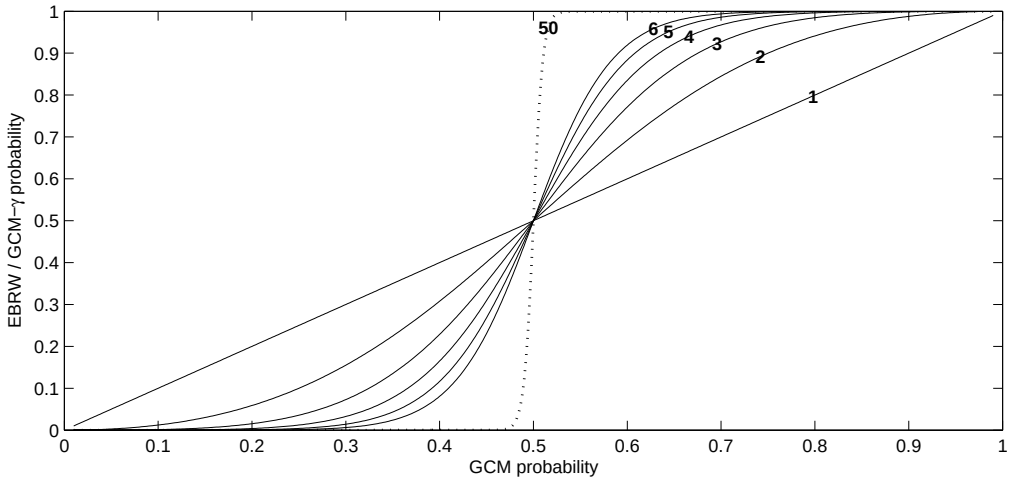


Fig. 2. Decision probabilities as a function of the underlying category membership probabilities, under an EBRW-SPRT interpretation of the GCM-γ model.

then the probability g_{ic}^* that the search process terminates at the upper boundary is

$$g_{ic}^* = \begin{cases} \frac{(q/p)^\gamma - 1}{(q/p)^{2\gamma} - 1} & \text{if } p \neq \frac{1}{2} \\ \frac{1}{2} & \text{otherwise} \end{cases}. \tag{7}$$

By applying Equation 4 for a two-category problem (with categories c and d), we observe that $q/p = s_{id}/s_{ic}$. It is then trivial to show that,

$$g_{ic}^* = \frac{{s_{id}}^\gamma - {s_{ic}}^\gamma}{{s_{ic}}^\gamma} \frac{{s_{ic}}^{2\gamma}}{{s_{id}}^{2\gamma} - {s_{ic}}^{2\gamma}} = \frac{{s_{ic}}^\gamma}{{s_{ic}}^\gamma + {s_{id}}^\gamma} = \theta_{ic}^*. \tag{8}$$

Thus, as noted by Nosofsky and Palmeri (1997), GCM-γ can be viewed as a special case of the EBRW model. With this interpretation in mind, it is illustrative to plot the response probabilities for the GCM-γ model as a function of the underlying category membership probabilities, and the value of γ. Figure 2 shows these curves for a range of γ values. Obviously, at $\gamma = 1$ the GCM-γ response probabilities exactly mirror the underlying GCM probabilities, and as γ becomes very large, the response curve will come to resemble a step function (as illustrated by the dotted line showing $\gamma = 50$).

It is not difficult to show that the EBRW model (and hence GCM-γ) is a special case of Wald's classic *sequential probability ratio test* (SPRT), a correspondence which turns out to be very useful for interpreting γ. The idea of using the SPRT as a psychological model originates with Stone (1960), though there are some difficulties associated with this (see Luce, 1986). The SPRT, like the EBRW model, assumes that evidence is accrued until some fixed threshold is reached, an accumulation process that is provably optimal,

in the sense that it makes the fastest possible decisions for a fixed error tolerance (Wald & Wolfowitz, 1948). The SPRT works as follows. Suppose we have a set of t independent observations $(y_1, y_2, \ldots, y_t)$, each of which corresponds (in the EBRW case) to an exemplar retrieved in response to the observation of a novel item. Additionally, we have two rival statistical hypotheses h_1 and h_2, which in this case correspond to the possible category decisions. Given this, the evidence provided by data (i.e., the retrieved exemplars) for h_1 over h_2 is given by the odds ratio,

$$\prod_{j=1}^{t} \frac{p(y_j \,|\, h_1)}{p(y_j \,|\, h_2)}. \tag{9}$$

Bayesians typically refer to this quantity as the Bayes factor. It is generally convenient, however, to work with log-odds, since these combine additively. If we let z_j represent the log-odds provided by the jth sample, we obtain the overall log-odds

$$Z_t = \sum_{j=1}^{t} \left[\ln \frac{p(y_j \,|\, h_1)}{p(y_j \,|\, h_2)} \right] = \sum_{j=1}^{t} z_j \tag{10}$$

So when data arise sequentially, the optimal way to accrue evidence is to add the log-odds from the current observation z_j to a running tally. As indicated above, in the EBRW each datum corresponds to a retrieved exemplar, which will obviously turn out to be a member of category c with probability θ_{ic}. Now, from the learner's standpoint, the actual value of θ_{ic} is relevant only insofar as it dictates the appropriate course of action (e.g., choose c). Accordingly, an optimal learner will propose statistical hypotheses that correspond to each possible action. In a two-choice task where the learner must pick between categories c and d, these hypotheses may be written as follows:

$$\begin{aligned} h_c &: \theta_{ic} \geq \tfrac{1}{2} \\ h_d &: \theta_{ic} < \tfrac{1}{2}. \end{aligned} \tag{11}$$

With these as sensible hypotheses, the learner needs to be able to evaluate the agreement between hypothesis and data. For instance, the learner needs to be able to assess the likelihood $p(y_j \in c \,|\, h_c)$ that an element of category c will be drawn from memory in the event that it really is the more likely category. To do so, the learner requires some prior belief (i.e., before sampling starts) about the possible values that θ_{ic} might take on (otherwise, the problems is poorly defined). Note that θ_{ic} is an unknown probability of success, so we are talking about setting a prior over a Bernoulli probability. Now, while an observer could adopt all kinds of priors, a uniform prior $p(\theta_{ic} = \theta) \propto 1$ is a principled choice here: Jaynes (2003, p. 382-386) makes a compelling case that a uniform prior is ideal for a learner who believes that either outcome (in this case categories c and d) is possible, but is otherwise ignorant. The two hypotheses are then formed by segmenting this prior according to which outcome is more likely. Under this assumption, it is straightforward

to calculate the chance that a sample belongs to category c according to hypothesis h_c: [2]

$$p(y_j \in c \,|\, h_c) = \int_0^1 p(y_j \in c \,|\, f_{ic} = \theta) p(f_{ic} = \theta \,|\, h_c)\; d\theta = \int_{\frac{1}{2}}^1 2\theta \; d\theta = 3/4. \tag{12}$$

Similarly, the probability of observing an element of category d under hypothesis h_c is simply $\frac{1}{4}$. Since hypothesis h_d corresponds to the other half of the uniform prior, it assigns probability $\frac{1}{4}$ to the possibility of drawing members of category c and probability $\frac{3}{4}$ to category d exemplars. The consequence of the uniform prior is that every datum provides a 3:1 odds in favor of the corresponding response,

$$\frac{p(y_j \,|\, h_c)}{p(y_j \,|\, h_d)} = \begin{cases} 3 & \text{if } y_j \in c \\ 1/3 & \text{otherwise} \end{cases} \tag{13}$$

Although the 3:1 rule is a consequence of the uniform prior specifically, there are a range of priors that produce a "constant evidence rule". For the current paper I will keep the 3:1 rule for convenience, but it has no substantive effect on the EBRW-SPRT model. To see this, note that when the log-odds are accumulated in the random walk process, we obtain:

$$z_j = \begin{cases} \ln 3 & \text{if } y_j \in c \\ -\ln 3 & \text{otherwise} \end{cases}. \tag{14}$$

However, the decision to use the natural (base e) logarithm is purely a convention. Accordingly, we can shift to a base 3, which gives the EBRW update rule:

$$z_j = \begin{cases} 1 & \text{if } y_j \in c \\ -1 & \text{otherwise} \end{cases}. \tag{15}$$

Essentially, any symmetric prior over possible category membership probabilities yields this update rule. All that will change is the base to which the logarithm is taken (which I keep fixed at 3 for the remainder of the paper).

This relationship is useful for two reasons. Firstly, it implies that the EBRW implements an optimal decision process, namely the SPRT. As discussed by Nosofsky (1998), the original GCM can be treated as a "rational" Bayesian model by converting the similarity function to a likelihood function. The SPRT equivalence implies that there is a similarly rational interpretation for the decision process built into GCM-γ. Secondly, it allows us to convert γ into an error tolerance parameter, in exactly the same sense that an α-level is used as a parameter in statistical tests. A convenient result by Wald (1947; see Schervish,

[2] The "2" in Equation 12 arises because the probability density for θ *given* the hypothesis h_c is a proper uniform density that integrates to 1, over the truncated range $[\frac{1}{2}, 1]$, and hence $p(\theta_{ic} = \theta \,|\, h_c) = 2$.

1995, p. 550 for the derivation) gives us the relationship,

$$\gamma = \log_3 \frac{1-\alpha}{\alpha}. \tag{16}$$

In general, this relationship is only approximate, since it assumes that the evidence tally Z_t exactly equals γ or $-\gamma$ when the process terminates. As it happens, the structure of the EBRW ensures that this is true for all integer-valued γ, and we obtain the error tolerance

$$\alpha = \frac{1}{1+3^{\gamma}}. \tag{17}$$

So the EBRW-SPRT approach to response scaling suggests that we should treat γ as a kind of error tolerance, and suggests that it may be more natural to convert γ to α when reporting parameter values [3].

5 Minkowski Measures of Category Similarity

The EBRW-SPRT interpretation of GCM-γ provides an elegant way of understanding how γ should be interpreted, if one were to treat it as an aspect of the decision process. However, the concerns raised by Smith and Minda (1998, 2002) operate at the category similarity and representational levels. It is natural, therefore, to ask whether sensible (or at least interesting) interpretations of the model can be generated by treating γ as a lower-level parameter. The remainder of this paper addresses this question.

Suppose γ were treated as a parameter that operates at the category similarity level. Steepened typicality gradients and enlarged prototype-enhancement are both similarity-based effects, so this is a natural way to examine Smith and Minda's concerns. In one sense, steepened typicality gradients are not a serious problem: if γ acts to steepen the typicality gradient, we can adjust λ to widen it again if we need to. So, to the extent that this effect occurs, it does not allow GCM-γ to "do" anything that is not already built into the GCM. However, it complicates the interpretation of parameter values, since γ may distort the estimates of λ. To examine the effect of γ at the similarity level, it may be helpful to rewrite GCM-γ in a manner that minimizes this problem. Fortunately this is not difficult. To do this, it is helpful to reparameterize the model using $\beta = \lambda\gamma$. If β is used as the generalization gradient instead of λ (the "β-reparametrization"), the stimulus similarity measure becomes,

$$s'_{ij} = \exp\left(-\beta d_{ij}\right) = \exp\left(-\lambda\gamma d_{ij}\right) = \left[\exp\left(-\lambda d_{ij}\right)\right]^{\gamma} = {s_{ij}}^{\gamma}. \tag{18}$$

[3] To be clear, it is worth pointing out that the α-level in question is not the tolerance for making incorrect classifications, it is the tolerance for accidentally choosing the category that is not most likely according to the original GCM.

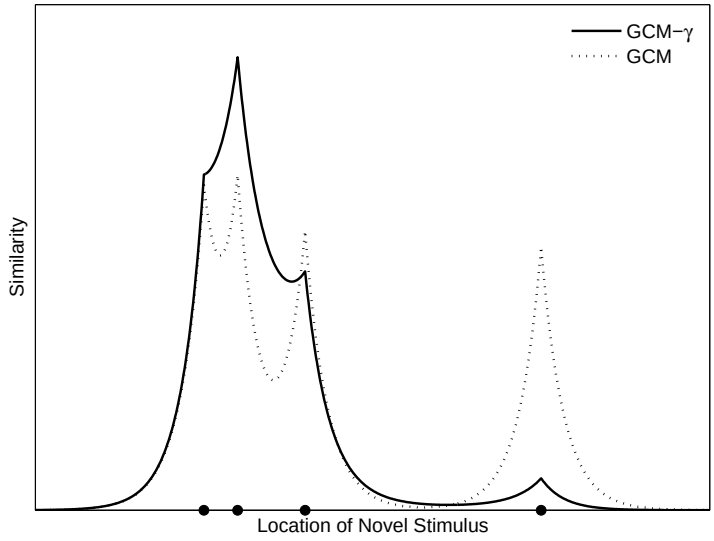


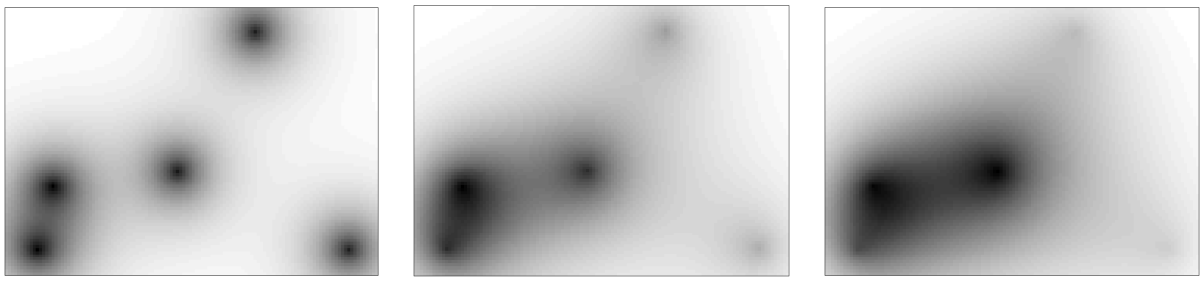
Fig. 3. Category similarity functions for GCM and GCM-γ with $\gamma = 4$. The exemplars are the same as for Figure 8, and (since this section is not concerned with the response scaling effect) the functions are normalized to have equal area.

More to the point, $s_{ij} = {s'_{ij}}^{1/\gamma}$. While this reparameterization helps neutralize the steepening effect, it does not sacrifice the interpretability of the stimulus similarities, since s'_{ij} is a perfectly legitimate stimulus similarity function. It describes an exponentially decaying function of the psychological distance between two represented entities, so the transformation from s_{ij} to s'_{ij} is entirely trivial and does not in any way influence the structure or interpretation of the stimulus similarity function. However, by using β to define the typicality gradient, GCM-γ predictions can be written as,

$$\theta^*_{ic} \propto \left(\sum_{j \in c} {s'_{ij}}^{1/\gamma} \right)^{\gamma} . \tag{19}$$

For ease of exposition, the normalizing term is dropped in this expression. By using β as the typicality gradient, the $1/\gamma$ term inside the sum 'balances' the γ term on the outside (roughly speaking). Varying γ should have less of an effect on β than it does on λ.

Curiously, this reparametrization affords us an understanding of how γ alters prototype-enhancement effects. Broadly speaking, prototype-enhancement effects in the GCM arise because stimuli that sit in the middle of an "exemplar cloud" will tend to be near a lot more exemplars than those stimuli on the fringes. So, these "central" or "prototypical" exemplars will tend to get more support from other exemplars. Accordingly, novel stimuli that lie near where one would expect to find a prototype will be judged to be more typical



(a) (b) (c)

Fig. 4. Similarity functions for a category with 5 exemplars (black circles), and γ values of 1, 3 and 5. Darker tones indicate greater similarity.

of the category, even in an exemplar model. The prototype enhancement effect for the original GCM is illustrated by the dotted line in Figure 3, which plots s_{ic} (the original category similarity function) as a function of x_i (the location of the novel stimulus i) for a simple unidimensional category with $\gamma = 1$. The two exemplars on the left side of the category cluster together very tightly, and so reinforce each other, while the two exemplars further to the right have little influence on one another.

Now, when we interpret γ as a category similarity parameter, we are required to restore the original GCM "probability matching" rule. We do this by assuming that $\theta^*_{ic} \propto s^*_{ic}$, where s^*_{ic} denotes a *revised category similarity* that results from allowing γ to shape the category similarity rather than the decision process. Under the β parameterization, this revised similarity is written:

$$s^*_{ic} = \left(\sum_{j \in c} {s'_{ij}}^{1/\gamma} \right)^{\gamma} . \tag{20}$$

By writing the model in this way, the effect of γ has been "pushed down" into the category similarity function s^*_{ic}. As suggested by Smith and Minda, this has a substantial influence on the category similarity function. The solid line in Figure 3 illustrates this, showing what happens when we increase γ to 4, and then plot s^*_{ic} against x_i. Clearly, the revised category similarity function has a rather different shape to the original. The density plots in Figure 4 provide another way of visualizing this effect, illustrating the way that the similarity function changes shape for a five-stimulus category represented in a Euclidean two-dimensional space.

As with Equation 1, the new category similarity function in Equation 20 takes the form of a Minkowski measure. [4] This observation allows the development of a theoretical basis for

[4] The Minkowski measures are closely related to the power means (also called Hölder means), in which $\bar{x} = \left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^p\right)^{1/p} = n^{-1/p}\|(x_1, \ldots, x_n)\|_{\ell_p}$ However, in the same way that the original GCM uses a sum measure rather than an arithmetic mean, the GCM-γ model uses Minkowski norms rather than power means. When two categories contain the same number of exemplars,

γ as a similarity-level parameter. The original GCM assumes that individual similarities combine additively, according to a city-block metric ($\gamma = 1$). By varying γ, we move to a different Minkowski measure (the $1/\gamma$ measure, to be precise). Letting $p = 1/\gamma$ denote the particular measure used by the model, we observe that for a category with n exemplars whose individual similarities to novel stimulus i are described by the vector $(s'_{i1}, \ldots, s'_{in})$, Equation 20 can be interpreted as a Minkowski category similarity, adopting the ℓ_p measure:

$$s^*_{ic} = \|(s'_{i1}, \ldots, s'_{in})\|_{\ell_p} \tag{21}$$

As noted by Cross (1965), when p is small, the Minkowski measure weights small values more heavily. In the limit, as the metric $p \to 0$, the Minkowski measure converges to the infimum measure. From a categorization standpoint, as $\gamma \to \infty$, we obtain a *minimum exemplar similarity rule*, in which the category similarity function becomes $s^*_{ic} = \min_{j \in c} s'_{ij}$. Note, however, that this limit is taken in the β-reparametrization and assumes a constant β. In the original parametrization, the limit needs to be taken slightly more carefully, to ensure that some stimulus generalization takes place. Thus the corresponding limit is $\gamma \to \infty$, $\gamma\lambda \to \beta$. In any case, there is a similar limit as $\gamma \to 0$ (and $\gamma\lambda \to \beta$), with a *maximum exemplar similarity rule* resulting. In this situation, the category similarity function becomes $s^*_{ic} = \max_{j \in c} s'_{ij}$. Interestingly, this rule has been shown to be effective in accounting for category-based induction tasks (e.g., Osherson et al. 1990).

While these limits provide a useful illustration of the effect of parameter variation, empirical parameter estimates do not correspond to the limiting cases, so it is important to examine the behavior of the model across a plausible parameter range. In particular, we would like to know how closely the revised category similarity mimics the different limiting cases across a reasonable range. In this investigation, I use the range $0 < \gamma \leq 10$ since it is broad enough to cover most empirical fits, and also suffices to demonstrate the important effects. Even so, this range is probably a little broader than the empirically observed range. The metrics of interest are the min-similarity rule, the max-similarity rule, and the original GCM sum-similarity rule. Additionally, since Smith and Minda (1998, 2002) are concerned with the possibility that the GCM-γ model behaves like a prototype model, I include a similarity-to-prototype rule, where the prototype is assumed to lie at the centroid of the exemplar set.

In the first simulation I generated 1000 categories consisting of 10 uniformly distributed stimuli varying only on a single dimension (range 0 to 10) and measured the similarity (with $\beta = 1$) of a novel item (also uniformly distributed) to the category across a range of values of γ. The max-similarity, min-similarity, sum-similarity and prototype-similarity values were calculated, and correlated with the GCM-γ category similarity. The results are shown in Figure 5a. Naturally, when γ is near 0, the Minkowski similarity correlates almost perfectly with a max-similarity rule (white squares), and when $\gamma = 1$ it correlates perfectly with a sum-similarity rule (white triangles). As γ increases, the correlations between GCM-γ and these two rules declines, and the correlation with the min-similarity

the two are equivalent.

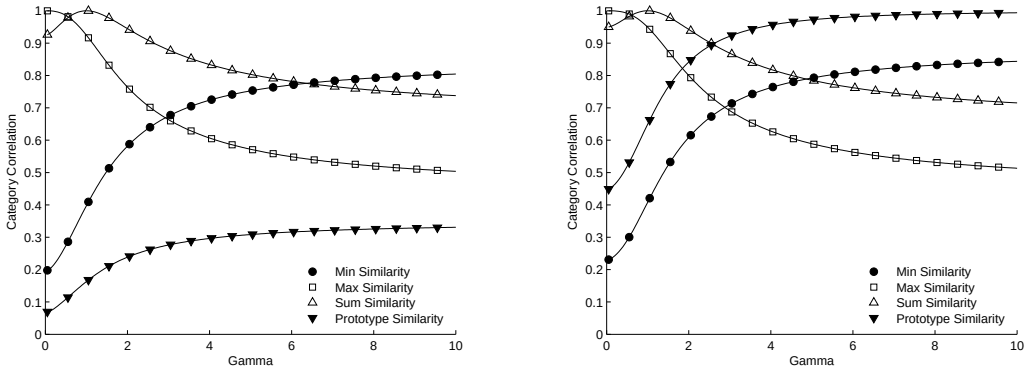


Fig. 5. The interpolation (panel a) and extrapolation (panel b) behavior of GCM-γ, with respect to a category consisting of 10 uniformly distributed stimuli varying only on a single dimension. Since there are 10 category exemplars, we might expect the choice of metric (via setting γ) to have a substantial effect on the category. When $\gamma = 1$ the GCM-γ predictions correlate perfectly with a sum-similarity rule, as one would expect. Similarly, at very small γ values the model mimics a max-similarity rule, but at large γ it starts to resemble a min-similarity rule. This holds for both interpolation and extrapolation. The correlation between the GCM-γ similarities and prototype similarities is always grows with γ, but is much stronger for extrapolations than interpolations.

rule (black circles) increases, so that when γ reaches 10, the min-similarity rule correlates most strongly. Interestingly, the prototype rule correlates poorly with the GCM-γ rule across most of the range. This is initially surprising, until one notices that the similarity assessment here is almost always an *interpolation* problem, since the novel stimulus will generally lie inside the range spanned by the exemplar set. If we redo the simulations with the novel stimulus fixed at 0 (on the edge of possible category members), the correlations for the min-, sum-, and max- measures do not change substantially, but the correlations with the prototype model increase dramatically. In other words, raising γ does appear to make the category similarity function look more like a prototype model, but for empirically-plausible values the effect is only large with respect to *extrapolation outside the range* spanned by the known exemplars. Nevertheless, this result should be interpreted with some caution, since it is based only on simple unidimensional categories.

An intuitive understanding of the partial prototype-mimicry effect is provided by Figure 6, which shows category similarity functions for a simple three-stimulus category located in two dimensional space (with a city-block metric in the top panels, and a Euclidean metric in the lower panels). As γ rises from 1 to 5, the similarity function shows the smoothing effect over the interior of the category, but the space between the exemplars is still judged to be less similar to the category than the original exemplars would be. It is only at the largest – empirically implausible – value of 500 that the interior of the similarity

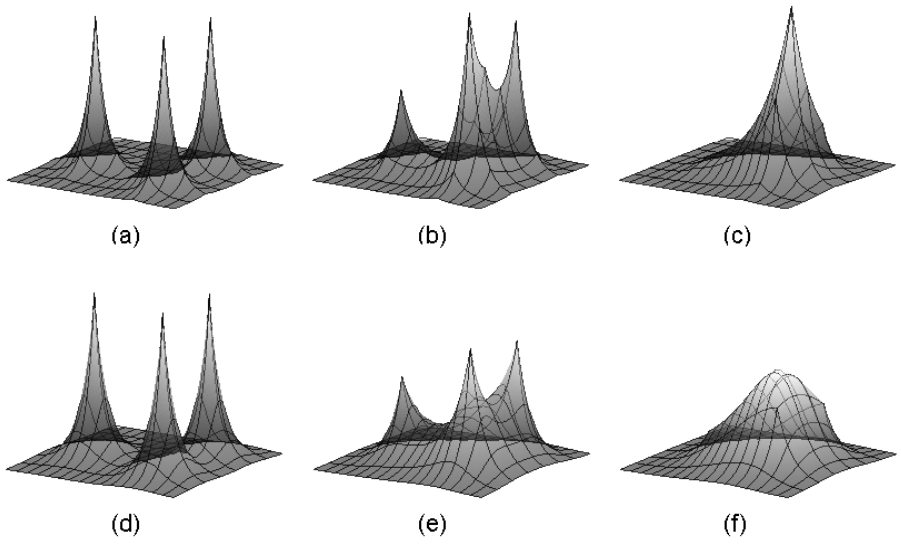


Fig. 6. Category similarity functions for a category with three exemplars embedded in a two-dimensional space, and assuming a moderate typicality gradient. In panels a–c the metric is city-block ($r = 1$), while in panels d–f the metric is Euclidean. In the leftmost panels, there is no smoothing ($\gamma = 1$). In the middle panels $\gamma = 5$, while in the rightmost panels $\gamma = 500$.

function resembles the generalization function from a prototype. However, at $\gamma = 5$, the region around the edges of the exemplar set has smoothed enough that it seems intuitively plausible that it could correlate strongly with the predictions of a prototype model, even though the interior looks quite different. For instance, notice that the peak associated with the "exception item" (the one on the left) has declined relative to the other two, which is what one would expect from a prototype.

6 Changes to the Underlying Representation

The development in the last section assumes that γ has no effect on the representations or the decision process, and operates only on the way that exemplar similarities combine with each other. Alternatively, the model could be rewritten to accommodate a representational interpretation, in which people still probability match, and individual similarities are assumed to combine additively, as in the original GCM. Under this view, varying γ produces a change to the set of internally-represented entities that the learner uses to describe a category. In essence, in order to see whether raising γ is "tantamount to introducing a prototype", we should push the effect of the parameter further down, into the representation itself.

Since I will now be talking about a range of different representations, with entities that may not correspond to specific exemplars or to generalized prototypes, I will adopt a neutral term, and refer to a represented entity as a *predicate*. A prototype model postulates a single predicate for each category, while an exemplar model incorporates a predicate for each stimulus. In order to build a representational interpretation of GCM-γ, we need to define the set of represented predicates as a function of γ. This is intended to be an illustrative discussion only, so I constrain γ to integer values for the sake of mathematical tractability, and return to the *original* parametrization of GCM-γ, in which

$$s^*_{ic} = \left(\sum_{j \in c} s_{ij} \right)^{\gamma} \tag{22}$$

This allows us to motivate a particular predicate set easily, by noting that this equation describes a polynomial function. Expanding this function using the standard expansion of a polynomial gives us the expression,

$$s^*_{ic} = \sum_{z} \left[\frac{\gamma!}{\prod_{j=1}^{n} z_j!} \prod_{j=1}^{n} {s_{ij}}^{z_j} \right]. \tag{23}$$

In this expanded expression, each z_j is an integer between 0 and γ, and the sum is taken over all possible vectors $z = (z_1, \ldots, z_n)$ such that $z_1 + z_2 + \ldots + z_n = \gamma$. The advantage to this expansion is that we now have a category similarity function expressed as a sum, as per the GCM. Moreover, since this is a polynomial expansion of Equation 22, it is straightforward to provide an interpretation of z. If we imagine sampling exactly γ exemplars with replacement from the set of n stored instances, then each possible outcome corresponds to one of the possible values for z, where we treat z_j as a count of the number of times that the jth exemplar is sampled. Thus, in attempting to provide a representational interpretation of γ, it is natural to associate each term (and thus each z vector) with a specific predicate. In doing so, we assume that the similarity s^*_{iz} between novel stimulus i and a particular predicate is given by:

$$s^*_{iz} = \frac{\gamma!}{\prod_{j=1}^{n} z_j!} \prod_{j=1}^{n} {s_{ij}}^{z_j}. \tag{24}$$

Assuming that there is a sensible basis for proposing Equation 24, we have a representational interpretation of GCM-γ that preserves the sum-similarity rule and probability-matching decision process used in the GCM. That is,

$$\theta^*_{ic} = \frac{\sum_{z \in c} s^*_{iz}}{\sum_d \left(\sum_{z \in d} s^*_{iz} \right)} \tag{25}$$

As this expression makes clear, the entire effect of γ has now been pushed down into the set of represented entities (and their corresponding individual gradients s^*_{iz}).

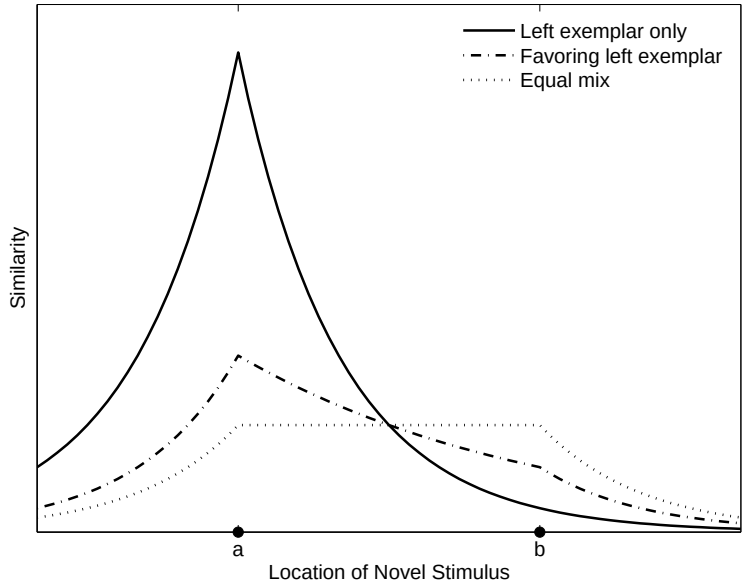


Fig. 7. Three different generalization gradients based on two exemplars. In one case (solid line), the generalization is based only on the exemplar on the left. In another case (dashed line), the "predicate" from which the learner generalizes is mainly (2/3) influenced by the exemplar on the left. In the third case (dotted line), the represented entity is an equal mix of the two exemplars.

The mathematical development in the previous paragraph serves only to discover what kinds of represented entities and similarity gradients would be needed if we required GCM-γ to employ a sum-similarity rule and a probability-matching decision process while still allowing $\gamma \neq 1$ (i.e., Equation 25). As demonstrated, the set of required predicates corresponds to the set of allowable z vectors, and the similarity gradients are described by Equation 24. Now, it seems rather unlikely that Equation 24 precisely describes a sensible rule for generalization from some internally represented predicate, so this rewrite of the model is only partly successful: it distorts the relationship between distance and generalization somewhat. Nevertheless, it is interesting to consider the basic shape of the generalization gradients that it produces. Suppose that we set $\gamma = 6$ for the very simple category shown in Figure 7, consisting of two previously observed exemplars, a and b (the black dots). The simplest possible predicate corresponds to the case when $z = (6, 0)$, since only one of the exemplars (in this case a) is involved. In our fictitious "sampling" process, all 6 sampled exemplars correspond to a, so not surprisingly the resulting generalization gradient produced by substituting this into Equation 24 looks exactly like the standard exponential similarity rule (shown by the solid line).

Continuing this example, suppose we look at the predicate corresponding to $z = (3, 3)$. In this case, half of the retrieved instances correspond to item a, and half correspond

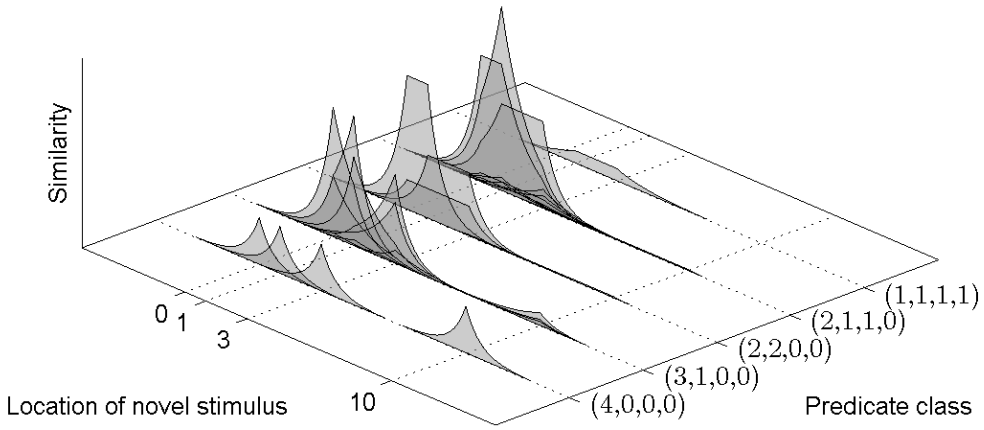
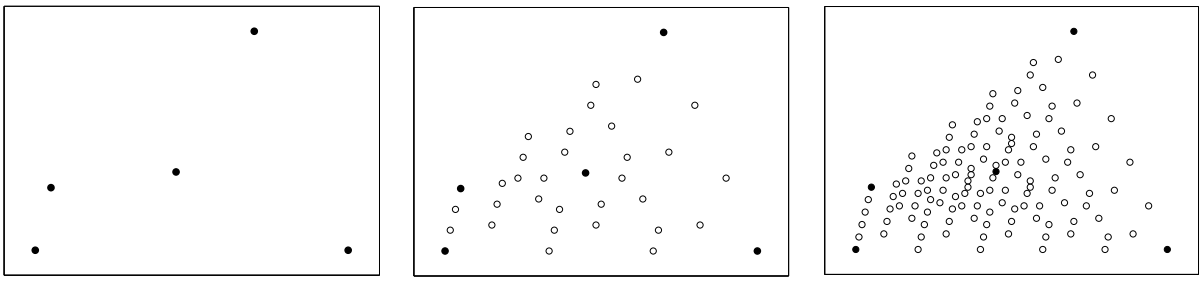


Fig. 8. An analysis of GCM-γ predictions for a one-dimensional category with four exemplars and $\gamma = 4$. The figure plots the generalization gradients for the various predicates in the model. Each row of functions corresponds to a predicate "class".

to item b. When the corresponding gradient is drawn out (dotted line) we obtain a flat-topped generalization curve, which bears a remarkable resemblance to the generalization curves predicted by Tenenbaum and Griffiths' (2001) extension of the exponential law. Moreover, when we extend this to the more complex case $z = (4, 2)$ where there is an uneven "mix" of the two original items, we obtain a skewed curve (dashed line) that bears a qualitative similarity to the skewed gradients that Tenenbaum and Griffiths (2001) refer to as the result of uneven sampling (see their p. 770). While one would not want to over-interpret these results, the basic point is that Figure 7 suggests that Equation 24 may well approximate some more principled generalization function.

This example can be extended somewhat, by drawing the generalization functions for all possible z predicates. This is illustrated in Figure 8, which plots the generalization gradients for every predicate that would be included for a simple one-dimensional category when $\gamma = n = 4$ (and with exemplars located at 0, 1, 3 and 10). The predicates are divided into five "classes", by noting that every possible z vector is a permutation of either $(4, 0, 0, 0)$, $(3, 1, 0, 0)$, $(2, 2, 0, 0)$, $(2, 1, 1, 0)$ or $(1, 1, 1, 1)$. If γ had been set to 1, only the first row of predicates would appear. However, the set of predicates increases with rising γ, since a host of "mixture predicates" are added. On the whole, the various generalization gradients involved look fairly plausible (in the sense suggested by Tenenbaum and Griffiths, 2001), which is somewhat reassuring. In fact, the aggregated category similarity functions (e.g., the solid line in Figure 3) are somewhat similar to some of the category distributions that are produced by a simple extension of Tenenbaum and Griffiths' model that can account for some of the basic category learning findings (see Navarro 2006, Figure 6).



(a) (b) (c)

Fig. 9. Expanded representations for a category with 5 exemplars (black circles). The top row shows the 'locations' of all predicates for γ values of 1, 3 and 5.

Table 1
Stimulus structure (left) for the two categories, and the corresponding similarity structure (right) that counts, for every pair of distinct items, the number of features they have in common.

A	1	0	0	0	1
	2	0	1	0	0
	3	1	0	1	1
	4	0	0	0	0
B	5	1	0	0	0
	6	1	0	1	0
	7	1	1	1	1
	8	0	1	1	1

		1	2	3	4	5	6	7	8
A	1	-	2	2	3	2	1	1	2
	2	2	-	0	3	2	1	1	2
	3	2	0	-	1	2	3	3	2
	4	3	3	1	-	3	2	0	1
B	5	2	2	2	3	-	3	1	0
	6	1	1	3	2	3	-	2	1
	7	1	1	3	0	1	2	-	3
	8	2	2	2	1	0	1	3	-

Another perspective is provided by Figure 9. If the centroid of the exemplars is taken to be the 'location' of a predicate in the representational space, the Figure depicts the locations of all predicates for the same two-dimensional category shown in Figure 4, again with γ values of 1, 3 and 5. As γ increases, the predicates 'fill in' the space within the convex hull of the exemplars, explaining why, as γ rises in Figures 3, 4 and 6, the probabilities associated with the interior of the category tends to rise.

7 A Simple Illustration

Having discussed the interaction between the response scaling process and the rest of the GCM, it is worth examining how these considerations play out in empirical data. The intent is not to make any strong claims about which interpretation is "correct", since all three views correspond to the same mathematical object (GCM-γ). Even so, some insights are possible by examining how parameter estimates change as a function of learning, for instance. To provide an illustrative example, data are taken from Smith and Minda's (1998) experiment 3 with non-linearly separable categories. This data set is convenient since the exemplar model is known to provide a good fit to these data, so we can examine the behaviour of γ with respect to a data set that the model accounts for quite well. The logical structure for the two categories is taken from Medin and Schwanenflugel (1981), and is shown in the left panel of Table 1. The panel on the right shows the relationships between stimuli, by counting the number of features shared by any two stimuli. In category A, stimulus 4 is clearly the most prototypical, and stimulus 3 is an obvious exception item. Category B is slightly less clear-cut, since stimuli 6 and 7 have the same amount of overlap with other category members, although stimulus 7 overlaps less with members of category A. Additionally, stimulus 5 is something of an exception item, since it overlaps fairly closely with category A members.

In this experiment, the stimuli were pronounceable nonwords with particular letters corresponding to stimulus dimensions. Participants were trained on the two categories using a standard supervised learning paradigm: stimuli were presented, participants were asked to predict the category memberships, and then feedback was provided. Learning curves for all 16 participants are shown in Figure 10. Each line corresponds to an individual participant, with each point corresponding to average performance over a 56-trial block. As is immediately clear, all 16 participants eventually learned to classify items correctly, but there are noticable individual differences in the task. When faced with individual differences, a common approach is to fit each participant's data separately, to avoid averaging artifacts. However, this approach carries risks of its own, since it leads to a proliferation of free parameters, and an increased risk of over-fitting the data. To address this trade-off, a natural approach is to pay attention both to the similarities and to the differences between people, and seek to find groups of participants with similar patterns of performance (e.g., Lee & Webb, 2005; Navarro, Griffiths, Steyvers & Lee, 2006). Using the Minimum Description Length clustering technique pioneered by Kontkanen et al. (2005) and extended to learning curves by Navarro and Lee (2005), it is clear that there are three distinct groups of participants whose data may be aggregated. Since this is an illustrative application, only the largest "gradual learner" group is analyzed in this paper, shown in solid lines in Figure 10.

While the analysis of learning curves does a reasonable job of extracting appropriate groups, it is important to check that the averaging within these groups does not produce

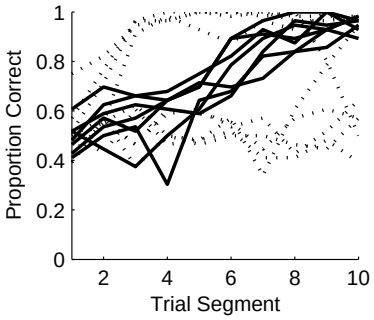


Fig. 10. Empirical learning curves for all 16 participants in experiment 3. Using the MDL approach to data clustering, the data segregate naturally into three groups. The data from the middle "gradual learning" group shown in solid lines are analyzed.

any systematic distortions to the patterns of performance across *stimuli*. This is shown in Figure 11, which plots individual learning curves for each of the 7 participants that belong to the gradual learning group, broken down by each of the 8 stimuli in the task. These curves are shown as dotted lines. Since each data point aggregates over only 7 trials for each participant, these data are very noisy. However, it appears unlikely that the averages across participants (the solid dots) systematically misrepresent the performance of any of the individuals. Given that the averages appear reasonable, the standard practice would be to fit the averaged data. However, since each of these data points represents only 49 observations, there is still some noise in the data. To minimize this noise, underlying trendlines were extracted using Kalman filtering (Kalman, 1960), shown by the solid lines. Both the averaged data and the filtered data are analyzed.

GCM-γ was separately fit to each of the 10 trial blocks, by extracting parameter values that maximize the correlation between model predictions and the data (i.e., ordinary least squares fitting). Consistent with Smith and Minda's findings, the fit of the model is best to the later stages of learning. However, the model predictions are reasonably accurate for all stages of learning, as shown in Figure 12 which plots the model predictions for all stimuli at all stages of learning, for both the averaged data (dotted lines) and the Kalman filtered data (solid lines). The filtered data are cleaner, making the model predictions easier to interpret, but the model handles the noisier data reasonably well. The key point is that GCM-γ captures the key trends in the data: stimulus 3 (and to a lesser extent, stimulus 5) is a noticable exception item and is learned slower than the more prototypical items.

For the purposes of the current paper, it is most interesting to look at the trajectory through the parameter space that is traced out across the 10 trial blocks. This path (for the specificity and "response scaling" parameters) is shown in Figure 13 for three different parametrizations of the model. The left panel plots γ against λ, the middle panel plots α against λ, and the right panel plots γ against β. The Minkowski measure interpretation on

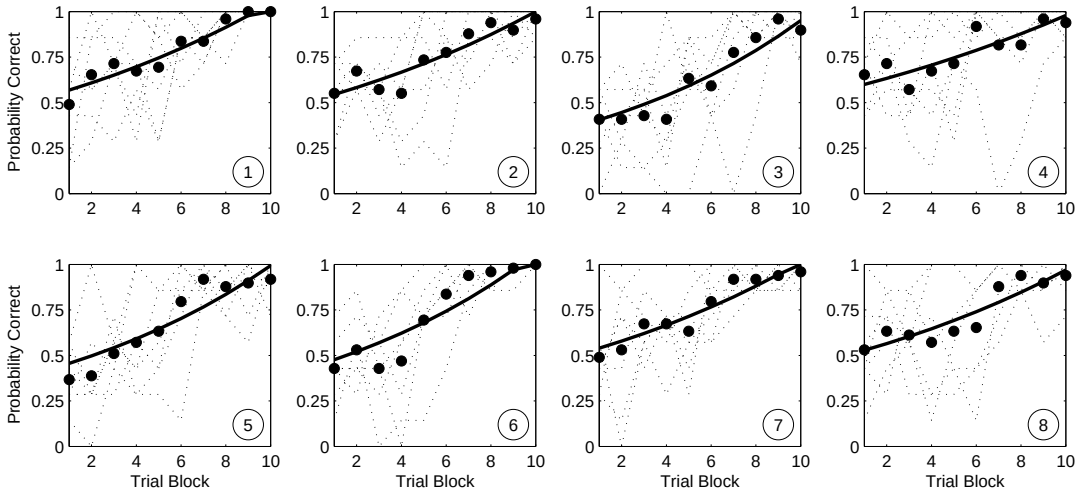


Fig. 11. Individual learning curves (dotted lines) for each participant in the group, broken down by stimulus (each subplot corresponds to a different stimulus). The underlying "trendline" (extracted by Kalman filtering) is shown in solid lines, while the average performance for each trial block is shown with black dots.

the right produces a smooth, monotonic path through the parameter space, which seems to be a desirable characteristic for a learning process. However, it is interesting to note that both γ and β tend to rise throughout learning: the typicality gradients narrow, and the similarity function tends to become somewhat more prototype-like.

The left and middle panels show curved trajectories that arise from a decision-process interpretation of the model. In this view, γ represents a rational decision process that leads the model to make predictions about response times (analytic expressions are derived by Feller, 1968). In the illustrative example presented here, these predictions seem somewhat unlikely, since γ rises steadily throughout the learning process. When γ is converted to an error tolerance α (as shown in the middle panel), a similar picture emerges, but with more emphasis placed on the changes early in learning. According to a decision process interpretation, a rise in γ (or α) implies that people become more conservative in setting their decision criteria, and accordingly should take longer to make decisions. Without response time data for this experiment, it is not clear whether this prediction is met, but it seems implausible in view of the more general tendency for people to make faster decisions as they become better practiced (e.g., Nosofsky & Alfonso-Reese, 1999).

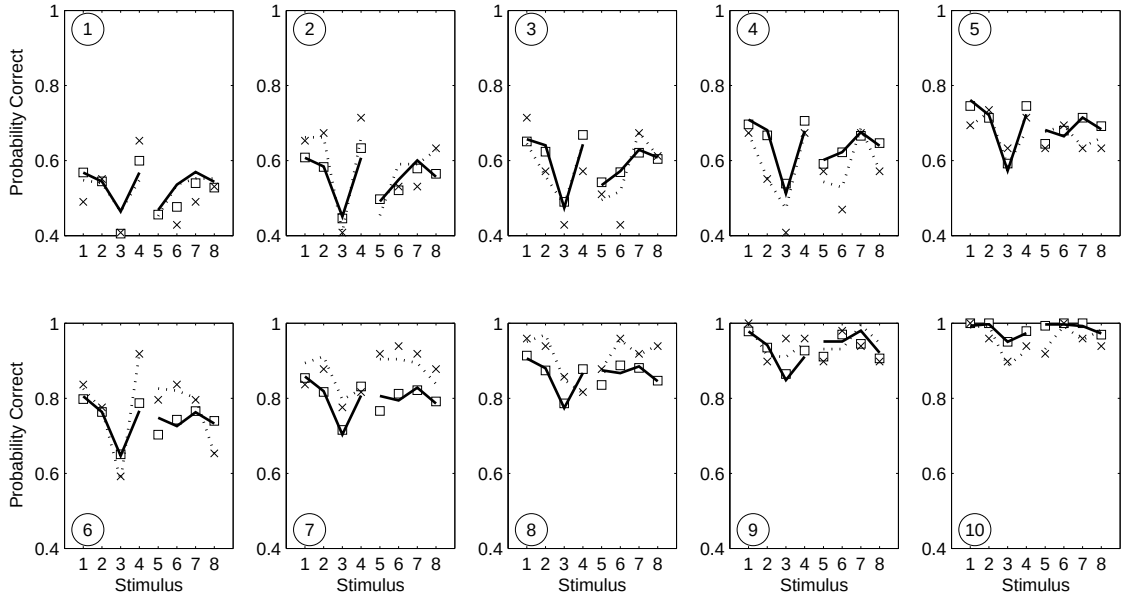


Fig. 12. Model fits to the filtered data and averaged data . Crosses denote the averaged data, and the dotted lines are the model predictions for those data. Squares denote the filtered data, and the solid lines are the corresponding model predictions. In all segments (each corresponding to a different subplot), GCM-γ captures the key trends in the data: stimulus 3 is a noticable exception item, while stimuli 4 and 7 are "prototypical" for the two categories.

8 Discussion

Given the recent emphasis on parsimony in cognitive psychology (e.g., Pitt, Myung & Zhang 2002), it is worth asking whether the γ parameter really contributes much to the GCM, especially in light of the concerns raised by Smith and Minda (1998, 2002). Empirically, there seems to be evidence supporting the additional complexity introduced by γ (e.g., Nosofsky & Zaki 2002). However, achieving better data prediction is not always preferable if it comes at the cost of rendering the model "theoretically ambiguous" as suggested by Smith and Minda (2002, p. 809). It is certainly not the case that the parameter lacks an interpretation, since its operation is comprehensible at the decision level, at the category similarity level, and even (with a certain degree of charity) at the representational level. The question of which interpretation makes most sense is difficult to answer, so some ambiguity exists. To some extent this is perhaps inevitable: as argued by the advocates of "bounded rationality" (e.g., Gigerenzer & Todd, 1999), simple decision processes based on clever representations can often approximate the outcomes of complex decision processes that use simpler representations. Even so, it is clear that disambiguation is still possible: similarity level parameters should be sensitive to different kinds of

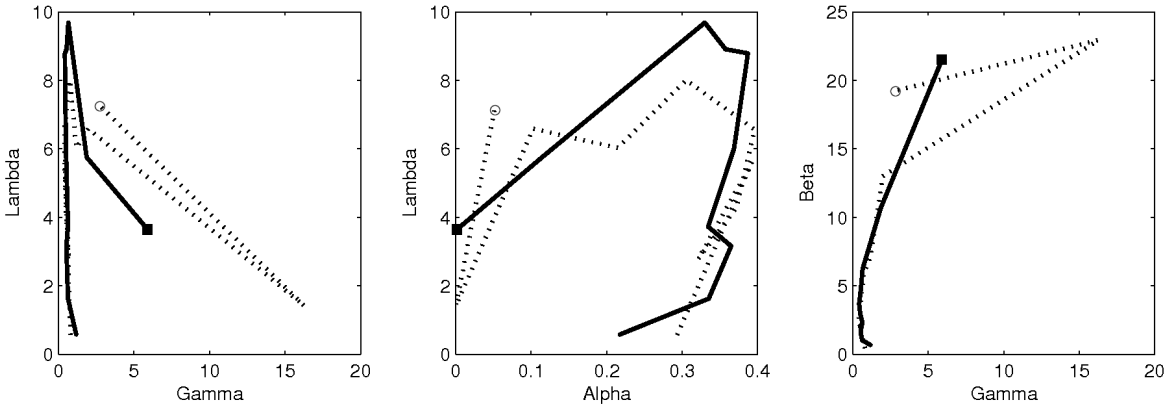


Fig. 13. Parameter values for the fitted models, for filtered (solid) and averaged (dotted) data. In the leftmost panel, the original GCM-γ parameters are reported, while the middle panel replaces γ with the error tolerance parameter α. Finally, the rightmost panel shows the "category similarity" interpretation of γ, using the β-reparametrization. Solid squares indicate the end of the path (trial block 10) for the filtered data, and the hollow circles indicate the end of the path for the averaged data.

experimental manipulations than decision level parameters. In the simple example presented, the trajectory of parameter values across learning trials has a different qualitative form depending on which interpretation is adopted. The most interesting phenomenon in this example is that γ increases as learning continues, which is slightly awkward under any interpretation of the model. This may not pose substantial problems for the model, but it does highlight the utility of more explicit models of representational changes (e.g., Love, Medin & Gureckis, 2004) and decision processes (e.g., Nosofsky & Palmeri, 1997).

Acknowledgments

Correspondence address: Dan Navarro, School of Psychology, University of Adelaide, SA 5005, Australia. E-mail: daniel.navarro@adelaide.edu.au, Tel: +61 8 8303 5265, Fax: +61 8 8303 3770, URL: http://www.psychology.adelaide.edu.au/personalpages/staff/ danielnavarro/. Variants on this paper have circulated for quite some time: I thank Kevin Murphy for the Kalman filtering code, John Paul Minda for making data available, and Nancy Briggs, Simon Dennis, Michael Lee, Rob Nosofsky, Mark Pitt, Matt Welsh and several anonymous reviewers for many helpful comments.

References

Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika, 56*, 567–587.

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

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A Study of Thinking*. New York: Wiley.

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

Feller, W. (1968). *An Introduction to Probability Theory and its Applications* (3rd ed). New York: Wiley.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering, Series D, 82*, 35-45.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112*, 500-526.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J. & Tirri, H. (2005). An MDL framework for data clustering. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 323–354). Cambridge, MA: MIT Press.

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

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

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush and E. Galanter (Eds) *Handbook of Mathematical Psychology, 1* (pp. 103-190). New York: Wiley.

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision, 8*, 229-254.

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

Medin, D. L. & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory, 7*, 241-253.

Minkowski, H. (1891). [On positive quadratic forms and on algorithms suggesting continued fractions]. *Journal für die reine und angewandte Mathematik, 107*, 278–297.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Navarro, D. J. (2006). From natural kinds to complex categories. In R. Sun & N. Miyake (Eds), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 621-626). Mahwah, NJ: Lawrence Erlbaum.

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.

Navarro, D. J. & Lee, M. D. (2005). An application of minimum description length clustering to partitioning learning curves. *The 2005 IEEE International Symposium on Information Theory.*

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. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition.* (pp. 218-247). New York: Oxford University Press.

Nosofsky, R. M. & Alfonso-Reese, L. A. (1999). Effects of similarity and practice on speeded classification response times and accuracies: Further tests of an exemplar-retrieval model. *Memory & Cognition, 27*, 78-93.

Nosofsky, R. M. & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

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.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185-200.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

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

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds), *Cognition and Categorization* (pp. 27-77). Hillsdale, NJ: Erlbaum.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories, *Cognitive Psychology, 7*, 573-605.

Schervish, M. J. (1995). *Theory of Statistics.* New York: Springer.

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. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 1411-1436.

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.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*, 251–260.

Tenenbaum, J.B. & Griffiths, T.L. (2001). Generalization, similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24*, 629-641 & 762-778.

Vickers, D. (1979). *Decision Processes in Visual Perception*. New York: Academic Press.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics, 19*, 326–339.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford, UK: Blackwell.