# Measuring category intuitiveness in unconstrained categorization tasks

Emmanuel M. Pothos[1], Amotz Perlman[1]

Todd M. Bailey[2]

Ken Kurtz[3]

Darren J. Edwards[1]

Peter Hines[4]

John V. McDonnell[5]

*in press*
**Running head**: category intuitiveness. **Word count**: 10,786

**Contact details**: [1]Department of Psychology, Swansea University, SA2 8PP, UK. email: e.m.pothos@swansea.ac.uk; [2]School of Psychology, Cardiff University, Cardiff CF10 3AT, UK. email: baileytm1@cardiff.ac.uk; [3]Department of Psychology, Binghamton University, Binghamton NY 13902-6000, USA. email: kkurtz@binghamton.edu; [4]Department of Computer Science, University of York, York YO10 5DD, UK. email: peter.hines@cs.york.ac.uk; [5]Department of Psychology, New York University, New York, NY 10003, USA. email: john.mcdonnell@nyu.edu.

## Abstract

What makes a category seem natural or intuitive? In this paper, an unsupervised categorization task was employed to examine observer agreement concerning the categorization of nine different stimulus sets. The stimulus sets were designed to capture different intuitions about classification structure. The main empirical index of category intuitiveness was the frequency of the preferred classification, for different stimulus sets. With 169 participants, and a within participants design, with some stimulus sets the most frequent classification was produced over 50 times and with others not more than two or three times. The main empirical finding was that cluster tightness was more important in determining category intuitiveness, than cluster separation. The results were considered in relation to the following models of unsupervised categorization: DIVA, the rational model, the simplicity model, SUSTAIN, an unsupervised version of the generalized context model (UGCM), and a simple geometric model based on similarity. DIVA, the geometric approach, SUSTAIN, and the UGCM provided good, though not perfect, fits. Overall, the present work highlights several theoretical and practical issues regarding unsupervised categorization and reveals weaknesses in some of the corresponding formal models.

## Introduction

Without concepts, human thought would be impossible as we know it. Concepts help us organize briefly and efficiently the information around us, but they are also at the heart of many abilities which we consider uniquely human, such as reasoning on the basis of abstract ideas (Murphy, 2004; Pothos & Wills, 2011). The question of how concepts arise is one of fundamental importance for our understanding of human behavior. Many concepts are taught, through language, social convention, or education. This tradition of supervised categorization inspired highly influential formalisms, such as prototype and exemplar theory (e.g., Hampton, 2000; Minda & Smith, 2000; Nosofsky, 1984; Vanpaemel & Storms, 2008). Equally, it seems that in many situations groupings can be constructed in an unsupervised manner, that is, without being guided by an external teacher signal. For some time now, researchers have been recognizing the importance of unsupervised categorization processes in the understanding of human concepts.

The focus of the present study is unsupervised categorization in the context of free sorting tasks such as the following: participants receive a set of schematic stimuli, presented individually on printed cards; they are asked to divide the stimuli in any way they like, with no constraints on the number of groups or the number of elements per group. That such a process is unsupervised is evident in that there are no external constraints to guide categorization; a participant can create any kind of groups he/she wants. Several researchers have employed free sorting tasks, mostly to examine the impact of various methodological variations on participant performance (Ashby et al., 1999; Handel & Preusser, 1969; Handel & Imai, 1972; Handel & Rhodes, 1980; Medin et al., 1987; Milton & Wills, 2004; Regehr & Brooks, 1995). For example, does it make a difference whether participants see all the stimuli at once instead of sequentially? Are there circumstances that

encourage participants to create classifications on the basis of a single stimulus dimension? This research has produced many important insights, even though the range of stimulus structures employed has been typically limited. One of the objectives of the present research is to motivate and test a wide range of stimulus structures.

We seem to have a natural tendency to organize information in the world. When exposed to a new domain, we implicitly or instinctively look to identify the basic 'kinds' that go together. This is a paradigmatic case of unsupervised categorization, though in adult thought it is often hard to separate out such unsupervised categorization processes from influences based on linguistic labels and existing categories. This is not to say that there are not everyday life situations when we engage in purely unsupervised categorization processes similar to those in the lab-based free sorting tasks: for example, arranging books in a bookcase, organizing administrative paperwork, archiving literature search articles, or arranging household items in a garage or garden shed. In all cases, the stimuli can be described with a set of dimensions (not all perceptual), so that there is a similarity structure for the stimuli. Also, in all cases there is a sorting problem, that of deciding which items go together. Such examples show the relevance of unsupervised categorization in limited problem-solving situations, but we contend that the impact of unsupervised categorization in human thought is both more profound and more pervasive.

A controversial issue in development psychology concerns the relation between linguistic and conceptual development. One view is that linguistic development guides conceptual development, as linguistic labels are employed to facilitate the acquisition of concepts. An alternative view is that children first develop concepts, so that at a later stage labels are matched to appropriate concepts (e.g., Nelson, 1974; Quinn & Eimas, 1986; Schyns, 1991). Such a view is supported by evidence that parent-child interaction may

involve limited or no corrective feedback, when it comes to children's inappropriate use of linguistic labels (e.g., Chapman et al., 1986; Nelson et al., 1993; see also, Brown & Hanlon, 1970; Demetras et al., 1986; Johnson & Riezler, 2002). The process which allows conceptual development in children is in some ways analogous to a free sorting task, in that in both cases it is recognized that some items go with others. Indeed, developmental psychologists have shown that children can perform free sorting tasks like the one described above (e.g., Gopnik & Meltzoff, 1997).

The sense that certain items go together, which is thought to drive behavior in free sorting experiments, appears relevant in perceptual organization as well. When we interpret a novel visual scene there is often a very strong intuition that certain elements form groups. For example, in the experiments of Compton and Logan (1993, 1999), participants were presented with arrangements of dots in two-dimensional spaces. In some cases, there was a very strong intuition about the presence of groups and most participants agreed in how the dots should be classified. This idea, that when a set of items can be classified in an intuitive way there should be more consistency in participants' classifications, is a key element of the present research as well. More generally, the link between perceptual organization and unsupervised categorization has been taken up by some researchers, who proposed models of unsupervised categorization based on perceptual principles (Compton & Logan, 1993, 1999; Pothos & Chater, 2002).

We can extrapolate this intuition of 'things going together' with adult concepts as well. The relative contribution of supervised (through language, social interaction etc.) and unsupervised processes in adult concepts is difficult to quantify (cf. Malt et al., 1999; Malt & Sloman, 2007, for assumptions about categories induced by linguistic labels and the impact of linguistic labels on categorization). But, we can observe that many of our categories

involve coherent collections of objects, that is objects which are similar to each other or, at the very least, make sense together (Murphy & Medin, 1985). What is the glue which binds together the members of a category? For example, why do we consider a category like 'chairs' as intuitive (coherent), a category like 'games' as less intuitive (in the sense that people disagree more about the membership of this category), and a category composed of 'babies, the moon, and rulers' completely nonsensical? We can call this the problem of what determines category intuitiveness and it is clearly a fundamental one for cognitive psychology. We would like to suggest that, at least part of the solution, relates to understanding performance in free sorting tasks. This is because there seems to be a fundamental equivalence between many of our concepts and the groups created in free sorting tasks, in that in both cases people recognize that certain items should be grouped together.

Research in supervised categorization has been much more extensive than research in unsupervised categorization. If one considers entirely unconstrained classification tasks with an aim related to category intuitiveness, there are few studies, which are methodologically limited (e.g., Compton, 1999; Pothos & Chater, 2002). Note that unsupervised categorization is not the same as unsupervised learning, though it is possible of course that the two processes are based on similar computational principles (after all, they are both instances of inductive inference). The former concerns the specific (empirical) objective of spontaneously grouping some stimuli together. The latter is more general and concerns all situations of generalizing from some initial stimuli without feedback (e.g., Billman & Knutson, 1996; Fiser & Aslin, 2005; Reber, 1967). Also, while there has been some work on unsupervised categorization, much of it is not appropriate for the study of category intuitiveness. For example, researchers have employed sequential or concurrent

presentation procedures for the stimuli. However, there seems to be a strong sense of category intuitiveness only in the case of concurrently presented stimuli. Equally, some researchers have asked participants to spontaneously group a set of stimuli into a specific number of categories (e.g., Medin, Wattenmaker, & Hampson, 1987; Milton, Longmore, & Wills, 2008). Such a procedure is less appropriate when studying category intuitiveness, since it can restrict participant performance. We employed an unsupervised classification task, with no constraints in the number of groups which could be created.

Part of the success of research in supervised categorization can be attributed to the existence of standard datasets (e.g., Medin and Schaffer, 1978; Shepard, Hovland, and Jenkins, 1961), specific dependent variables (e.g., classification probability of novel instances or speed of learning), and detailed computational comparisons between competing formal models (Minda & Smith, 2000; Nosofsky, 1990). With the present work, we try to make progress in unsupervised categorization in all these respects. First, we motivate a dependent variable appropriate for the study of category intuitiveness. Second, we specify a range of stimulus sets, created so as to contrast various factors possibly relevant in free sorting performance, and collect data from a large population sample. Third, we concurrently apply several computational models of unsupervised categorization. Medin et al. (1987) provide an eloquent statement motivating a modeling effort in unsupervised categorization (p.43): “The categories which people normally create and use represent a tiny subset of the many possible ways in which entities and experiences could be partitioned. Therefore, a central question is what basic principles underlie category construction.” The objective of the unsupervised categorization models is exactly this, to provide hypotheses about the computational principles (and mechanisms) which underlie category construction.

## The dependent variable

In *supervised* categorization, researchers typically study the probability with which novel instances are classified to the different trained categories. We think that having such a specific, simple dependent variable has facilitated the development of supervised categorization models. What is an appropriate empirical measure of classification intuitiveness? That is, under what circumstances can we say that a particular classification is psychologically more intuitive than another?

Consider Figure 1, assuming that such diagrams correspond to psychological spaces and each point to a physical stimulus. The A and B panels show two different classifications for the same stimulus set. Which classification is more intuitive? We expect that participants will produce the more intuitive classification more *frequently* (cf. Compton & Logan, 1993, 1999). Therefore, in comparing alternative classifications for the same stimulus set, higher relative frequency can be interpreted as higher intuitiveness. We can now extend this discussion for different stimulus sets. Let's assume that the classification in panel A is the *preferred* classification for this stimulus set (the one which is produced most frequently) and likewise for panel C. Inspection of these classifications leads to an impression that the classification in panel C is less intuitive (e.g., the clusters are closer to each other). In other words, in panel C there is a less striking/ obvious best way to classify the stimuli. Accordingly, we expect participants to disagree more on how the stimuli should be classified and (therefore?) that the preferred classification will be produced with a lower frequency, compared to the one in panel A.

Overall, we propose that the main dependent variable in unconstrained unsupervised categorization experiments should be the frequency of the preferred

classification (in fact, our results show that this variable correlates nearly perfectly with the number of distinct classifications produced for each stimulus set). Two qualifications are in order. First, we are not proposing that there is no other information in unsupervised classification results. Rather, our point is that this dependent variable is the most practical way with which to examine participant performance and formal models of unsupervised categorization. Second, how well this variable 'works' is ultimately an empirical issue and it may turn out that an alternative measure of category intuitiveness is better. However, there are no indications from the present results that this is the case.

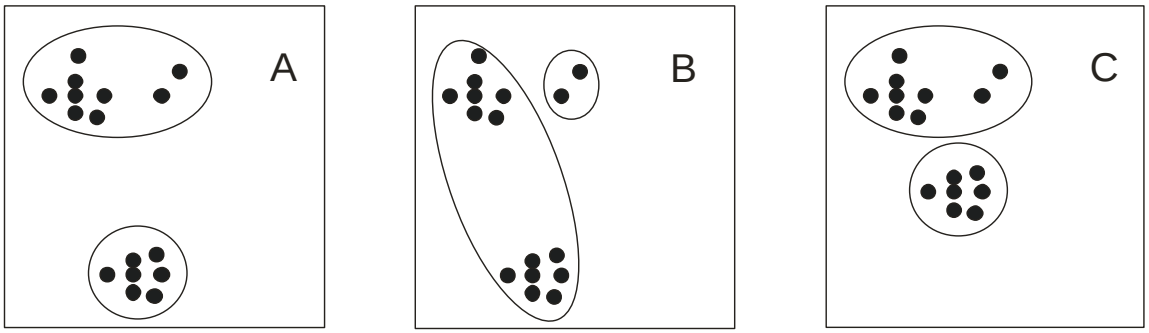


Figure 1. Hypothetical stimulus groupings that vary in intuitiveness. Classifications A should be more intuitive than classification B, since it involves more cohesive clusters. Classification A should also be more intuitive than classification C, since the clusters in the latter are less discriminable compared to the former.

## The empirical challenge

The space of possible classifications is vast. For ten stimuli there are about 100,000 classifications (Medin & Ross, 1997) and for 16 stimuli 10.4 billion possible classifications. It is remarkable that ordinarily only a tiny fraction of the possible classifications are psychologically relevant, and one has to wonder about just how intuitive a particular

classification has to be in order to stand out amongst so many alternatives. The large problem space could potentially lead to high performance variability, so that large sample sizes would be required. We adopted a population sample of 169 participants and assessed the adequacy of the sample size by comparing classification results across two conditions, manipulating the instructions, but employing identical stimuli. Note that in unsupervised categorization there are no aspects of the procedure to prevent idiosyncratic strategies and participant performance can be very variable. By contrast, in supervised categorization, the task forces conformity into participants' responses, since participants are required to learn a division of stimuli into the same categories.

Regarding the selection of stimuli, note first that in supervised categorization, researchers have been able to specify stimulus sets for which there are very precise expectations about how the predictions of categorization models would differ (e.g., Feldman, 2000; Medin and Schaffer, 1978; Shepard et al., 1961). In supervised categorization, an experimenter is free to specify classifications of arbitrary complexity. This is why it has been possible to, for example, identify classifications (such as the 5-4), which can discriminate between flexible computational models. In unsupervised categorization, a researcher cannot employ arbitrarily complex classifications. She is constrained to the study of classifications which are *likely* to be produced spontaneously. The design objective then becomes one of trying to anticipate the factors which are likely to influence the spontaneous preference for different classifications.

We identified four characteristics which are likely to be relevant. The first such characteristic is the proximity of clusters, whereby the straightforward intuition is that when clusters are closer to each other the corresponding classification would be less intuitive. The second one is the number of clusters. Is it the case that classifications having more clusters

are more intuitive than equivalent ones having few, larger clusters? Some unsupervised categorization models make this prediction (Pothos & Chater, 2002). The third characteristic is the relative size of clusters, since it is possible that classifications involving equally sized clusters may appear more intuitive than ones having clusters of varying sizes. The final characteristic is the spread (or tightness) of clusters, that is the degree of similarity between the members of each cluster. For perfectly separated clusters, this characteristic distinguishes between clusters which are 'tightly packed' versus ones which are less so (cf. Rosch &Mervis, 1975). We stress that the above characteristics are *hypotheses* of the factors which might be affecting the intuitiveness of classifications. The empirical results may show all or none of these factors to be relevant in determining differences in unsupervised categorization.

Finally, we chose to create two-dimensional stimuli, as this allowed more flexibility in manipulating the above characteristics. In unsupervised categorization tasks, with two-dimensional stimuli there is the possibility of emergent configural dimensions, or, conversely unidimensional biases (Medin et al., 1987). We think these are unlikely possibilities for our data, as the instructions emphasized to participants that they should employ both stimulus dimensions and because unidimensional biases have been demonstrated primarily in spontaneous classification tasks in which participants were asked to divide items into a particular number of groups (Pothos & Close, 2008). Our results (the observed classifications and multidimensional scaling analyses) confirmed that the experimenter-assumed dimensions were equivalent to the psychological ones.

## The modeling challenge

Categorization models can help us understand the underlying psychological process. We applied baseline versions of six unsupervised categorization models, with the view to identify the principles which appear most promising in the formalization of unsupervised categorization and category intuitiveness. This is the first comprehensive comparison of unsupervised categorization models (for limited previous efforts see Pothos & Bailey, 2009, and Pothos, 2007). One problem is that the implementation of the models has yet to converge (contrast with comparisons of exemplar and prototype models). Nevertheless, we can identify the essential aspect of each model (Table 1 provides a more detailed overview of the main model properties).

Most models of categorization involve an assumption that categories should be preferred if they maximize within category similarity, while minimizing between category similarity (Rosch & Mervis, 1975). We created a geometric approach model, so as to examine how far we can get in terms of modeling the empirical results with just this basic intuition. The DIVA model (Kurtz, 2007) is a connectionist model, which assumes that different statistical structure can be extracted from different categories. Thus, categories can be flexibly represented. The SUSTAIN model (Love, Medin, & Gureckis, 2004) also assumes that categories can be flexibly represented, but SUSTAIN explores flexibility in terms of a continuum between purely exemplar and purely prototypical representations. The Unsupervised Generalized Context Model (UGCM) emphasizes a different kind of flexibility, that of stimulus representation. According to the rational model (Anderson, 1991), categorization is about consistency with an underlying statistical model for category structure. Finally, according to the simplicity model (Pothos & Chater, 2002), categorization is a process of data compression. We will revisit the issue of model intuitions in the General Discussion. Once we have explored the models' application to the present empirical data

set, it will be possible to consider, and evaluate, in more specific terms the explanation about category intuitiveness from each model.

Table 1. A summary of the key characteristics of the models of unsupervised categorization considered in this work.

| | Geometric | DIVA | Rational | Simplicity | SUSTAIN | Uns. GCM |
|---|---|---|---|---|---|---|
| Formal principle | Similarity rules! | Connectionist auto-encoding: clusters reflect different statistical structure | Classification depends on Bayesian posterior of new stimulus given categories | More efficient information encoding of similarity→ more intuitive classification | Similarity with additional assumptions about dimensional selection | Similarity, but psych. space can be flexibly transformed. |
| Item presentation | Concurrent | Trial by trial | Trial by trial | Concurrent | Trial by trial | Concurrent |
| Sensitivity to similarity | Both within and between category | Flexible | Indirectly encoded via the likelihood term | Both within and between category | Formulated in terms of within category | Formulated in terms of within category |
| Relative size of clusters | Neutral | ? | Bigger favored | Bigger favored | ? | Neutral |
| Spread out clusters | Bad | ? | Should lead to less conservative extensions/ less probable classifications | Does not matter as long as clusters well-separated | Does not matter | Depends on the sensitivity parameter |
| Dimensional selection | None | Flexible | None | None | Fewer dimensions preferred | Flexible |
| Psychological space | Euclidean | N/A | N/A | Non-metric; only relative magnitude of similarities matters | Euclidean | Any power metric allowed; sensitivity parameter allows blurring |
| Dependence between dimensions | Independent | Any allowed | Independent | Independent | Independent | Depends on power metric (city block= independence) |
| Parameters | None | Spawning parameter | Coupling parameter | None | Tau parameter | Upper limit of the sensitivity parameter |

Notes: 'Spread out clusters' relates to the least similarity between any two items in the same cluster and so is different from within category similarity, which is an average. 'Dependence of dimensions' refers to whether the dimension values of cluster members are related. For example, if you see 'claws', do you expect to see 'fur' as well? The 'Number of parameters' row refers to parameters whose value has to be chosen by the experimenter.

### *DIVA*

The DIVA model (Kurtz, 2007) consists of a three-layer, feedforward neural network with a bottleneck hidden layer that is trained auto-associatively using backpropagation. The model operates by recoding the input at the hidden layer and then decoding (reconstructing the original input) in terms of different channels consisting of a set of output units (separate weights connect the units of each channel to the hidden layer). Each channel corresponds to a different category. In supervised learning tasks, DIVA assesses how well the input is reconstructed by each category (channel) and reconstructive success determines classification. That is, the model assumes that an exemplar belongs to a category if it can be reconstructed by the category. According to DIVA, a category can be any collection of exemplars which can be successfully reconstructed by the same category channel, which means that the exemplars must share some statistical regularity. For example, one category can correspond to all items that have value 1 on feature F1, or all items for which F1 and F2 are perfectly correlated, or all items such that feature F1 has value 1 unless features F2 and F3 each have value 0. Psychologically, this means that categories can be flexibly represented (e.g., as overall similarity or rules or any combination of critical features).

In unsupervised categorization, the model begins with a single channel and additional channels are recruited whenever the existing ones yield reconstructive errors below the spawning threshold (lower values make it easier for new categories to be created). After the evaluation of a stimulus and selection of a category channel, one supervised training trial with the input equal to the target is conducted and the error signal is applied only to the selected channel. A classification arises in the form of category channels that specialize in reconstructing sets of stimuli with similar properties.

### *A geometric similarity approach*

Rosch and Mervis (1975) extensively considered what determines category prototypes and the basic level of categorization, and their corresponding proposal, involving maximizing within category similarity, while minimizing between category similarity, has had a major influence in categorization research. This idea can be considered as a general proposal for category intuitiveness. We created a 'geometric approach' on the basis of this max within, min between similarity principle, so as to explore how far we can get in terms of modeling the empirical results with just this basic intuition. Note the geometric approach is not meant to be a proper model for the range of Rosch and Mervis's (1975) ideas; several researchers have tried to create such a model, showing that this is not a straightforward exercise (e.g., Corter & Gluck, 1992; Gosselin & Schyns, 2001; Jones, 1983; Medin, 1983; Murphy, 1991; note that these models are specified in terms of features, not distances in psychological space).

Within category similarity was measured as the average similarity of all stimulus pairs such that both stimuli are in the same category (computing an average is appropriate since in this way within category similarity does not depend on the size of the category). For categories with only one element, within category similarity was considered the similarity between the category member and its nearest neighbor (otherwise, the greatest preference is for clusters with a single item). Between category similarity corresponded to the average similarity of all pairs of stimuli such that the two stimuli in each pair were in different categories. The ratio of within category similarity to between category similarity was taken to be an index of category intuitiveness. A ratio approach for combining within and between category similarities has been employed in previous categorization work (e.g., Estes, 1994) and also guarantees that the resulting index of category intuitiveness will always be

sensitive to both within category similarity and between category similarity. By contrast, in a difference approach within category similarity would typically dominate (cf. Mervis & Crisafi, 1982; Murphy, 1991). This approach is really the most basic arithmetic description of the similarity structure in a partitioned stimulus set; the other models can be seen as ways to employ the same similarity information in more elaborate ways.

***The rational model***

The rational model is an incremental, Bayesian model of categorization, which classifies a novel instance into the category which is most likely given the feature structure of the instance (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2010). For example, the category 'cats' may be highly likely for a novel instance with features ('has fur', 'can purr'), because cats are fairly common (high prior probability) and these particular features are typical for cats. In the continuous version of the rational model, the probability of classification of a novel instance with feature structure *F* into category *k* depends on the product $P(k)P(F|k)$. $P(F|k) = \prod_i f_i(x|k)$, where *i* indexes the stimulus dimensions and *x* indicates the different values dimension *i* can take. Each $f_i(x|k)$ term corresponds to the probability of displaying value *x* on dimension *i* in category *k*, and feature values within a category are assumed to be independent (for a more general approach see Heller, Sanborn, & Chater, 2009). Finally, the probability that an object comes from a new category is given by $P(0) = \frac{1-c}{(1-c)+cn}$, where *n* is the number of classified stimuli, and *c* is the coupling parameter. Lower values of the coupling parameter make it less likely that dissimilar stimuli will be included in the same cluster and vice versa. We implemented the continuous version of the rational model, as described in Anderson (1991; Pothos & Bailey, 2009; Sanborn et al., 2010).

### *The simplicity model*

The simplicity model (Pothos & Chater, 2002) is an implementation of Rosch and Mervis's (1975) proposal within the information-theoretic framework of minimum description length (note that minimum description length and Bayesian updating can be formally related; e.g., Chater, 1996). The model first counts all similarities within categories and all similarities between categories. It then postulates that all within category similarities should be greater than all between category similarities. The more these constraints, and the more correct they are, the more intuitive the corresponding classification is predicted to be. The model takes into account the information cost of identifying and correcting erroneous constraints and also that of specifying a particular classification. Because of the latter, the simplicity model automatically determines the appropriate number of clusters. In other words, the simplicity model encodes the similarity information between the stimuli in a set either without or with categories and assesses category intuitiveness in terms of the difference between the respective codes.

### *SUSTAIN*

SUSTAIN aims to capture both supervised and unsupervised categorization in the same framework (Gureckis & Love, 2003; Love, Medin, & Gureckis, 2004), but of interest here is only its unsupervised component. SUSTAIN favors clusters of similar items. When a to-be-categorized item is presented to the model, it activates each existing cluster in memory, in a way based on the similarity of the item to each cluster. In addition, learned attention weights in the model can bias this activation in favor of dimensions which are more predictive for categorization. Note that SUSTAIN is biased to focus on a subset of stimulus dimensions. The most activated cluster is the one to which the new instance is assigned. If

no cluster is activated enough (as determined by the tau parameter), then a new cluster is created. Finally, category representation in SUSTAIN is adaptively determined and can be anything between the extremes of a purely exemplar representation and a purely prototype one.

### *The UGCM*

The UGCM (Pothos & Bailey, 2009) is a modification of the standard GCM (Nosofsky, 1984). With the standard GCM, behavioral data are typically fit by adjusting the model parameters until the classification probability GCM predicts for a test stimulus is as close as possible to the empirically observed one. An error term for the GCM can be computed as $\sum(O_i - P_i)^2$, whereby $O_i$ are the observed probabilities and $P_i$ are the predicted probabilities. In its unsupervised mode, the intuitiveness of a classification is estimated by considering how well each stimulus is predictable given the assignment of the other stimuli to their intended categories. Suppose we are interested in evaluating a classification for a set of stimuli, {1 2 3}{4 5 6 7 8 9} (the numbers are stimulus ids). We can consider each item in turn as a test item whose classification is to be predicted, and all the other items as training items whose classification is given. UGCM parameters are adjusted until the predicted classification probabilities for individual 'test' items are as close as possible to 100% for the classification of interest. The lower the sum of all the corresponding error terms, the more intuitive a classification is predicted to be, according to the UGCM. The parameters of the UGCM are automatically set so as to make the examined classification as intuitive as possible—but this does not imply that participants will likewise consider the classification intuitive. Thus, none of the parameters of the UGCM are manipulated with a view to achieve better correspondence with empirical results and so this model can be parameter free, as is the

geometric model and the simplicity model (but note that in practice we manipulated the upper limit for the sensitivity parameter).

## Experimental investigation

### *Participants and design*

Participants were 169 students at Swansea University, who took part for a small payment. Each participant classified nine stimulus sets, one after the other (the order of stimulus sets was randomized for each participant). A between-participants condition related to whether the stimuli were described in a neutral way (87 participants) or as real-world objects (82 participants).

### *Stimuli*

We created nine stimulus sets so as to reflect four intuitions about the considerations which might be relevant in unsupervised categorization: proximity of clusters, number of clusters, relative size of clusters, and spread of clusters (Figure 2). The 'two clusters' stimulus set is a baseline stimulus set of two well-separated, equally sized clusters. The 'three clusters' and the 'five clusters' stimulus sets involve greater numbers of (nearly) equally sized clusters. The 'unequal clusters' stimulus set alters the relative size of the two clusters. The 'spread out clusters' stimulus set involves two perfectly separated and equally sized clusters, but whose spread is broader compared to the 'two clusters' one. The 'poor two clusters' stimulus set retains only one aspect of classification structure which might correspond to classification intuitiveness: the existence of a simple linear boundary separating the two clusters. The 'random' stimulus set involves stimuli randomly sampling the available psychological space. Participants might be consistent in their classifications in such a case if

they simply employ some kind of proximity strategy anchoring each cluster on extreme stimuli. Finally, the 'embedded' stimulus set is there to examine whether participants might be able to pick out a fairly cohesive cluster amidst noise. The main intended characteristics of the stimulus sets are tabulated in Table 2. The proximity of clusters for a stimulus set was approximated as the average distance between all prototypes in a classification and the spread of clusters as the maximum distance between any two points in the same cluster.

The stimuli were made from two continuous dimensions, which were mapped to the length of a 'body' (horizontal dimension) and the length of the 'legs' after the joint (vertical dimension) of schematic spider-like stimuli (Figures 2, 3). By choosing such stimuli, both dimensions of physical variation were lengths, and so a Weber fraction in mapping the Figure 2 values to physical values could be assumed (8%; Morgan, 2005; note we do not claim that 8% increases correspond to the smallest noticeable differences in the physical dimensions, rather that they correspond to comfortably noticeable differences). For both dimensions, the actual lengths were between 40mm and 80mm. Also, the stimulus dimensions were such that it appears that no analytic effort is required to perceive them concurrently.

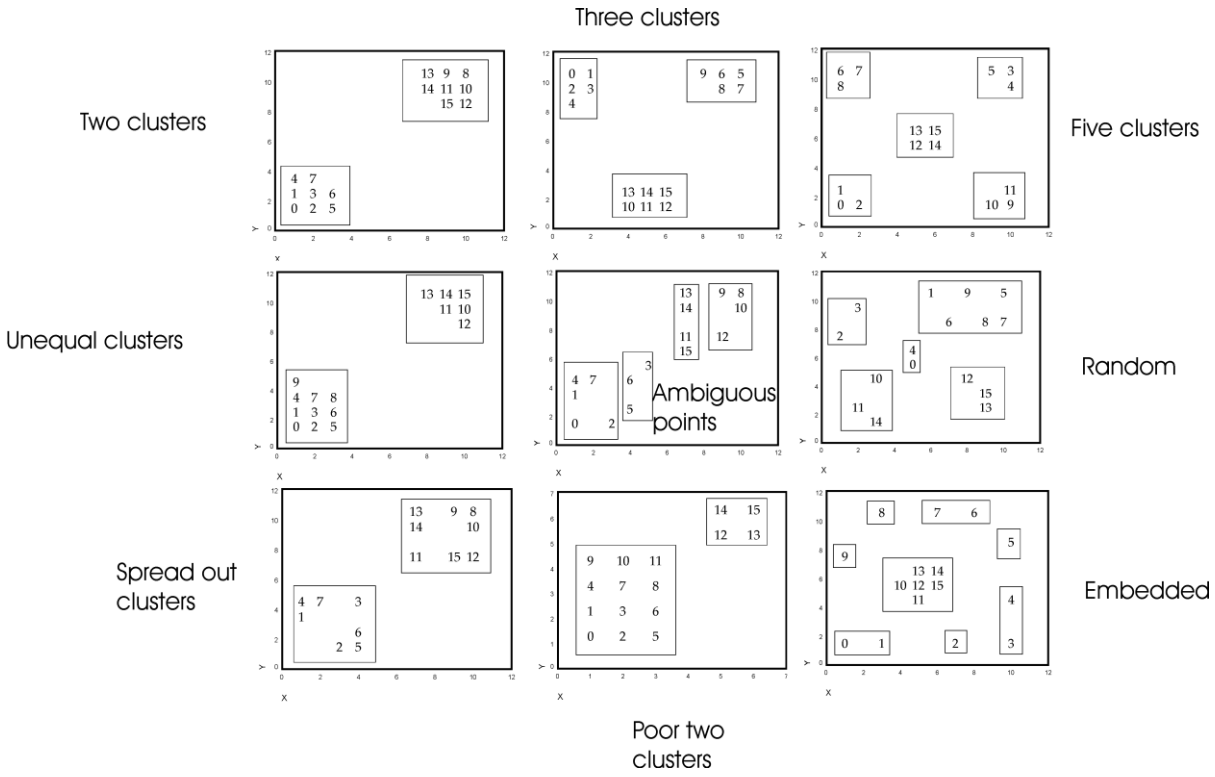


Figure 2. The nine stimulus sets employed in this study. The stimuli in a set are indexed by a number from 0 to 15. The grouping of points indicates the preferred classifications. For the Ambiguous Points and Embedded stimulus sets there were two and six, respectively, preferred classifications with the same frequency; we randomly chose one of these classifications for the figure.

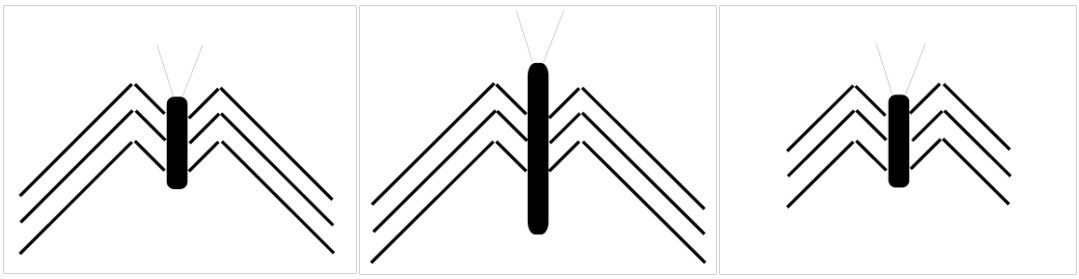

Figure 3. Some examples of the stimuli used.

Table 2. A summary of the characteristics of the nine stimulus sets employed in this study.

| | Number of well-defined clusters | Relative size of clusters | Between cluster separation | Av. dist. between cluster prototypes | Cluster spread | Max dist. between any points in the same cluster |
|---|---|---|---|---|---|---|
| Two clusters | 2 | Same | High | 10.25 | Medium | 2.83 |
| Unequal clusters | 2 | Different | High | 10.30 | High/ Medium | 3.61 |
| Spread out clusters | 2 | Same | High/ Medium | 8.66 | High | 4.24 |
| Three clusters | 3 | Approximately same | High/ Medium | 8.48 | Medium/ Low | 2.24 |
| Ambiguous points | 2 | Same | Medium/ Low | 6.10 | High | 4.24 |
| Poor two clusters | 2 | Different | Low | 4.61 | High/ Medium | 3.61 |
| Five clusters | 5 | Approximately same | Medium | 8.05 | Low | 1.41 |
| Random | 5 | Different | Low | 5.97 | High | 4.47 |
| Embedded | 8.2 | Different | Medium | 6.30 | High | 4.74 |

Note: The number of clusters for the Embedded stimulus set was computed as the average number of clusters for all the classifications which were produced with the same, highest frequency.

We examined whether the similarity structure of the actual stimuli conformed to the assumed coordinate representation. We created 12 stimuli which randomly spanned the coordinate space. We then asked 30 experimentally naïve participants (Swansea University students) to provide similarity ratings for these stimuli. Each participant was shown all possible stimulus pairs in this set of 12 stimuli, excluding identities: there were 12x12 – 12 (identities) = 132 trials. Stimulus presentation and response recording were computer controlled. Each trial started with a fixation point for 250ms, followed by the two stimuli in a pair one after the other for 1000ms each, followed by a 1-9 Likert ratings space. Similarity results from all participants were then averaged and subjected to a Euclidean distance, 2D multidimensional scaling (MDS) procedure (stress 0.115). The correspondence between the coordinate representation and the one based on similarity ratings can be quantified with the Orthosim procedure (Barrett et al., 1998), which allows the computation of various similarity indices between two sets of coordinates for the same set of items. Based on recommendations in the Orthosim documentation, we adopted a 'procrustes' approach, according to which the coordinate configurations to be compared are first normalized and rotated/ reflected to remove any of the arbitrariness in MDS solutions, and the 'double-scaled Euclidean distance' coefficient, for which 0 corresponds to complete dissimilarity, 1 to identity. This coefficient was 0.911, indicating very close correspondence between the assumed coordinates and the similarity-ratings based representation. This analysis shows that participants perceived the stimuli in the assumed way.

***Procedure***

Participants received the stimuli in each set in a pile; the stimulus sets were presented in a random order. The two dimensions of the stimuli were described and it was emphasized

that they were equally important. In one instructional condition, the stimuli were described as 'objects' and the two dimensions as 'rectangle length in the center and thin parallel lines length on the sides'. In the other, a scenario was presented saying how new spiders are discovered all the time around the world. Participants were then told about a recent expedition to the Amazon, during which several new spiders were identified. All these new spiders had broadly similar structure, but differed in terms of the length of their bodies and legs. In both instructional conditions, participants were told to consider the stimuli in each set independently, that is, as if the current stimulus set was the only one they had received. They were asked to spread the items in front of them and classify the items in a way that seemed natural and intuitive, using as many groups as they wanted, but not more than necessary. It was stated that more similar objects should end up in the same group, so as to prevent participants from adopting idiosyncratic strategies. Participants were told to indicate their groupings by arranging the objects in each group in separate piles. After the participant had left, the experimenter recorded the participant's classifications by noting the stimuli which were grouped together (each stimulus was associated with a number, written on the back of the stimulus).

***Empirical results – Frequency of the preferred classification***

A few classifications contained errors in how they were recorded (e.g., an item might be missing from the classification transcript). In some cases, we were able to conservatively carry out a correction, but where this was not possible a classification was not included in the analyses. Table 3 shows the classifications recorded correctly for each stimulus set (the max is 169). There was a problem with missing classifications only for the 'embedded' stimulus set. Our experience with the task is that it is more difficult to record classifications

which involve more erratic clusters. Therefore, it is not appropriate to scale the dependent variable (frequency of the preferred classifications), because the missing classifications are more likely to be ones which were more random. Notwithstanding this point, the results hardly change if some appropriate scaling is carried out.

Tables 3 shows that we observed more than 1100 unique classifications. This result was surprising, given that in some cases participants were asked to classify stimuli conforming to a simple two- or three-cluster classification, and informs the complexity of the task of analyzing this data. Indeed, many of the classifications produced appear to reflect random individual variation in classification strategy (note that restricting the number of allowed categories and employing fewer stimuli decreases performance variability; Milton et al., 2008). We would like to argue that the best way (or at least the most practical way) to make sense of this large and noisy dataset is by focusing on the frequency of the preferred classification in each stimulus set.

Note first that the instructional manipulation can be used to check whether the frequency of the preferred classifications is a robust variable. If classification intuitiveness drives participants' classification preferences, as we would like to assume, then we should observe the same pattern of results, regardless of whether participants received the realistic or neutral instructions. This was the case (Table 3). Correlating the frequency of the preferred classification for different stimulus sets, as a function of instructions, we obtained $r$=.92, $p$<.0005.

We then examined how the frequency of the preferred classification relates to measures of classification variability in each stimulus set. One such measure is the number of distinct classifications and another one is entropy, computed as $-\sum_i p_i log_2 p_i$, where $p_i$ is the probability of each classification in a stimulus set. Entropy is a measure of how easy it is

to identify a particular item in a set and so it is highest when there are many equiprobable classifications. The frequency of the preferred classification correlated highly with both measures (-.98 and -.99, in both cases $p<.0005$). Thus, the frequency of the preferred classification captures entirely two measures of variability in participants' classifications and nothing new would be gained by considering classification variability separately.

The frequency of the preferred classification dominates for all stimulus sets for which there is a salient classification structure: the frequency of the *next* most preferred classification is much lower than the frequency of the preferred one (Table 3). Most of the other classifications produced for a stimulus set would have a frequency of just one. The distributional properties of all the classifications in each stimulus set can be examined with the Rand index of classification similarity (Rand, 1971). It is computed as the pairs of items that are both in the same cluster, or both in different clusters, divided by all pairs, and so it ranges from 0 to 1, corresponding to totally different or identical classifications respectively. We examined the Rand similarity of all classifications produced for a stimulus set with the preferred one (in the few cases when there was more than one classification with the same highest frequency we selected one at random) and computed the mean and standard deviation of these Rand index values. The correlations between the mean Rand index and the standard deviation of the Rand values with the frequency of the preferred classification for each stimulus set were .31 and -.30 respectively; as both results were non-significant, we do not discuss them further.

Despite the encouraging analyses above, future work may reveal aspects of the data in unsupervised categorization tasks not captured by the frequency of the preferred classifications. An important advantage of the frequency of the preferred classifications is

that it is a *practical* dependent variable, in that it makes application of the computational models relatively straightforward.

Table 3. Summary of the empirical results of the study.

| Stimulus set | Frequency of most preferred[1] | Frequency of next most preferred[1] | Distinct classifications produced | Classifications recorded |
|---|---|---|---|---|
| Two clusters | 32 (20, 12) | 5 | 122 | 169 |
| Unequal clusters | 33 (17, 16) | 7 | 113 | 169 |
| Spread out clusters | 8 (5, 3) | 3 | 149 | 168 |
| Three clusters | 55 (32, 23) | 4 | 100 | 167 |
| Ambiguous points | 3 (1, 2) | 3 | 158 | 167 |
| Poor two clusters | 17 (11, 6) | 3 | 140 | 167 |
| Five clusters | 60 (27, 33) | 8 | 81 | 168 |
| Random | 3 (2, 1) | 2 | 158 | 168 |
| Embedded | 2 (1,1) | 2 | 148 | 154 |

Note: [1] 'Preferred' corresponds to the classification preferred by participants for the corresponding stimulus set. In parentheses we show the frequency of the preferred classification, as a function of the two types of instructions participants received; the first number corresponds to realistic instructions and the second to neutral instructions.

***Empirical results – key findings***

Our results indicate that the number of clusters does not affect classification salience. In fact, we were surprised to find that the preferred classifications for the stimulus sets 'three clusters' and 'five clusters' were produced with a higher frequency than the one for the 'two clusters stimulus set'. Relatedly, it appears that smaller, tighter clusters are preferred to more spread out ones, even in cases where the latter are very well separated.

The above intuitions can be made more precise with the information in Table 2. We carried out a regression analysis with frequency of the preferred classification for each stimulus set as the dependent variable and four independent variables: cluster spread, the number of well defined clusters, between cluster separation, and whether the size of the clusters is balanced or not (a binary variable). Accordingly, this regression analysis can show which characteristics of the stimulus sets contribute more to the frequency (and so salience) of the preferred classification. The overall regression was significant ($F(4,8)=29.80$, $R^2=.97$, $p=.003$) and the standardized betas were -.98, .19, -.16, .02 for the factors cluster spread, between cluster separation, size balanced, and the number of clusters respectively.

Finally, note that the preferred classifications in the 'three clusters' and 'five clusters' stimulus sets showed sensitivity to both dimensions of stimulus variation, since the corresponding three and five cluster classifications could not have been produced unless participants were attending to both dimensions. This finding (together with the MDS results) provides evidence against a hypothesis that there might have been an emergent feature driving classification performance and against the presence of a unidimensional bias in our experiments. This was as intended, given that the instructions emphasized that both dimensions should be considered (cf. Medin et al., 1987; Milton, Longmore, & Wills, 2004; Pothos & Close, 2008).

### *Modeling*

We assume that the higher the frequency of the preferred classification, the more intuitive this classification should be relative to alternative classifications for the same stimuli. The objective of the models is then to predict differences in the frequency of the preferred classification between stimulus sets. For example, why were participants fairly consistent in identifying the preferred classification in the 'two clusters' and 'unequal clusters' cases, but in the case of the, seemingly equivalent, two-group classification for the stimulus set 'spread out clusters', the preferred classification was identified with a much lower frequency?

We have standardized the comparison between the models as much as possible by employing the same test for the performance of each model. Specifically, the nine stimulus sets give us nine data points, the frequency of the preferred classification for each stimulus set. The application of the models aimed at predicting the differences between these frequencies. For example, if the preferred classification for one stimulus set was produced more frequently than the preferred classification for another stimulus set, does a model correctly predict that the former classification should be more intuitive than the latter?

All models which assume sequential presentation of the stimuli (DIVA, the rational model, and SUSTAIN) were applied by manipulating a single parameter, which has an equivalent function in each of the models. Specifically, for DIVA we manipulated the spawning threshold, for the rational model the coupling parameter, and for SUSTAIN the tau parameter. In all cases, these parameters correspond to whether the models are conservative as regards their category extensions or not. Regarding the models which assume concurrent presentation, the geometric approach and the simplicity model are

parameter free, while for the UGCM we manipulated the upper limit of the sensitivity parameter. The sections below briefly discuss some application details for the models.

### *DIVA*

We employed the latest version of DIVA (Kurtz, in preparation) which uses a linear activation function at the output layer and a default value for learning rate of 0.15. Other model details follow from Kurtz (2007): the hidden layer consisted of two nodes and the initial weights were randomized within a range of zero +/- 0.5. To simulate a spontaneous classification task, DIVA generated a sort based on two learning passes, i.e., two evaluations of each stimulus set (the second pass allows learning to be applied to examples that were experienced at the beginning of the first pass, i.e., before the structure of the sort had taken form). There was one free parameter: the spawning threshold (a single value of the spawning threshold was employed for all simulations). Pilot testing revealed good performance with a spawning threshold of 0.067. Note that similar performance was observed across a range of values around 0.067, and there was no evidence that another range of values would produce a qualitatively better fit to the human data.

The experimental stimuli were encoded as input patterns consisting of the two continuous dimension values, scaled by 1/10. For each of the nine datasets, DIVA was tested 170 times with random assignment of initial weights and random order of presentation for the two passes through the stimulus set. To generate category intuitiveness values, we counted the frequency with which the preferred classification was produced for each stimulus set (in the second pass). Where the frequency of the preferred classification is higher, DIVA predicts the classification to be more intuitive.

### *A geometric approach to similarity*

DIVA is a trial by trial model, so that a measure of category intuitiveness can be derived by employing several different orders and counting the number of times the empirically preferred classification is produced. The geometric approach assumes concurrent presentation of all items and, so a different approach is necessary. For each stimulus set, we considered the intuitiveness value of the empirically preferred classification (when there were more than one classifications which were produced with the highest frequency, we computed the intuitiveness value of all these classifications and considered the prediction of the model to correspond to the best possible intuitiveness value; an analogous approach was adopted for the simplicity model and the UGCM). Psychological similarity was equated with Euclidean distance.

Note that the geometric approach could be augmented by an attentional salience mechanism, such that instead of computing distance as $d(x,y) = (\sum_i (x_i - y_i)^2)^{1/2}$ it is computed as $d(x,y) = (\sum_i w_i \cdot (x_i - y_i)^2)^{1/2}$, where $w_i$ are dimensional weights. We considered all possible combinations of weights from 0 to 1 in 0.1 increments for the weights of the first dimension, so as to identify the dimensional weights leading to best correspondence with empirical results. In brief, this approach did not lead to more accurate predictions (see the Appendix). Perhaps retrospectively it is unsurprising that attentional selection does not aid the geometric approach. All stimulus sets were designed either so that their most obvious classification structure would be evident in two broadly equally weighted dimensions (e.g., ‘three clusters’, ‘five clusters’) or so that it would not matter whether the stimuli were classified on the basis of both dimensions or one (e.g., ‘two clusters’, ‘unequal clusters’).

### *Rational model*

The rational model is a trial by trial model of unsupervised categorization and so it was applied in a way analogous to that for DIVA and SUSTAIN. For each stimulus set, we counted the number of times the model produced the preferred classification across 5000 randomly selected stimulus presentation orders (and confirmed replicability of the results across a second set of 5000 random presentation orders). We explored a range of values for the coupling parameter, *c* (0.1, 0.25, 0.333, 0.4, 0.5, 0.667, 0.75, and 1) and the best correspondence between rational model results and predictions was identified for c=0.333. Note that we also examined the Gibbs Sampler (Sanborn et al., 2010) algorithm, but this procedure did not improve the rational model predictions.

### *Simplicity model*

The simplicity model assumes concurrent presentation of all stimuli. It was used to compute the intuitiveness value for the preferred classification for each stimulus set (the simplicity model has been designed to do exactly this). The model's predictions are expressed in terms of the ratio [codelength with categories] / [codelength without categories], as is typically the case.

### *SUSTAIN*

SUSTAIN assumes trial by trial presentation (Gureckis & Love, 2002; Love, Medin, & Gureckis, 2004). Stimuli were presented to the model as coordinate pairs, such that each dimension was scaled between 0 and 1. Initially, attention along both dimensions was set to be equal (initial lambda= 1.0) but SUSTAIN could adjust these values to emphasize differences along either dimension. Once stimulus presentation has finished, the

classification SUSTAIN produced was extracted by examining which stimuli activated the same cluster. For each stimulus set this process was repeated 170 times and we counted the number of times the preferred classification was produced. SUSTAIN was applied with only one free parameter, the tau parameter. We explored 40 potential values for tau, equally spaced between 0.18 and 0.9, choosing the solution (tau=0.586) that led to closest correspondence with empirical results. Note that SUSTAIN's operation depends on parameters other than tau, but these were not fitted in this application, rather we recycled a single set of global parameters from a previous study (the ones employed in the unsupervised fits of SUSTAIN in Love, Gureckis, & Medin, 2004).

Note that the proponents of SUSTAIN have argued that the tau parameter reflects individual differences in the preference for more conservative or liberal classifications. Thus, the SUSTAIN fit to a single stimulus set should involve a range of tau values, drawn from a particular tau distribution. However, applying SUSTAIN on the basis of a single tau value for all stimulus sets was consistent with how the other sequential models were applied. Our approach means that SUSTAIN predictions tend to show less variability and that they are a little less accurate, compared to an approach involving estimating the mean and variance of the tau distribution and using a range of different tau values even for the same stimulus set.

### *Unsupervised GCM*

We computed a sum of squared residuals value for the preferred classification for each stimulus set. When examining classifications which involve well-separated clusters, the UGCM can always find a value of the sensitivity parameter high enough to perfectly predict the target classifications. Accordingly, for the UGCM to produce meaningful predictions for the present stimulus sets the sensitivity parameter, *c*, had to be restricted (Pothos & Bailey,

2009). We examined several upper limits for the sensitivity parameter (0.001, 0.005, 0.01, 0.05, 0.075, 0.1, 0.15, 0.20, 0.25, 0.30. 0.35). The best fit was identified for an upper limit for *c* or 0.20. For each stimulus set and classification, we optimized the UGCM parameters 300 times with random parameter starting values (the starting value for a parameter was always within its allowed range). Without exception, the value for the sensitivity parameter identified in this way was equal to its upper limit. Note that these 300 runs of the UGCM per stimulus set are not like the number of presentation orders for DIVA or SUSTAIN, they are simply a measure of the computational effort for achieving UGCM global fits (even though the model has only one free parameter, it does have several other parameters which need to be optimized for each classification and stimulus set).

***Model comparison and discussion***

Predictions from all models are shown in Table 4, Figure 4. We first examined the Pearson correlations between the model predictions for category intuitiveness (Table 4) and the frequency of the preferred classification, for the different stimulus sets (employing Kendall's tau or Spearman's rho correlation coefficients leads to a nearly identical pattern of results). However, some of the models benefit from a free parameter and others do not. We therefore next computed the Akaike Information Criterion (AIC; Akaike, 1974) for each model, which allows meaningful comparisons between models with different numbers of parameters. When applied to Pearson correlations, $AIC_{r^2} = n \cdot ln\frac{1-r^2}{n} + 2k$. In this equation, *n* and *k* correspond to the number of data points and model parameters respectively, and *r* is the unadjusted correlation (Table 5). DIVA, the geometric approach, SUSTAIN, and the UGCM all provide a reasonable, but not perfect, account of the data, and the rational model and the simplicity model do not do as well (note that as the models are

not nested it is not meaningful to examine whether one model performs significantly better than another; e.g., Ashby, 1992; Ashby & O'Brien, 2008; Pitt et al., 2002). This is an important finding of the present research. Our analysis is the first extensive application of categorization models to spontaneous classification results and Table 5 illustrates that most models are in need of some revision. To illustrate the stimulus sets for which different models had difficulty, we normalized empirical results and model predictions on a zero (least intuitive) to one (most intuitive) scale and plotted the predictions of each model against empirical results (Figure 4).

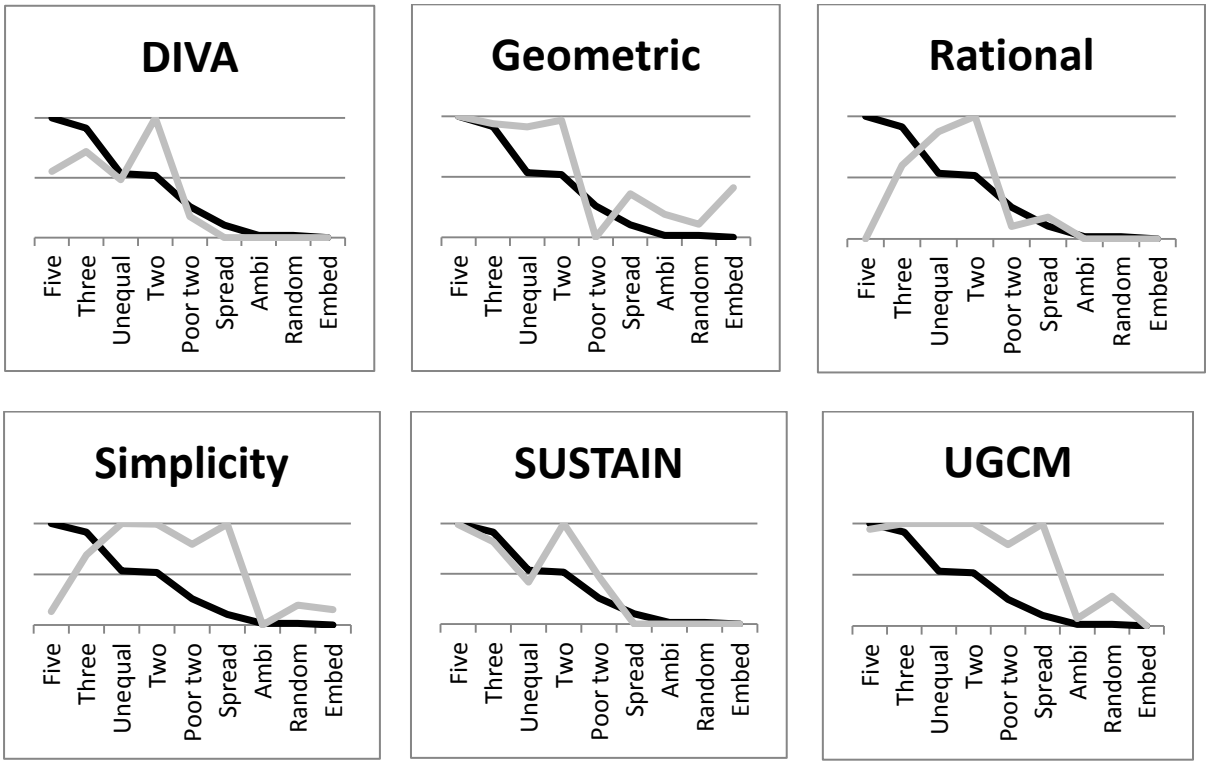


Figure 4. The empirical data and model predictions, normalized on a zero to one scale.

We next examined the relation between model predictions and the four stimulus characteristics, so as to better understand why some models performed better than others and, more generally, the properties of the different models (the general properties of models are not always obvious from their implementational details). Table 6 shows the

correlations between the intuitiveness predictions from each model and the four main characteristics of the stimulus sets, noting that the most and least important characteristics in determining category intuitiveness were cluster spread and the number of well-defined clusters respectively. The worst performing models, the simplicity model and the rational model, make predictions sensitive to the number of clusters. But, sensitivity to this characteristic of stimulus structure is 'misleading', since empirical results showed that it does not influence category intuitiveness. By contrast, the best performing models are most sensitive to cluster spread. Note that the sensitivity of the UGCM to between-cluster separation is a direct result of manipulating the upper limit of the sensitivity parameter (if this is set to a high enough value, then the UGCM shows no sensitivity to between-cluster separation). Note also that the rational model is insensitive to cluster spread. But, the general Bayesian intuition is that more spread out clusters would allow less conservative extensions (see, e.g., Stewart & Chater, 2002), so that tighter clusters should be preferred. This intuition was not born out in the rational model analyses; we further discuss this issue in the General Discussion Section.

Finally, the models differ considerably in their implementation, but sometimes such differences can be misleading. We therefore next examined how much different models converged in their predictions. Where there is a high degree of convergence, we may perhaps conclude that the performance of two models should be considered equivalent, despite any differences in AIC. This was the objective of Table 7 and Figure 5, where it can be seen that DIVA, the Geometric approach, and SUSTAIN all correlate highly with each other, and less so with simplicity and the rational model. This is an inevitably graded conclusion, but it does inform a distinction between better and worse performing models. We next created a similarity matrix for the models by computing Pearson correlations for

their predictions; i.e., the similarity between model X and model Y would be the absolute value of the correlation between the predictions of the two models (Table 7). We then derived a two-dimensional MDS representation, shown in Figure 5, so as to further illustrate model similarities (the stress of the MDS solution was 0.006).

Table 4. Model performance for the category intuitiveness results of the present study. The entries for each model correspond to model output linearly transformed onto a range from the lowest (2) to the highest (60) observed classification frequencies.

| Stimulus set | Human data | DIVA[1] | Geometrical approach[2] | Rational model[3] | Simplicity[4] | SUSTAIN[5] | UGCM[6] |
|---|---|---|---|---|---|---|---|
| Two clusters | 32 | 60 | 58 | 60 | 60 | 60 | 60 |
| Unequal clusters | 33 | 30 | 55 | 53 | 60 | 26 | 60 |
| Spread out clusters | 8 | 2 | 23 | 12 | 60 | 2 | 60 |
| Three clusters | 55 | 44 | 56 | 37 | 42 | 49 | 60 |
| Ambiguous points | 3 | 2 | 13 | 2 | 2 | 2 | 6 |
| Poor two clusters | 17 | 12 | 2 | 8 | 48 | 29 | 48 |
| Five clusters | 60 | 34 | 60 | 2 | 10 | 59 | 57 |
| Random | 3 | 2 | 8 | 2 | 13 | 2 | 19 |
| Embedded | 2 | 2 | 26 | 2 | 11 | 2 | 2 |
| Raw model output for: | | | | | | | |
| Five clusters | 60 | 47 | 0.145 | 0 | 74.9% | 168 | 0.25 |
| Embedded | 2 | 0 | 0.288 | 0 | 74.3% | 0 | 4.41 |

Note: the predictions from each model were based on: [1]The number of times the preferred classification has been produced in 170 runs of the DIVA model. [2]Average of all within category distances divided by average of all between category distances. [3]The number of times the preferred classification is produced across 5000 random presentation orders of the stimuli. [4]Percentage codelength of the preferred classification. [5]The number of times the preferred classification was produced in 170 runs of SUSTAIN. [6]Sum of squares error term.

## Note, this is the original table (from cat intuitiveness 6):

Table 4. The predictions for category intuitiveness from all the models of unsupervised categorization.

| Stimulus set | Human data | DIVA[1] | Geometrical approaches[2] | Rational model[3] | Simplicity[4] | SUSTAIN[5] | UGCM[6] |
|---|---|---|---|---|---|---|---|
| Two clusters | 32 | 85 | 0.153 | 5000 | 50.2% | 170 | 0 |
| Unequal clusters | 33 | 41 | 0.166 | 4374 | 50.0% | 71 | 0 |
| Spread out clusters | 8 | 0 | 0.300 | 885 | 50.2% | 0 | 0.02 |
| Three clusters | 55 | 61 | 0.160 | 3005 | 58.9% | 139 | 0 |
| Ambiguous points | 3 | 0 | 0.341 | 0 | 78.7% | 0 | 4.09 |
| Poor two clusters | 17 | 15 | 0.387 | 510 | 55.9% | 80 | 0.91 |
| Five clusters | 60 | 47 | 0.145 | 0 | 74.9% | 168 | 0.25 |
| Random | 3 | 0 | 0.361 | 0 | 73.1% | 0 | 3.13 |
| Embedded | 2 | 0 | 0.288 | 0 | 74.3% | 0 | 4.41 |

Notes: [1]The number of times the preferred classification has been produced in 170 runs of the DIVA model. [2]Average of all within category distances divided by average of all between category distances. [3]The number of times the preferred classification is produced across 5000 random presentation orders of the stimuli. [4]Percentage codelength of the preferred classification. [5]The number of times the preferred classification was produced in 170 runs of SUSTAIN. [6]Sum of squares error term.

Table 5. Correlations between model predictions (as shown in Table 7) and the number of times the preferred classification was produced for each stimulus set (the first column in Table 3).

| Model | Correlation | Free parameters | $AIC_{r^2}$ |
|---|---|---|---|
| DIVA | .799** | 1 | -26.93 |
| Geometric approach | -.834** | 0 | -30.48 |
| Rational model | .434 | 1 | -19.65 |
| Simplicity | -.207 | 0 | -20.17 |
| SUSTAIN | .896** | 1 | -32.39 |
| UGCM | -.709* | 1 | -24.06 |

Note: A '*' indicates two-tailed significance at the .05 level and a '**' at the .01 level. Lower values of $AIC_{r^2}$ indicate better correspondence between empirical results and model predictions.

Table 6. Correlating the intuitiveness predictions for each model with the four characteristics of the stimulus sets, as shown in Table 2. The characteristics have been ordered in terms of their empirical importance.

| Characteristic | Correlation with model predictions | | | | | |
|---|---|---|---|---|---|---|
| Cluster spread | SUSTAIN<br>-.93** | DIVA<br>-.78* | Geometr.<br>.74* | UGCM<br>.67* | Rational<br>-.32 | Simplicity<br>.12 |
| Between cluster separation | Geometr.<br>-.85* | Rational<br>.82* | DIVA<br>.70* | UGCM<br>-.66* | Simplicity<br>-.59 | SUSTAIN<br>.49 |
| Size balanced | Geometr.<br>-.44 | SUSTAIN<br>.41 | DIVA<br>.41 | UGCM<br>-.35 | Rational<br>.14 | Simplicity<br>-.03 |
| Number of clusters | Simplicity<br>.63 | UGCM<br>.57 | Rational<br>-.48 | DIVA<br>-.29 | SUSTAIN<br>-.21 | Geometr.<br>.06 |

Note: A '*' indicates two-tailed significance at the .05 level and a '**' at the .01 level.

Table 7. The correlations between the predictions of the models.

| | DIVA | Geometrical approach | Rational model | Simplicity | SUSTAIN | UGCM |
|---|---|---|---|---|---|---|
| DIVA | | -.85** | .80** | -.45 | .93** | -.68* |
| Geometric approach | | | -.67* | .30 | -.77* | .61 |
| Rational model | | | | -.74* | .55 | -.62 |
| Simplicity | | | | | -.30 | .78* |
| SUSTAIN | | | | | | -.69* |

Note: A '*' indicates two-tailed significance at the .05 level and a '**' at the .01 level.

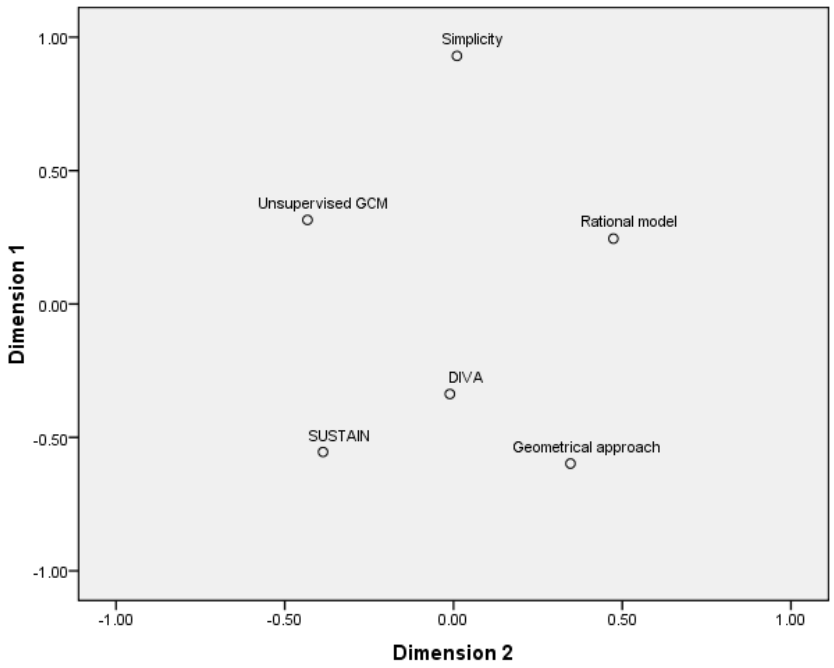


Figure 5. A schematic impression of which models make more similar predictions, for the stimulus sets employed in this study; smaller distance in this space reflects greater similarity in model predictions.

## General discussion

It is often the case that cognitive process cannot be guided by an external, supervisory signal, so that it is unsupervised. Research onto unsupervised cognition has been extensive and relates to most key cognitive processes (e.g., Chater, 1996; Elman, 1990; Feldman, 2009; Fiser & Aslin, 2005; Quinn & Eimas, 1986; Schyns, 1991). We have suggested that unsupervised categorization specifically is an important aspect of category intuitiveness, that is, our impression that it makes sense to place certain objects in the same category. Very little research has been done on unsupervised categorization with unconstrained free sorting tasks, perhaps because of the difficulty associated with organizing participant

responses into a meaningful dependent variable. We argued that the frequency of the preferred classification for a stimulus set is a suitable dependent variable.

The main empirical finding of the study was that the tightness of clusters appears to outweigh the separation between clusters, in determining category intuitiveness. This is a significant finding because current theorizing about the psychology of categorization frequently emphasizes the latter factor over and above the former. For example, Table 4 shows that many models managed to capture the relative goodness between the 'two clusters' and 'poor two clusters' stimulus sets (for which separation varies), but no model predicted a superiority for the 'five clusters' stimulus set over the 'two clusters' one. What is the source of participants' preference for tight clusters? With both adult and child participants, it is generally the case that more variable categories encourage more liberal category extensions (Hahn et al., 2005; Mareschal et al., 2002; Rips, 1989; Smith & Sloman, 1994; Stewart & Chater, 2002). We can thus speculate that classifications which involve tight clusters are more intuitive exactly because there is less ambiguity about cluster membership. If this intuition proves correct, it would favor categorization models for which classification depends directly on the distributional properties of categories (e.g., some revision of the rational model).

Category intuitiveness was influenced by cluster separation nearly as much as by whether clusters were equally sized or not. The importance of the latter characteristic in our data possibly follows from the result regarding cluster tightness. For example, compare the stimulus sets 'spread out clusters' and 'poor two clusters' for which the frequencies of the preferred classifications were 8 and 17 respectively. In the case of 'poor two clusters', perhaps higher category intuitiveness was observed because participants could easily identify the tighter, smaller cluster, which presumably then enabled a consistent

classification of the rest of the items in the stimulus set. Note that this speculation goes exactly against the predictions of the rational model and the simplicity model: both these models predict that category intuitiveness depends on cluster size (e.g., these models had difficulty predicting the superiority of the 'five clusters' stimulus set, compared to the 'two clusters' one, even though clusters were larger in the latter case).

Relatedly, the number of clusters proved to not affect classification intuitiveness at all. For example, the preferred classification in both the 'five clusters' stimulus set and the 'random' one had five clusters, but the category intuitiveness of the former was 60 and of the latter only three. From a psychological point of view, perhaps it makes sense that category intuitiveness is not affected by the number of categories, as this would mean that acquiring new, well-separated categories, does not interfere with the category intuitiveness of existing categories. But this finding, again, presents a challenge to formal models which depend on cluster size (because we wanted to have the same number of stimuli in each set, increasing the number of clusters decreases the number of stimuli per cluster). Moreover, most current categorization studies (modeling and experimental) involve only two clusters. The present findings recommend the importance of extensions with more clusters.

Finally, performance with the 'three clusters' and 'five clusters' stimulus sets showed that there was no unidimensional bias in participants' classifications (see also Milton & Wills, 2004; Milton et al., 2008; Pothos & Close, 2008). It is possible that unidimensional biases in category formation are more common with paradigms which require participants to form a specific number of categories (Murphy, 2004, p.129). Also, perhaps participants were encouraged to employ both stimulus dimensions in the present study because of the instructions encouraging them to do so. Note, however, that in previous work without such instructions, but with a similar unconstrained categorization task, we also failed to observe a

general unidimensional bias (Pothos & Close, 2008). In line with our present conclusion, intuition regarding natural categories suggests that they are not typically based on a single feature or dimension. Equally, it is important to explore whether our finding generalizes to more realistic categorization situations (e.g., more stimuli and categories). Of course, this point applies to all our empirical conclusions, though it is worth noting the consistency in the pattern of results for, e.g., the 'two clusters', 'three clusters', and 'five clusters' stimulus sets, for which tighter clusters led to increased category intuitiveness, regardless of the number of clusters.

The role of the computational models is to provide an explanation of psychological process in terms of formal principles. In an ideal world, the characteristics of category intuitiveness we empirically identified (such as the importance of cluster tightness), would follow naturally from a formal model of the categorization process. We applied six models, a geometric model (which was a simple formalization of Rosch and Mervis's, 1975, max-within, min-between similarity proposal), DIVA (Kurtz, 2007), the rational model (Anderson, 1991), the simplicity model (Pothos & Chater, 2002), SUSTAIN (Love et al., 2004), and the UGCM (Pothos & Bailey, 2009). This is currently the most comprehensive comparison of unsupervised categorization models, so it is interesting to examine how model differences in theory and implementation translate to differences in prediction. We standardized model application, by extracting the same measure from all models, that is numbers which can be interpreted as category intuitiveness. Note that some models are better suited to computing category intuitiveness than others (notably the simplicity one), but it turned out this did not make any difference in the accuracy of predictions.

In unsupervised categorization, a key intuition is that of Rosch and Mervis's (1975), that naïve observers should prefer categories which maximize within category similarity and

minimize between category similarity. The geometric model was an arithmetic expression of this intuition and it was meant to help explore how good a coverage of results can be achieved from just this basic intuition, without the additional elaborations of the 'proper', so to say, models. It turned out to be one of the three best performing models, together with SUSTAIN (Love et al., 2004) and DIVA (Kurtz, 2007). The success of the geometric model raises questions regarding the added value of the computational models of unsupervised categorization.

Both SUSTAIN and DIVA allow for flexible category representations and their success with the present data indicates that this is a theme which should be explored further in unsupervised categorization. Specifically, in SUSTAIN, at one extreme, category representation is analogous to that in exemplar models, at another extreme to prototype models. Category representations evolve adaptively, depending on the complexity of the learning problem. DIVA is based on an autoassociative connectionist architecture, so flexibility in category representation is consistent with the properties of such models. In DIVA, categories can be represented as overall similarity or rules or any combination of critical features, though it is not currently clear whether the full representational flexibility allowed by DIVA is needed in unsupervised categorization.

The UGCM (Pothos & Bailey, 2009) also provided a good account of empirical results. Its main characteristic is the flexibility it allows in the form of psychological space (and the way similarity is computed). According to the UGCM, categories are more intuitive if the stimuli in different categories are better separated. However, separation can be considered both with respect to the original stimulus representations and with the alternative representations which are allowed by the UGCM mechanisms (e.g., attentional weighting, the stretching or compression of psychological space, response-biasing of different

categories, and changes in both the metric structure of the space and the similarity function). Note that the UGCM's performance is dependent on employing a particular value for the upper limit of the sensitivity parameter. For low upper limits for *c* there is a high sum of squares error for the preferred classifications for all stimulus sets; this is because for low *c* values, the UGCM cannot differentiate between any clusters, however well separated. Conversely, for high upper limits for *c*, the clusters in all classifications can be made perfectly separated (note that all preferred classifications were linearly separable). For intermediate values for the upper limit for *c*, as the upper limit is gradually increased, the sum of squares error term drops more rapidly for the preferred classifications for the structured stimulus sets, than for the unstructured ones. Therefore, there is an optimal value for the *c* upper limit which allows a differentiation between the model predictions for the structured and unstructured stimulus sets. The success of the UGCM suggests a possible convergence between supervised and unsupervised categorization processes (cf. Colreavy & Lewandowsky, 2008).

The rational model (Anderson, 1991; Sanborn et al., 2010) understands categorization as a process of Bayesian inference. A category extension is acceptable if a novel instance is probable given the distributional characteristics of the category members and depending on the prior probability of the category as well (this prior depends on the category cardinality). The rational model did less well. We think the problem has to do with the category priors, which are set up so that the more the stimuli which have been seen by the model, the greater the penalty for the creation of new clusters. When there are several well-formed clusters, this seems to prevent the model from identifying the correct classification. For example, for the 'five clusters' stimulus set, with c=0.42, in the predicted

preferred classification, one of the intended clusters is broken up to clusters with individual items, but with c=0.43 the preferred classification consists of an all-inclusive category.

The simplicity model (Pothos & Chater, 2002) also did less well. The simplicity model assumes that categorization is a process of information compression: we organize experience into categories because this allows us to represent the same information in a more compact way. Such an informational approach implies that, as long as two categories are well-separated, it does not matter how much further separated they are and that classification intuitiveness should rapidly increase with increasing cluster size, for well-separated clusters. Especially the latter prediction was not supported in our results and some hierarchical version of the model might be more appropriate (cf. Hines et al., 2007).

The AIC values for model performance (Table 5) and the data on the correlations between model predictions (Table 7) led us to identify a class of well-performing models (geometric approach, DIVA, and SUSTAIN) and a class of poorly-performing models (rational model, simplicity model), with the UGCM in between and closer to the better-performing models. Note that in Table 6 SUSTAIN, DIVA, and the geometric approach are the three models which best capture the 'cluster spread' characteristic of the stimulus sets (this was the characteristic which affected empirical results the most). Thus, overall, our results indicate that a promising avenue for understanding the psychology of category intuitiveness is maximizing within category similarity, while minimizing between category similarity, and a process of adaptive/ flexible category representation (as embodied in DIVA or SUSTAIN).

Finally, the present research required us to consider which is a good dependent variable for studying category intuitiveness. Arguably, the success of the research tradition in supervised categorization is partly due to the fact that researchers agreed on what is a pertinent, practical dependent variable (classification probabilities for new instances). We

proposed that frequency of the preferred classification in different stimulus sets is an appropriate variable. First, it corresponds to a simple intuition regarding the raw empirical data, that is, that if more participants prefer a particular classification, then this classification is more likely be intuitive/ obvious. Second, it is a practical variable, in that it is both easy to measure and it is easy to employ in computational modeling. Third, it correlates extremely highly with measures of variability in participants' classifications, so that it fully captures an important characteristic of the raw empirical data. Fourth, the preferred classification for a stimulus set typically dominates participant performance, in that, for structured stimulus sets, its frequency is many times over the frequency of the next preferred classification. Finally, it appears that classifications other than the preferred one are just too noisy; our pilot examinations revealed little structure. Overall, it is of course possible that useful information could be extracted from a more detailed dependent variable. However, there was no evidence that this might be the case from our analyses.

In conclusion, the present research has identified several important challenges in the study of category intuitiveness. For example, there were methodological issues regarding the motivation of appropriate stimulus sets and measurement variables, empirical findings which contrasted intuition from formal categorization models, and the complexity of assessing categorization models which, though purporting to address the same cognitive process, vary widely in implementation. But, we have argued that category intuitiveness is a key aspect of our effort to understand the psychology of categorization and so deserving further consideration from the research community.

## Acknowledgements

This research was supported by ESRC grant R000222655 to EMP. We would like to thank Greg Ashby, Jerome Busemeyer, Nick Chater, Rob Goldstone, Todd Gureckis, Brad Love, Kim Levering, Amy Perfors, and Adam Sanborn for their comments and Paul Barrett for providing the Orthosim software, which can be obtained from his website: http://www.pbarrett.net/. A related preliminary report was made at the 2008 Annual Meeting of the Cognitive Science Society. The stimulus sets developed for this study were employed in another study reporting two supervised categorization experiments, matched to the present experiments, and comparing the results from these experiments to the frequencies of the preferred classifications (Pothos, E. M., Edwards, D. J., & Perlman, A. (in press). Supervised vs. unsupervised categorization: Two sides of the same coin? *Quarterly Journal of Experimental Psychology*).

## References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.

Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. Psychological Review, 98, 409-429.

Ashby, F. G. (1992). Multivariate probability distributions. In F. G. Ashby (Ed.), Multidimensional models of perception and cognition (pp. 1-34). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G. & O'Brien, J. B. (2008). The Prep statistic as a measure of confidence in model fitting. Psychonomic Bulletin & Review, 15, 16-27.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. Personality and Individual Differences, 25, 805-819.

Billman, D. & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 458-475.

Boucher, L. & Dienes, Z. (2003). Two ways of learning associations. Cognitive Science, 27, 807-842.

Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), Cognition and the development of language. New York: Wiley.

Chapman, K. L., Leonard, L. B., & Mervis, C. B. (1986). The effect of feedback on young children's inappropriate word usage. Journal of Child Language, 13, 101-117.

Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. Psychological Review, 103, 566-591.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. Memory & Cognition, 36, 762-775.

Compton, B. J. & Logan, G. D. (1999). Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation. Perception Psychophysics, 61, 1320-1335.

Corter, J. E. & Gluck, M. A. (1992).Explaining basic categories: Feature predictability and information. Psychological Bulletin, 2, 291-303.

Demetras, M. J., Post, K. N., & Snow, C. E. (1986). Feedback to first language learners: the role of repetitions and clarification questions. Journal of Child Language, 13, 275-292.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Estes, W. K. (1994). Classification and Cognition. Oxford: Oxford University Press.

Feldman, J. (2009). Bayes and the simplicity principle in perception. Psychological Review, 116, 875-887.

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

Fiser, J. & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. Journal of Experimental Psychology: General, 134, 521-537.

Gopnik, A. & Meltzoff, A. (1997). The Development of Categorization in the Second Year and Its Relation to Other Cognitive and Linguistic Developments. Child Development, 58, 1523-1531.

Gosselin, F. & Schyns, P. G. (2001). Why do we SLIP to the basic-level? Computational constraints and their implementation. Psychological Review, 108, 735-758.

Gureckis, T. M., Love, B.C. (2002). Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions. In Proceedings of the 24th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum: Hillsdale, NJ.

Gureckis, T.M. and Love, B.C. (2003). Towards a Unified Account of Supervised and Unsupervised Learning. Journal of Experimental and Theoretical Artificial Intelligence, 15, 1-24.

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. Memory & Cognition, 33, 289-302

Hampton, J.A. (2000) Concepts and Prototypes. Mind and Language, 15, 299-307.

Handel, S. & Imai, S. (1972). The Free Classification of Analyzable and Unanalyzable Stimuli. Perception & Psychophysics, 12, 108-116.

Handel, S. & Preusser, D. (1969). The Effects of Sequential Presentation and Spatial Arrangements on the Free Classification of Multidimensional Stimuli. Perception & Psychophysics, 6, 69-72.

Handel, S. & Rhodes, J. W. (1980). Free Classification: Element-level and Subgroup-level Similarity. Perception & Psychophysics, 28, 249-253.

Heller, K. A., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. Neural Information Processing Systems.

Hines, P., Pothos, E. M., & Chater, N. (2007). A non-parametric approach to simplicity clustering. Applied Artificial Intelligence, 21, 729-752.

Hubert, L. & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193-218.

Johnson, M. & Riezler, S. (2002). Statistical models of language learning and use. Cognitive Science, 26, 239-253.

Jones, G. V. (1983). Identifying basic categories. Psychological Bulletin, 94, 423-428.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. Psychonomic Bulletin & Review, 14, 560-576.

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

Malt, B. C. & Sloman, S. A. (2007). More than just words, but still not categorization. Cognition, 105, 656-657.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., and Wang, Y. (1999). Knowing Versus Naming: Similarity and the Linguistic Categorization. Journal of Memory and Language, 40, 230-262.

Mareschal, D., Quinn, P. C. & French, R. M. (2002). Asymmetric interference in 3- to 4-month-olds' sequential category learning. Cognitive Science, 26, 377-389.

Medin, D. L. (1983). Structural principles of categorization. In B. Shepp & T. Tighe (Eds), Interaction: Perception, development and cognition (pp. 203-230). Hillsdale, NJ: Erlbaum.

Medin, D. L. & Ross, B. H. (1997). Cognitive psychology. (2nd Ed.). Fort Worth: Harcourt Brace.

Medin, D. L. & Schaffer, M. M. (1978). Context Theory of Classification Learning. Psychological Review, 85, 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. Cognitive Psychology, 19, 242-279.

Mervis, C. B. & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. Child Development, 53, 258-266.

Milligan, G. L. & Cooper, M. C. (1986). A Study of the compatibility of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21, 441-458.

Milton, F. & Wills, A. J. (2004). The influence of stimulus properties on category construction. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 407-415.

Milton, F., Longmore, C.A. and Wills, A.J. (2008). Processes of overall similarity sorting in free classification. Journal of Experimental Psychology: Human Perception and Performance, 34, 676-692.

Minda, J. P., & Smith, J. D. (2000). 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.

Morgan, M. J. (2005). The visual computation of 2-D area by human observers. Vision Research, 45, 2564-2570.

Murphy, G. L. (1982). Cue validity and levels of categorization. Psychological Bulletin, 91, 174-177.

Murphy, G. L. (1991). More on parts in object concepts: Response to Tversky and Hemenway. Memory & Cognition, 19, 443-447.

Murphy, G. L. (2004). The big book of concepts. MIT Press: Cambridge, USA.

Murphy, G. L. & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. Psychological Review, 92, 289-316.

Nelson, K. (1974). Concept, word, and sentence: interrelations in acquisition and development. Psychological Review, 81, 267-285.

Nelson, K., Hampson, J., & Shaw, L. K. (1993). Nouns in early lexicons: evidence, explanations, and implications. Journal of Child Language, 20, 61-84.

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. (1990). Relations between exemplar-similarity and likelihood models of classification. Journal of Mathematical Psychology, 34, 393-418.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3-27.

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

Pothos, E. M. (2007). Occam and Bayes in predicting category intuitiveness. Artificial Intelligence Review, 28, 257-274.

Pothos, E. M. & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the Generalized Context Model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1062-1080.

Pothos, E. M. & Chater, N. (2002). A Simplicity Principle in Unsupervised Human Categorization. Cognitive Science, 26, 303-343.

Pothos, E. M. & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. Cognition, 107, 581-602.

Pothos, E. M. & Wills, A. J. (Eds., 2011). Formal approaches in categorization. Cambridge University Press.

Pothos, E. M., Perlman, A., Edwards, D. J., Gureckis, T. M., Hines, P. M., & Chater, N. (2008). Modeling category intuitiveness. In Proceedings of the 30th Annual Conference of the Cognitive Science Society, LEA: Mahwah, NJ.

Quinn, P. C. & Eimas, P. D. (1986). On categorization in early infancy. Merrill-Palmer Quarterly, 32, 331-363.

Rand, W. M. (1971). Objective Criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. Journal of Verbal Learning and Verbal Behavior, 6, 855-863.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 21–59). New York: Cambridge University Press.

Rosch, E. & Mervis, B. C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. Cognitive Psychology, 7, 573-605.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review, 117, 1144-1167.

Schyns, P. G. (1991). A Modular Neural Network Model of Concept Acquisition. Cognitive Science, 15, 461-508.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs, 75, whole no 517.

Smith, E. E. & Sloman, S. A. (1994). Similarity- versus rule-based categorization. Memory & Cognition, 22, 377-386.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 893–907.

Vanpaemel, W. & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. Psychonomic Bulletin & Review, 15, 732-749.

## Appendix. Augmenting the geometric approach with attentional weights (supplementary online material)

Dimensional weighting in the geometrical approach would correspond to:

Instead of computing distances as

$$d(x,y) = \left(\sum_i (x_i - y_i)^2\right)^{1/2}$$

To computing distances as:

$$d(x,y) = \left(\sum_i w_i \cdot (x_i - y_i)^2\right)^{1/2}$$

We explored a straightforward scheme, whereby we simply attempted to identify a set of attentional weights for all stimulus sets, which would lead to best fit. That is, attentional weights are free parameters (just one free parameter, since the attentional weights for the two dimensions are constrained to sum to 1). Also, we assume that participants employed the same attentional weights for all 9 stimulus sets. We could not have different free parameters for dimensional weights for each stimulus set separately, since in this way we would be introducing 9 free parameters for 9 data points.

To implement this scheme, we explored all combinations of attention weights from [0,1] to [1,0] in 0.1 steps. In the cases of the stimulus sets Ambiguous Points and Embedded, we examined the classification which was favored in the basic version (without attentional weighting) of the geometric approach.

The table below shows the correlations between the empirical data and the geometric approach predictions, for different values of the weight for the first dimension.

| **Correlation** | **$w_1$** |
|---|---|
| -0.7469 | 1 |
| -0.7678 | 0.9 |
| -0.7902 | 0.8 |

| | |
|---|---|
| -0.8081 | 0.7 |
| -0.8224 | 0.6 |
| -0.8323 | 0.5 |
| -0.8364 | 0.4 |
| -0.8301 | 0.3 |
| -0.805 | 0.2 |
| -0.7507 | 0.1 |
| -0.6089 | 0 |

The case for which $w_1$=0.5 corresponds to the case reported in the paper (modulo some minor rounding error), for which the dimensional weights of the two dimensions are equal. It can be seen that the highest possible correlation (-0.8364) is only very slightly higher than the correlation without dimensional weighting (-0.8323). This analysis clearly shows that this attentional weighting scheme does not improve the ability of the geometric approach to account for the empirical results.