ELSEVIER

Cognitive Science 26 (2002) 303–343

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

A simplicity principle in unsupervised human categorization

Emmanuel M. Pothos[a,*,1], Nick Chater[b]

[a]Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
[b]Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

Received 16 January 2001; received in revised form 21 September 2001; accepted 20 February 2002

Abstract

We address the problem of predicting how people will spontaneously divide into groups a set of novel items. This is a process akin to perceptual organization. We therefore employ the simplicity principle from perceptual organization to propose a simplicity model of unconstrained spontaneous grouping. The simplicity model predicts that people would prefer the categories for a set of novel items that provide the simplest encoding of these items. Classification predictions are derived from the model without information either about the number of categories sought or information about the distributional properties of the objects to be classified. These features of the simplicity model distinguish it from other models in unsupervised categorization (where, for example, the number of categories sought is determined *via* a free parameter), and we discuss how these computational differences are related to differences in modeling objectives. The predictions of the simplicity model are validated in four experiments. We also discuss the significance of simplicity in cognitive modeling more generally.

Keywords: Psychology; Concepts; Machine learning; Mathematical modeling

1. From perception to unsupervised categorization

Confronted with an unfamiliar sensory scene, such as a novel computer game, or material viewed in a microscope, we can rapidly interpret the scene in terms of different kinds of groups. This can be viewed as a process of perceptual organization, i.e., finding patterns and structures in

[*] Corresponding author. Tel.: +44-30106726747.
E-mail addresses: e.pothos@ed.ac.uk (E.M. Pothos), nick.chater@warwick.ac.uk (N. Chater).
[1] Tel.: +44-02476523537.

0364-0213/02/$ – see front matter © 2002 Cognitive Science Society, Inc. All rights reserved.
PII: S0364-0213(02)00064-2

the sensory input. But it can also be viewed as a process of unsupervised categorization, because its output involves a grouping of the perceived elements. In this section, we wish to further establish the relationship between perceptual organization and unsupervised categorization, so as to utilize theoretical insights in the former as a basis for computational modeling of the latter.

If we perceive a set of objects differently, then we will categorize them differently as well; in other words, categorization is defined as an operation on the perceived similarity structure of a set of objects, so that the influence of perceptual organization on categorization is taken for granted. However, what is the evidence that categorization affects perceptual organization? For example, Goldstone (1994) presented a series of studies whereby it was shown that categorizing a set of items in different ways appeared to affect the relative similarity of these items, in a way that goes beyond simple predifferentiation (Gibson, 1991), or preexposure (Hall, 1991) effects for the stimuli (that is, effects relating to differences in processing the stimuli by virtue of exposure to them, not categorization). As another example, there is an extensive literature on categorical perception (Harnad, 1987), showing that when a category boundary is introduced in a dimension of variation for a set of stimuli this results in the stimuli becoming more discriminable along this dimension at the category boundary, and less discriminable on either side of the boundary (but see Massaro, 1987).

There has recently been significant theoretical effort to motivate the idea that perception and categorization cannot be understood as independent cognitive processes. Schyns (1998) argued that the similarity of object recognition and categorization can be best explained by observing how both processes involve matching object information in memory. Schyns (1998) utilized this intuition to suggest an approach to understand both processes in terms of "diagnostic recognition." Schyns, Goldstone, and Thibaut (1997) complement the work of Schyns (1998) by suggesting that new features will dynamically arise so as to make a set of categories more diagnostic; one of the consequences of this feature creation process is that the perception of the categorized objects is altered (see also Goldstone, 2000). More directly, Compton and Logan (1993, 1999; see also, van Oeffelen & Vos, 1982) examined how the proximity law from the Gestalt approach to perceptual organization could used as the basis for predicting perceptual grouping behavior.

The above studies amply suggest that perceptual organization and some aspects of categorization might reflect equivalent cognitive processes. Thus, it might be possible to model both perceptual organization and unsupervised categorization within the same computational framework. In perception, the main theoretical issue concerns how to choose between different possible organizations (interpretations) of sensory stimulus; this has been addressed mainly *via* two competing principles. The first, initiated by von Helmholtz (1910/1962), advocates the *likelihood principle*: sensory input will be organized into the most probable distal object or event consistent with that input. The second advocates what Pomerantz and Kubovy (1986) call the *simplicity principle*: the perceptual system is viewed as finding the simplest, rather than the most likely, perceptual organization consistent with the sensory input. The relevance of simplicity in perceptual organization has arisen in alternative forms, as well. For example, the physiologist Barlow (1974) advocates simplicity in terms of a drive to reduce redundancy in the representation of the stimulus: "[In perception] in the absence of *a priori* expectations about the structure of the world, the only means we have to understand it, is to encode it with as little redundancy as possible." Chater (1996; see also, Pomerantz & Kubovy, 1986;

Mach, 1959/1906) showed that the simplicity and likelihood principles are equivalent, on the basis of very general assumptions. Simplicity will be the basis for the present modeling effort in unsupervised categorization. At a more general level, there are theoretical reasons to suggest simplicity as an appropriate cognitive principle (Chater, 1999), beyond the domain of perception and categorization, but we will defer this discussion until the end.

At an intuitive level simplicity predicts the interpretation of the sensory input that enables us to encode it as briefly as possible. In a simplicity framework, the notions of "interpretation" and "encoding" are central: the sensory input is some information that the cognitive system must encode, whereby encoding is processing in a way that the information will be useful in the prediction of other compatible information, etc. (if you have recognized an object as an 'apple,' you would like to be able to recognize other objects as 'apples' as well). Interpretation is a theory about the structure of the information: if we have a sequence like "ababababab" we could interpret it as "5 × (ab)"; but, clearly, there are many alternative interpretations (e.g., "a, 2 × (baba), b"). In general, theories differ in terms of how complicated they are (how easy or difficult it is to specify them) and also in terms of how complicated is the specification of the data with a given theory. The simplicity principle provides us with a specific guide in choosing a particular theory (interpretation) for some data (information). The preferred theory minimizes the sum of (1) the complexity of the theory and (2) the complexity of the data when encoded with the theory; also, we cannot identify a theory for the data when the sum of the above complexities is greater than the complexity of the data without a theory.

The simplicity principle has a very straightforward application in unsupervised classification. In categorization, we have a set of objects or, more specifically, information about how similar different objects are to each other. For example, a banana and an apple will be less similar to each other than an apple and a pear, but more similar to each other than a banana and a boat. In perceptual organization grouping appears to be grounded primarily on physical similarity (Pomerantz, 1981); that is, the groups that are recognized as part of the process of encoding a visual scene reflect collections of similar objects. We thus suggest that spontaneous grouping in unsupervised categorization likewise has a component that reflects categories that maximize within category similarity, while minimizing between category similarity (Rosch & Mervis, 1975; we discuss in future directions other possible influences on unsupervised categorization, most notably general knowledge). This allows us to view categorization as imposing default constraints on the similarity relations between a set of objects: we suggest that having a category 'banana and apple' and a separate category 'boat' implies, by default, that the similarity of bananas and apples is higher than that of bananas and boats. These constraints could help us encode briefly the similarity structure of a set of objects if they are numerous and, in general, correct. The classifications that allow a brief encoding for the similarity structure of a set of objects will be the ones that are preferred, according to the simplicity principle (note that one can prove this objective to be compatible with other stated objectives of categorization, such as feature predictiveness, or accurate classification of new objects).

1.1. Relation to other categorization research—supervised categorization

Categorization is one of the most extensively researched areas in cognitive psychology, giving rise to a host of influential models. For example, in definitional accounts of concepts (e.g., Bruner, Goodnow, & Austin, 1956; Katz, 1972; Katz & Fodor, 1963), categories are characterized by necessary and sufficient conditions that determine which items are category members (see Pothos & Hahn, 2000, for a recent evaluation). In exemplar theories (e.g., Brooks, 1987; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1989, 1988a, 1988b, 1985), a concept is represented by a set of known instances of that concept; new instances are therefore assigned to different categories in terms of their similarity to the members of each category. In prototype theories assignment is also determined by a similarity process, this time to the prototypes of each category, where prototypes encapsulate some measure of central tendency across the exemplars of a category (e.g., Homa & Vosburgh, 1976; Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968; Reed, 1972). In the general recognition theory (Ashby & Perrin, 1988) categorization effects are understood in terms of intrinsic noise properties of perception and representation: learning a set of categories is assumed to be equivalent to trying to estimate an optimal decision boundary in some psychological space that enables discriminating the instances corresponding to the different categories (see Ashby & Alfonso-Reese, 1995 and Nosofsky, 1990 for a comparison of the general recognition theory, exemplar, and prototype models).

An obvious question is how the present model can relate to this research tradition. We attempt to explicate this in terms of a distinction between supervised and unsupervised classification, and arguing that these two types of classification have complementary explanatory objectives. All the models mentioned above are models of supervised categorization, whereby there is a ‘correct’ way of dividing up items into categories; the learner must infer the underlying category structure, in order to generalize as successfully as possible in categorizing new items. By contrast, in unsupervised categorization, which is the focus of the present work, there is no pre-specified classification; rather, the learner identifies a classification for a set of objects that is intuitive or natural.

Experimentally, in supervised categorization participants are typically presented with a set of artificial items, labeled in a some way by the experimenter; their task is to correctly learn the labels of the items. Classification of new instances and category extensions are issues that frequently arise in the context of supervised classification. The paradigm case of supervised learning in daily life is learning the meaning of a new word (i.e., the concept corresponding to the meaning of the word), from examples of items to which the word applies (this is learning by “ostension”). For example, a child might be told that certain objects are called apples while others oranges; her task would be to infer enough about the category structure of the concepts “apples” and “oranges” so as to correctly classify new instances. The above is to be contrasted with unsupervised classification. Experimentally, unsupervised classification would be investigated by examining spontaneous groupings of participants; the real-life analogue of this process would be situations whereby we seem to spontaneously recognize groupings of objects, when there is no *a priori* category structure in place.

One might claim that all our concepts reflect semi-arbitrary classifications, learned *via*, e.g., the use of linguistic labels; this would make the study of unsupervised categorization superfluous. There are many reasons why our conceptual structure is unlikely to be established exclusively on the basis of supervised categorization mechanisms. First, note that learning categories in supervised categorization typically assumes that people have seen enough exemplars from the different groups to be able to infer the boundaries between them. However, in learning new words, both children and adults often generalize effectively from a small number of examples (e.g., Feldman, 1997; Tenenbaum & Xu, 2000). This suggests that supervised learning of linguistic categories may be guided by rich prior constraints on what categories are plausible; and unsupervised learning provides a potentially important source of such constraints. Second, while there is some variation in the categories used in different cultures, there also appear to be strong commonalities between schemes of categorization (e.g., López, Atran, Coley, Medin, & Smith, 1997). Thus, in some cases at least, unsupervised categorization may provide the groundwork on which supervised learning can occur.

1.2. Related previous empirical work—unsupervised classification

The fundamental issue in unsupervised classification is the extent to which people recognize certain classifications as more natural or intuitive than others. There have seen several studies in which participants are typically asked to sort items into groups under different experimental conditions (e.g., Zippel, 1969; Imai & Garner, 1965). The objective of this research tradition (called free classification) is not directly to predict spontaneous classifications, but rather to identify the factors that appear to influence performance in sorting tasks, such as different types of instructions/experimental procedures and the structure of the stimuli (e.g., whether they are made of integral or separable dimensions, and the extent to which this affects the number of dimensions used in the classification task; e.g., Handel & Preusser, 1970; Smith & Baron, 1981; Wills & McLaren, 1998; Kaplan & Murphy, 1999).

For example, Handel and Preusser (1969) report that when a set of stimuli are presented simultaneously participants are likely to classify these spontaneously in similar ways, regardless of the actual spatial arrangement of the stimuli. On the contrary, when the stimuli are presented sequentially, the particular order used would affect the way they are classified. Handel and Imai (1972) further provide experimental results showing that when the stimuli were created from a set of *separable* dimensions people would classify them on the basis of just one of these dimensions, but when the stimuli were made from *integral* dimensions spontaneous sorting would reflect an overall similarity bias (that is, maximizing within category similarity, while minimizing between category similarity). More recently Regehr and Brooks (1995; see also Medin, Wattenmaker, & Hampton, 1987a) argued that spontaneous classification is most frequently performed on the basis of a single dimension, unless participants are encouraged to compare pairs of stimuli (but note that Regehr and Brooks always asked their participants to classify the given stimuli into two categories, in contrast with the standard free classification task where no constraints are imposed). Compton and Logan (1993, 1999) used a more unconstrained classification procedure: participants were presented with arrangements of dots and were asked to divide them into any groups in any way they liked. Compton and Logan (1993, 1999) used this procedure to examine the plausibility of various parameters, embodied in a model of perceptual grouping (for example, how proximity between elements in a perceptual pattern would affect a perception of the elements as being in the same group).

Overall, this research has amply demonstrated that people presented with a set of stimuli would spontaneously classify these stimuli on the basis of their similarity in fairly consistent ways. However, the emphasis in this unsupervised classification literature is to identify manipulations that can influence performance in spontaneous sorting, not to actually *predict* the particular way in which people would classify a set of items (the Compton and Logan studies are exceptions that will be further discussed below). By contrast, the objective of the present study is to enable prediction of the preferred classification for a set of objects, on the basis of information about how people perceive these objects. In this sense, we expect that when naïve observers produce one-dimensional sorts they must be perceiving the corresponding objects in an analogous way. It has to be noted that an alternative possibility is that classifications based on a single dimension reflect some kind of a parametric bias about category shape. With future empirical work we hope to evaluate these possibilities.

1.3. Related previous theoretical work—category coherence

What makes the category of birds or cups a coherent category, but disallows a non-sensical category consisting of dolphins born on Tuesdays together with pink tulips within 20 miles of London, and the Eiffel Tower? Category coherence concerns the core topic of this research, that is the observation that certain groupings appear to represent better concepts than others. This is a problem of unsupervised categorization, as it relates to how categories originate—a process which, necessarily, cannot be guided by a ‘supervisor.’

There have been several hypotheses about what constitutes category coherence. For example, it has been pointed out that some categories, typically artifacts, are grouped together by their common function; hence corkscrews are categorized together because they all have the function of opening bottles, even though their appearance, mechanism, size, color and so forth are widely varying (Barsalou, 1985). Of more relevance to the present work, explaining some aspects of category coherence by similarity would imply that categories are held together by the fact that the items they contain are judged to be similar to each other (Rosch, 1975; Wittgenstein, 1957; but see Goodman, 1972, and Quine, 1977). According to this viewpoint, ‘bird’ is a coherent category because birds are similar to each other.

The most dominant view on category coherence has been the hypothesis put forward by Murphy and Medin (1985; see also Gentner & Brem, 1999; Lakoff, 1987; Medin & Wattenmaker, 1997), that theoretical knowledge provides the “glue” holding sets of exemplars together. Their view was that a concept is a lot more than a collection of features, or similarity information: a concept is an element of people’s “naïve theories” about the world. For example, the concept “water” could not be explained just in terms of its chemical structure, similarity to other objects, prototype representation, etc. Rather, it has a particular “naïve” meaning in our everyday life, and this is how it is psychologically grounded. In support of this notion, it has been shown that young children can generalize on the basis of theoretical knowledge, rather than just perceptual similarity (e.g., Gelman & Wellman, 1991). For instance, they would generalize biological properties such as “having a spleen” between apparently dissimilar animate creatures (e.g., a person and a worm) in preference to generalizing to an inanimate object (e.g., a toy monkey, which is perceptually much more similar to a person than a worm is).

Murphy and Medin’s (1985; see also Medin & Wattenmaker, 1997) arguments above constituted an influential line of reasoning as to why a model of conceptual coherence cannot be grounded on similarity alone.[1] The simplicity framework we present in principle can provide an account of unsupervised categorization in terms of both background knowledge and similarity information (see Section 10.3). However, general knowledge effects have been notoriously difficult to formalize (Dreyfus & Dreyfus, 1986; Heit, 1997; Heit & Bott, 1999; McDermott, 1987; Oaksford & Chater, 1991, 1998; Pickering & Chater, 1995). In this work we avoid influences of general knowledge by using novel, abstract, perceptual stimuli (a strategy employed in studies of supervised categorization as well; e.g., Ashby & Perrin, 1988; Nosofsky, 1989).

1.4. Related previous theoretical work—basic level categories

A possible relation can be established between basic level categorization and unsupervised classification. Rosch and Mervis (1975) noted that out of the hierarchy of more or less general categories into which we may place an item (as being a Scottish highland terrier, a terrier, a dog, an animal, a living thing, and so on), there appears to be a privileged 'basic' level. This seems to be the default level for identifying new objects with linguistic labels (seeing Fido, the default is that we say or think "A dog!" rather than "An animal!" or "A terrier!"). The notion of basic level categories has been supported from a range of converging sources of evidence. For instance, basic level categories lead to rapid picture naming in comparison with subordinate or superordinate categories, and there is less between-participant variation concerning what attributes objects have, if they belong to basic categories (Rosch, Mervis, Gray, Johnson, & Boyles-Braem, 1976). Similarly, Mervis and Crisafi (1982) showed that basic categories are privileged in naming and other category-related behavior of children (see also Horton & Markman, 1980).

Basic level categorization can relate to unsupervised classification if basic level categories are viewed as *especially* coherent—i.e., whatever measure of coherence is assumed to underlie the formation of categories, is also assumed to distinguish basic from non-basic level categories. Indeed, there have been sophisticated attempts to provide computational accounts of basic level categorization performance (e.g., see Corter & Gluck, 1992; Gluck & Corter, 1985; Gosselin & Schyns, 1997, for some prominent models and reviews).

Broadly speaking, unsupervised classification models and basic level categorization ones have different predictive scopes. With basic level categorization there is an assumption that there is a hierarchy of concepts, so that the predictive objective is to identify the category level that would be privileged in this hierarchy; ordinarily, this involves identifying the basic level among three or four category levels, but no attempt is made to predict the exact way in which items should be partitioned in categories within the basic level. Conversely, in unsupervised classification, we aim to identify the particular preferred classification for a set of objects among all possible classifications for these objects. Also all basic level categorization models operate on some type of featural representations for the objects that are categorized. Unsupervised categorization concerns primarily spontaneous classification of objects that are novel and for which, by definition, we have no prior schemas (and by consequence it is very frequently either difficult or impossible to specify features). Thus, the simplicity model is computationally implemented independent of features; while we can use the simplicity model on objects specified in terms of features, we cannot apply models of basic level categorization with the datasets we investigate in this paper.

2. The simplicity model of unsupervised classification

Let us now consider how a simplicity principle can be applied to grouping items into categories (a formal presentation of the model is provided in the Appendix A). This requires specifying the data and hypotheses. Hypotheses correspond to possible groupings of the items. Let us assume that the data correspond to information about the similarity structure of the items. From standard information theory, the code length required to specify the similarity structure for a set of objects in terms of a particular grouping will be the sum

$$\begin{aligned}&\text{code length to specify similarities in terms of grouping}\\&\quad+\text{code length to specify grouping}\end{aligned}\tag{1}$$

Different groupings will result in a different overall codelength for (1). According to the simplicity principle (e.g., see Rissanen, 1978, for a particular mathematical formalization), groupings associated with a short codelength (high compression) will be favored. To formulate a simplicity model of unsupervised categorization we need to devise a scheme whereby specifying the similarity structure of a set of items with groups may result to a reduction in codelength.

2.1. Form of the data

Many different kinds of representation assumptions have been made in categorization research. In spatial models of representation (e.g., Nosofsky, 1985; Shepard, 1980, 1987), it is assumed that items can be embedded in a multidimensional space, and that similarities are negatively monotonically related to distances in this space. Such an approach requires that the similarity relations of the objects to be categorized adhere to the metric axioms, but there are situations where similarity information violates the metric axioms (e.g., Bowdle & Gentner, 1997; Tversky, 1977, see Nosofsky, 1991, for a very interesting discussion of these issues). A common alternative is to create representations of objects in terms of features. Items are taken to correspond to bundles of features, so that similarity is a function of the degree to which features are, or are not, shared between the items (Tversky, 1977). However, it is difficult to specify the features corresponding to novel, abstract objects (e.g., meaningless drawings). Unsupervised categorization very often concerns objects with which we have no prior experience, so that a model relying on features would be very restricted.

In the simplicity model what matters for accurate prediction of unsupervised categorization is how people perceive the similarity of the objects to be categorized. In some cases it would be more accurate to describe similarity information in terms of internal spaces, or even abstract similarity relations, in other cases features might be better. With this in mind, the present formulation of the simplicity model has been made in a way that is compatible with different types of representation assumptions. Thus, we avoid the immensely difficult, and largely irrelevant to this work, problem of psychological representation (see Goldstone, 1993; Goodman, 1972; Hampton, 1999; Quine, 1977). The approach of specifying some computational principle of cognition independent of specific representational assumptions has been pursued in other areas of cognitive science (e.g., Anderson, 1990; Marr, 1982; van der Helm & Leeuwenberg, 1996, 1999).

The simplicity model assumes that similarity information is in the following form: consider four objects *A*, *B*, *C*, and *D*; then, similarity information would be specified as similarity between (*A*, *B*) less or greater than similarity (*C*, *D*). The simplicity model was formulated in this work so that such similarities are never equal, and also that similarities between objects obey minimality, that is, similarity $(A, A) = 0$, and symmetry, that is, similarity $(A, B) =$ similarity (*B*, *A*); the similarities, however, may violate transitivity. The simplicity model can be specified assuming any combination of the metric axioms (if all the metric axioms can be assumed for some similarity relations, then these relations can be specified with less information; see Hines, Pothos, & Chater, submitted for publication). The computational implementation of the model is very similar whether symmetry and minimality are assumed or violated; since the stimuli we employed in this work were simple geometric shapes there would be no psychological reason to assume that symmetry or minimality are violated, (Tversky, 1977, discusses some of the factors that could underlie violations of the metric axioms; see also Bowdle & Gentner, 1997). However, it is difficult to modify the computer code of the version of the model that assumes transitivity to the code of the version in which transitivity could be violated. Since in the future we would like to examine datasets where transitivity could be violated, we programmed the simplicity model without the assumption of transitivity. It is important to note that unless similarity information is collected as specified above (is similarity between (*A*, *B*) less or greater than similarity (*C*, *D*)?), then transitivity will be necessarily obeyed. However, this does not affect the computation of the optimal categorization according to the simplicity model. The assumption of transitivity (and of all metric axioms) is equivalent to assuming extra constraints in the similarity information for a set of objects: for example, if we have objects *A*, *B*, *C*, and *D*, and we also know that $(A, B) > (B, C)$ and $(B, C) > (C, D)$, then assuming transitivity means that we can infer $(A, B) > (C, D)$. Consider now a case of comparing two classifications for the same set of objects, so that transitivity could be violated but in fact we know that it is obeyed (this is the case for the datasets examined here). How many extra constraints are there due to transitivity will depend on the number of groups and the number of elements in each group. This means that in situations where the classifications compared have similar groups and similar number of elements in each group, the extra constraints due to transitivity will be the same, so that the identification of the optimal classification will not be influenced by the inappropriate assumption of transitivity (this applies to all the "structured" datasets in this work in looking for the optimal classification). In summary, in the experiments to follow we analyze our data as if transitivity could be violated when, in fact, this is not the case. However, in no circumstances does this affect identification of the optimal classification.

Deciding whether for two pairs of objects *A*, *B*, and *C*, *D* similarity (*A*, *B*) is smaller or greater than similarity (*C*, *D*) is a binary choice, and so (from information theory) involves one bit of information. In the case where we have *r* items, or $s = r(r - 1)/2$ similarities between pairs of these *r* items, there are $s(s - 1)/2$ comparisons between the similarities for pairs of items. If the data are specified directly, without using a grouping, we require $s(s - 1)/2$ bits.

Suppose, for example, that the items are simply points in a Euclidean space as in Fig. 1. In this case, distances correspond to dissimilarities between the items (whether the similarity structure of the items is in terms of similarities or dissimilarities is irrelevant).

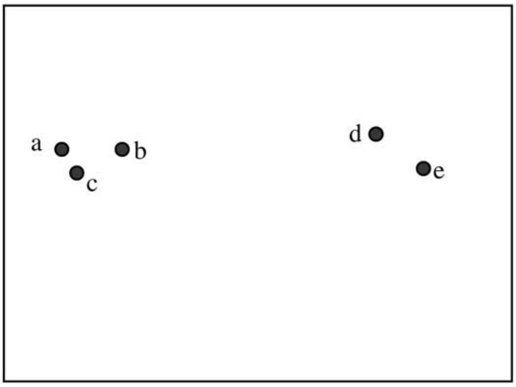

Fig. 1. A simple arrangement of points in a Euclidean space.

There are (5 × 4)/2 = 10 distances between these points, which can be expressed as (10 × 9)/2 = 45 inequalities; for example

$$d(a,c) < d(b,c);\ d(a,c) < d(a,b);\ d(a,c) < d(d,e);\ d(a,c) < d(b,d);$$
$$d(a,c) < d(b,e);\ d(a,c) < d(c,d);\ d(a,c) < d(a,d);\ d(a,c) < d(c,e);$$
$$d(a,c) < d(a,e)$$

and so forth. The redundancy of the data, that is the fact that there are very obvious regularities in the specification of these inequalities, means that there is a shorter description which captures the structure of the data—the simplicity model is one way of attempting to model this structure.[2]

2.2. Clustering by simplicity step 1: coding groups

The second term in (1) reflects how difficult it is to specify the category membership of a set of items. To compute the codelength required to specify how r items are allocated into a particular set of n categories, we must consider all possible allocations of the items into different classifications. The number of all possible allocations is given by $\sum_{v=0}^{n}(-1)^{v}((n-v)^{r}/(n-v)!v!)$ (this is Stirling's number, e.g., Graham, Knuth, & Patashnik, 1994; see also Feller, 1970). By standard information theory, the code length required to identify one out of D possibilities, assuming each one is equally probable, is $\log_2(D)$. Thus, a code of length $\log_2 \sum_{v=0}^{n}(-1)^{v}((n-v)^{r}/(n-v)!v!)$ will be needed to specify a particular allocation of r items into groups. In general, this term represents a minor contribution in the overall computation.

Note that in general certain category structures, for example, category structures that involve a number of clusters equal to the number of items divided by two, would be more likely than other ones (for example, a classification where all items are in one cluster, or one where each item is in its own cluster). In the above computations, our motivation was to choose a classification coding scheme which would be as neutral as possible with respect to the probability of different category structures and one that would appear to be consistent with the simplicity approach in general (for example, on the basis of the above scheme a classification where all objects are in one large category is considered very simple, albeit not very informative). It is possible that future work will identify constraints over the relative likelihood of different category structures that will require a particular, non-uniform, prior probability distribution for these category structures to be incorporated into the simplicity model. Finally, in principle, this is where general knowledge effects could be introduced since certain groupings will be psychologically more plausible, and hence will require shorter code lengths, than others (for example, biological vs. non-biological objects). With novel, abstract stimuli, however, we can ignore such effects and compute the second term in (1) in terms of the combinatorics of object allocation to categories (see also Pomerantz, 1981).

2.3. Clustering by simplicity step 2: specifying the data in terms of groups

We now turn to the question of how a group can be used to encode similarity data. We define a cluster or a category as a collection of objects such that all within-cluster similarities are greater than any between-cluster similarity (in other words, groupings are specified so that within category similarity is as great as possible, and between category similarity as low as possible; Rosch & Mervis, 1975). A particular grouping will therefore place default constraints on the similarities between items. The first term in (1) would be reduced for a grouping if these constraints are strong (i.e., many comparisons between distances are specified by them) and they are generally correct (i.e., relatively few of the constraints have to be "corrected" to reconstruct the data).

The above definition of a category allows us to formulate a way to specify similarity data in terms of a particular grouping. We first need to describe all the similarity inequalities that are *not* specified by the grouping (i.e., comparisons between two within-cluster similarities, or between two between-cluster similarities); if there are t of these, this will require a code of length t bits. Then we correct errors in the constraints imposed by the grouping—i.e., cases where a within-groups similarity is less than a between-groups one.

Suppose that the groups specify u constraints, of which e are incorrect. We first encode e, which must be between 0 and u, and hence requires a binary code of at most length $\log_2(u+1)$ bits. Then we must specify which e constraints out of the u constraints specified by the clusters are incorrect. By standard combinatorics, there are ${}_uC_e = (u!/e!(u-e)!)$ ways to choose e items from a set of u. Thus, the total code for correcting erroneous constraints, E, is $\log_2(u+1) + \log_2({}_uC_e)$.

Note that a short codelength is needed to specify the errors either when there are very few errors or when there are very many, and the greatest codelength is needed when the errors are half the number of constraints. We stipulate that the number of errors is less than half the entire number of constraints; if this is not the case, no clustering is defined (since otherwise such a clustering will group dissimilar items, not similar ones). This seems to be a mild additional assumption, from a computational point of view, because any reasonable algorithm for finding clusters will tend to group similar items (and this tendency, even if slight, will rule out the 'pathological' cases of classifications that involve several errors). Psychologically this corresponds to the plausible assumption that learners start classifying a set of objects by first creating a few categories that involve no errors, rather than by creating many categories that involve several erroneous constraints.

2.4. Summary of the simplicity model of categorization

The similarity structure of a set of objects can be represented in terms of pairwise similarity inequalities between pairs of objects. Using categories is equivalent to reducing the number of inequalities that need to be specified, since all within category similarities are defined (in this model) to be greater than between category similarities. However, this advantage of using categories needs to be balanced against the codelengths required to describe the particular set of categories used, and also correct any errors in the inequalities determined by the categories (the constraints). In general, using categories may simplify the description of the similarity structure of a set of objects. We suggest that the greater the simplification (or compression), the more psychologically intuitive the category structure should appear. To evaluate this proposal we use the simplicity principle to identify the psychologically preferred classification for different sets of items, and examine whether people's performance is compatible with the predictions we make.

The presentation of the simplicity model enables us to identify the defining features of the simplicity model. It enables predicting the most appropriate classification for a set of items without any information either about the number of categories sought, or the distributional properties of the items. The second feature, non-parametricity, is what makes the simplicity model particularly suitable for modeling classification of novel objects, for which there are no prior expectations. Note that the specification of the model is such that it is non-metric as well, that is the model works with only the ordinal relationships of the pairwise similarities. The fact that the model is non-metric implies a certain degree of scale invariance: if all similarities are multiplied by the same constant then the classification predictions of the simplicity model do not change; all that matters is relative magnitude of similarities. This is to be contrasted with the much more common parametric (e.g., Bayesian) approaches to unsupervised categorization where changing the magnitudes of similarities will generally affect classification predictions. Scale invariance has been supported as a valid assumption in modeling work in perceptual grouping (Compton & Logan, 1999; van Oeffelen & Vos, 1982, 1983), which is a process very similar to unsupervised categorization. For the purposes of the present investigation, with simple datasets of novel objects, scale invariance seems an appropriate assumption within reasonable bounds (for example, if for a set of objects naïve observers perceive a particular classification and then the relative similarities were increased by so much so that it would be hard to distinguish some objects from each other, then we would expect the perceived classification to change as well). More generally, however, it is a very interesting issue to examine the psychological accuracy of scale invariance in unsupervised categorization, and the simplicity model provides a means to help carry out such an investigation. Most likely, there are situations where scale invariance matters and situations where it does not; stimulus novelty certainly appears as a condition necessary (but by no means sufficient) for scale invariance of the perceived relative similarities and the corresponding spontaneous classifications.

3. Experimental investigation—general issues

Four experiments were conducted to test the simplicity principle of categorization. In all cases participants were asked to group together sets of meaningless items, to

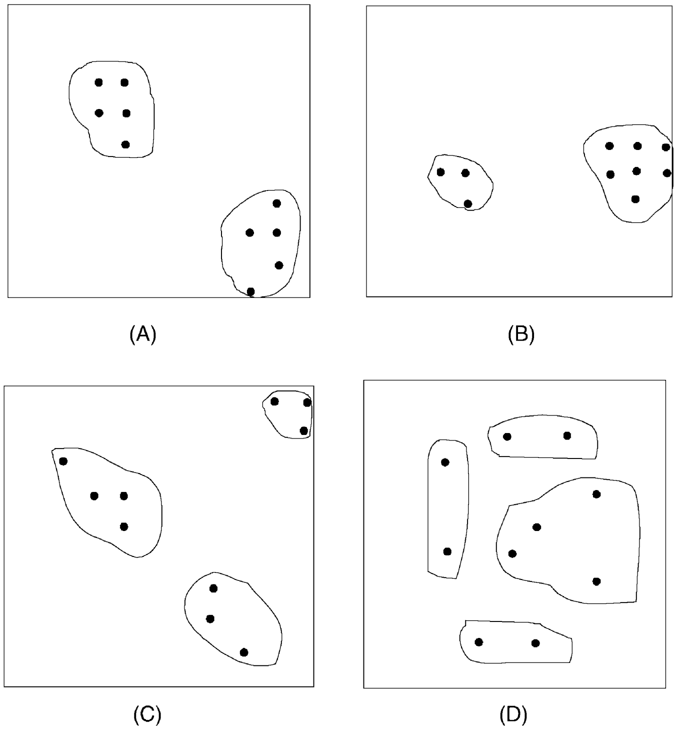

Fig. 2. Four datasets and the optimal classifications, as predicted by the simplicity principle. Dataset A will be referred to as the *two clusters* dataset, B as *big, small cluster*, C as *three clusters*, and D as the *little structure* dataset.

assess whether these groupings correspond to the ones favored by the simplicity principle.

What is the predictive objective for a model of unsupervised classification? Consider the datasets in Fig. 2 that we will properly present later: to a naïve observer, some of these classifications appear more intuitive or obvious than others. For example, in the two clusters case, the two categories are so obvious that they seem to emerge before even any information becomes accessible about the distribution of the dots within each category. On the other hand, this is less the case for the *three clusters* dataset, and a lot less so for the *little structure* one. In general, we are trying to model exactly these intuitions: what is it about the *two clusters* or the *big, small cluster* datasets that makes us recognize some particular groupings instantly, that does not apply to the *little cluster* dataset? As we will see later, such differences in the intuitiveness of classifications can be captured by differences in compression.

For a 10 item dataset there are nearly 100,000 alternative classifications (e.g., see Medin & Ross, 1997; Heit & Bott, 1999). The seemingly effortless psychological judgment of separating the items in the *two clusters* dataset into two categories reflects a choice of one classification among 100,000 possibilities, and a successful model of unsupervised classification needs to predict this choice. We now discuss ways in which the predictions of an unsupervised model of classification could be statistically examined. Note first that all the classifications possible with a dataset can be ordered in terms of how much compression they provide (higher compression is the same as greater simplification). The best possible compression associated with a classification for each dataset will vary; for example, as we will see later, the best possible compression for the two clusters dataset is higher than the best possible one for the *three clusters* one or the *little structure* one. However, the *average* compression associated with all classifications for any dataset will be the same, since the high compression classifications make up only a very tiny proportion of all possible classifications (so that they do not affect the average; explicit computations were conducted). Thus, a test of the simplicity principle would be to have a group of participants classify the items in different datasets and compute for each dataset the average compression associated with all the ways participants classified the dataset. There would be a simplicity bias if for each dataset the average compression of participants' classifications correlates with the best possible compression (as this would imply that participants are preferring the very rare high compression classifications). No correlation would imply that participants are classifying in accord with some other principle for grouping (or randomly). In cases where there is a sufficiently large number of participants classifying a dataset, we expect that the classification favored by simplicity would be produced more frequently than alternative classifications. Finally, for datasets for which a classification of very high compression is possible there should be lower variability in the way participants group the data. This is because high compression classifications are predicted to be more obvious, so that more participants would readily select them (rather than some alternative high compression classification). For example, considering again the Fig. 2 dataset, we would predict that participants would be a lot more consistent in classifying the *two clusters* dataset than they would be with the *little structure* one.

In terms of sources of error in examining the simplicity model, the most important limitation is that the simplicity model operates at the level of internal representations: that is, if we had an accurate measure of how an individual perceives the similarity structure of a set of objects, then the hypothesis is that the simplicity model could be used to predict how this individual would classify this set of objects, modulo constraints on cognitive optimization detailed next in this section. However, we do not have such a measure, and so deviations between predictions and performance could well reflect our inability to accurately infer how people perceive the similarity of a set of objects. This problem is ubiquitous in the study of categorization (e.g., Nosofsky, 1992; Tversky, 1977). However, it is possible that it will be more pronounced in the study of unsupervised categorization, given that there are so many possible classifications even for small datasets. If participants represent the similarity structure of the objects we use in a way different from what we assumed, then it is unlikely the simplicity model predictions will be supported.

Even if simplicity is an accurate cognitive constraint it is possible that the cognitive system cannot always compute the simplest solution. In this work, we do not investigate how compression is optimized, or any other dynamic aspect of the unsupervised classification process (Anderson, 1991). Thus, a person may not always classify a sets of objects with the classification providing the highest compression, in the sense that the cognitive system does not (or cannot) always perform optimally (cf. Osherson, 1990; Oaksford & Chater, 1991). In other words the only claim is that people would spontaneously classify groups of items in a way that reflects a simplicity bias, even if the optimal classification is not always produced.

4. Experiments 1–3: the items to be clustered

The simplicity principle was used to identify the preferred categorization in four simple datasets; the predictions of these models were evaluated in three categorization experiments (in one more experiment we used a different set of stimuli).

The datasets are shown in Fig. 2. Information about the relative similarity of objects can easily be derived from the object coordinates (in terms of relative distances). To identify the classification favored by simplicity we employed an algorithm which locally optimizes cluster goodness, and is akin to agglomerative clustering methods. It is important to note that the particular simplicity optimization algorithm is not meant to have any psychological significance whatsoever. The algorithm works by initially assigning each item to a separate cluster and subsequently successively merging pairs of clusters. At each step, all possible cluster pairs are formed and the total codelength that is needed to describe the data is calculated, assuming that a pair is merged and the other clusters are left intact. The cluster pair associated with the greatest reduction in code length is then merged, and the process is repeated. Unlike agglomerative clustering methods that continue merging clusters to provide a complete hierarchical classification tree, the algorithm identifies the level when no more compression occurs and then stops. There are many choices in the way one can select which clusters to merge at each step. In exploratory analyses we carried out the best choice appears to be to merge the pair of clusters which minimizes "costs"/"1 + constraints," where "costs" is the codelength required for encoding the classification and the errors, and "constraints" is how many inequalities are specified by the classification ("ratio" method). Note that we use "constraints + 1" rather than "constraints" in the "ratio" method because in degenerate cases a clustering may not specify any constraints at all. For the datasets shown in Fig. 2 informal examination of other algorithms did not lead to better solutions.

The classifications favored by simplicity that we identified are shown in Fig. 2, and in Table 1 the corresponding compression/codelength values are reported. Compression refers to the reduction in codelength achieved by using a classification to describe the similarity structure between a set of items. Thus, the greater the compression, the less the final codelength for a particular dataset; smaller codelengths should correspond to more obvious groupings, according to the simplicity principle.

Table 1
Best compression achieved with each of the datasets in Fig. 2

Dataset	Compression	Codelength
Two clusters	491	499
Big, small cluster	495	495
Three clusters	383	607
Little structure	181	809

Note: "Compression" refers to the maximum possible reduction in codelength achieved by using a classification to describe the similarity structure between a set of items, for each of the datasets. Thus, the greater the compression, the less would be the final "codelength" for a particular dataset; smaller codelengths should correspond to more obvious groupings, according to the present model.

As an example of how the numbers in Table 1 are computed, consider the *two clusters* dataset. In this, as well as in all the other datasets, there are 10 points, so that there are (10 × 9)/2 or 45 distances between the points. Distances are assumed to represent dissimilarities, so if we were to describe the similarity structure of these points without categories, we would be needing $(45 \times 44)/2 = 990$ inequalities, corresponding to 990 bits. According to the simplicity model, categories can reduce this codelength by specifying some of these inequalities. In particular, for the *two clusters* dataset, the categorization shown in Fig. 2 involves two groups of five points each. A category is defined so that all similarities within a category are greater than all similarities between categories. In other words, all distances within a category are meant to be less than all the distances between categories. In each of the two categories we have five points, or $(5 \times 4)/2 = 10$ distances. So, overall, there are 20 distances "inside" categories. There are 25 distances across categories (that is, distances between points belonging to different groups), so that this two group category structure specifies $(20 \times 25) = 500$ inequalities (constraints). However, if we are to use categories we need to consider the codelength associated with specifying the particular set of categories we use (in this case about nine bits), and also the codelength required to correct any errors in the category constraints (in this case there are no errors, so this term is zero; this is a very good classification). Overall, when we describe the similarity structure of the points in the *two clusters* case with the two groups as shown in Fig. 2, we have a compression of $500 - 9 = 491$ bits, or a codelength of 499 bits; recall that greater compression and smaller codelengths correspond to psychologically more intuitive category structures.

To briefly summarize the predictions of the simplicity model with the Fig. 2 results, in the first three cases the data points are partitioned in a way that seems to reflect most faithfully the structure of the domains. In the fourth case there is very little structure, and this intuition is confirmed in our account since the best categorization in this dataset is associated with a compression value that is much lower compared to what we obtained with the other datasets (and, thus, the resulting description of the dataset in terms of categories will require a longer codelength relative to the other cases; see Table 1).

5. Predictions of other models

Categorization is an umbrella term for several different kinds of processes, addressing different objectives. The diversity of the processes that make up categorization implies a diversity in the models that are proposed to explain these processes. In proposing a new model it is therefore important to identify its scope for meaningful comparisons with other models. We have already discussed at a broad level the relation of the simplicity model with other research in categorization (e.g., supervised categorization, basic level categories, category coherence). In this section, we discuss other computational models within unsupervised categorization, statistical clustering, and unsupervised learning more generally, in light of the Fig. 2 predictions.

Fried and Holyoak's (1984) is an influential early study on unsupervised learning that will help us illustrate some of the features of the simplicity model that make it distinctive. In Fried and Holyoak's (1984) approach each object is represented in terms of a number of independent feature dimensions and categories are described as density functions for these objects over a feature space. When a learner is presented with a set of exemplars, he/she can use these exemplars as a sample with which the actual category density function can be inferred; thus, learners can generalize from the presented sample of exemplars to potentially an infinite number of instances. For this approach to work, one needs to assume that category density functions have a particular form (Fried and Holyoak illustrate their theory with normally distributed category populations); also, the number of categories sought must be specified externally. Fried and Holyoak's (1984) research is an excellent model to contrast with the simplicity approach. In Fried and Holyoak's model, category learners know *a priori* that they are looking for a given number of categories, and that the category exemplars conform to certain distributional properties. This learning scenario corresponds to situations where learners have some *a priori* expectations about the exemplars they encounter; for example, this would be the case when we try to infer new categories in a domain we are already familiar with (e.g., a birds expert identifying bird categories in a new domain). By contrast, the simplicity model makes no assumptions at all either about the number of categories sought or the parametric properties of these categories, so that it addresses an aspect of unsupervised learning different to that of Fried and Holyoak (1984).

A similar, more recent proposal, for unsupervised classification from the machine learning literature has been that of Cheeseman and Stutz (1995). Cheeseman and Stutz's model, called AutoClass, involves two components. The first one is a probability distribution that determines which category each object belongs to (objects are not assigned to a category, rather they belong to different categories with different probabilities). For each category, the second component is a probability density function for the distribution of the attributes of the objects that belong to the category. Thus, unlike Fried and Holyoak's model, in AutoClass category attributes can be distributed in several different ways. Clearly, as AutoClass is not restricted to only one type of probability distribution, it can model many more types of attribute distributions within categories and category distributions. But the particular modeling scope in any AutoClass version is determined by the range of probability density functions AutoClass can select from. These features of AutoClass make it more similar to Fried and Holyoak's (1984) model, but distinct from the simplicity one, since in the simplicity model a particular distribution for categories or category attributes are not assumed.

As with AutoClass, there are several other Bayesian approaches to unsupervised learning that are not restricted by lack of knowledge for the number of categories sought. Although slightly outside the scope of the present paper, Ghahramani and Beal's (2000) research is still relevant in this context, as it shows how the basic factor analysis procedure can be extended in a Bayesian framework to automatically determine the number of factors required to model the data (factor analysis is somewhat similar to clustering if one considers the number of instances associated with each factor as belonging to the same cluster). This is achieved by a method that allows components to be rejected from the model, under particular circumstances. Ghahramani and Beal's (2000) computations though assume that the distribution of information to be modeled is Gaussian in form, in contrast to the non-parametric method used in the present approach.

Another way in which parametric features can be used to guide a model's search for how to classify a set of objects is van Oeffelen and Vos's (1982, 1983) CODE model, later modified by Compton and Logan (1993, 1999). The model has been presented as a formalization of the proximity principle of the Gestalt approach to perception, with respect to the problem of perceptual grouping of dot patterns (but presumably the model can be readily extended to grouping of similar, primitive assemblies of elements). In CODE each element in a pattern is seen to have a "strength" that originates from the element and propagates in an isotropic manner away from it. At its location in the pattern the strengths from the different elements are added and if the result is above a certain threshold then the elements are considered as belonging to the same group. Clearly, the function determining strength spread is of crucial importance to determining the classifications predicted as accurate (as is the case with the AutoClass model). In the original CODE formulation, the threshold was a fixed parameter so that the model predicted a single classification for a set of objects. But Compton and Logan (1993, 1999) modified CODE so that it produces a set of nested classifications for a set of objects. Compton and Logan justified their modification on the basis of general considerations relating to the psychological features of the perceptual grouping process. This is an additional illustration of how the appropriateness of model features such as parametricity or fixing the number of categories sought depends entirely on which cognitive process is being modeled.

Other well known models of unsupervised categorization are Ahn and Medin (1992) and Schyns (1991). Ahn and Medin's (1992) two-state model of category construction is a model for free classification. The model was primarily used to evaluate the relative compellingness of a hypothesis according to which spontaneous grouping is driven by an overall family resemblance process, relative to an alternative suggesting that people would tend to sort stimuli on the basis of a single dimension (which is an issue that has received considerable attention in the free classification literature). But, as is common in the free classification literature, there is no predictive interest in determining the number of categories. In the Ahn and Medin (1992) most sorts were obtained for two groups; more generally, the model's prediction is that there will be as many groups as the number of (feature) values along the most salient dimension of the objects (Ahn, personal communication; no attempt is made to predict the most salient dimension).

Schyns (1991) used a neural network made of two modules to study the problem of how children spontaneously discover categories in the instances in their experience and subsequently associate these categories with the category labels provided by their family context. The category learning aspect of Schyns' work is relevant to us here: the way categories were spontaneously discovered was by using a Kohonen neural network architecture to reduce the high dimensionality input vectors to lower dimensionality (two dimensions) ones at the output. A Kohonen neural network enables the segregation of the output space into distinct regions, that can effectively be identified with different categories. This segregation into distinct regions is spontaneous in the sense that it is not guided by any external constraints, but rather by the similarity structure of the input distances. Schyns' work, like Ahn and Medin's, is an excellent example of models of unsupervised categorization where the categories can be spontaneously identified given information about how many categories are sought: for a Kohonen neural network to function, the number of categories sought needs to be known in advance. As Schyns (1991) notes, the operation of a Kohonen net is very similar to that of a *K*-means clustering algorithm (discussed below).

Anderson's (1991) rational model of human categorization is worth mentioning here so as to further illustrate the different modeling requirements that arise in the study of categorization. Given a set of objects presented sequentially, the model is capable of identifying a categorization of the objects that optimizes the accuracy of estimating unseen features of the items in the categories. The focus of Anderson's model is on how differences in presentation order of the to-be-categorized stimuli affect the final classification produced; this is in contrast to the present model which assumes that all stimuli are processed simultaneously (Handel & Preusser, 1969). In Anderson's model a new cluster is formed if a newly encountered item is sufficiently dissimilar from the items in the existing clusters; and hence the number of clusters produced depends on a free parameter, the coupling parameter, controlling the level of dissimilarity that triggers the construction of a new cluster. Thus, for Anderson the coupling parameter is an essential aspect of his modeling problem: since we do not know how many objects are there to be classified, there has to be an *a priori* measure of "closeness" of instances for the purposes of determining whether objects should be classified together or not. By contrast, for the simplicity model predicting the number of categories is a central aspect of the modeling problem.

Anderson's (1991) rational model is different from the simplicity one in that it is a model for the dynamic aspects of unsupervised categorization. Analogous is the difference between the simplicity model and Fisher's COBWEB system (Fisher, 1987, 1996; Fisher & Langley, 1990; Gennari, Langley, & Fisher, 1989; Gennari, 1991). This approach uses Corter and Gluck's (1992) category utility, which has been proposed as a measure to explain what is special about basic level categories. Category utility (and by extension COBWEB) can, in theory, be used to predict both the number of clusters appropriate for a set of objects, and how items should be divided among these clusters. However, category utility is a principle for basic level categorization; the relation between basic level categorization and the aspect of unsupervised categorization the simplicity model addresses is currently not clear and a comparison between category utility and the simplicity model should follow a clarification at the theoretical level of basic level categorization and unsupervised categorization more generally.[3]

Unsupervised clustering has been intensely studied in the context of statistics and data mining as well (e.g., Arabie, Hubert, & de Soete, 1996; Fisher, Pazzani, & Langley, 1991; Everitt, 1993; Hartigan, 1975; Krzanowski & Marriott, 1995). There is a large and important line of research concerned with hierarchical agglomerative cluster analysis (e.g., Jardine & Sibson, 1971): at the first step of such an analysis, all items are assumed to be individual clusters. At the next step, two items are combined in a single cluster, and so forth, until the combination of items/clusters leads to an all-inclusive category. This procedure will always result in $n - 1$ groups, for n items, regardless of the actual algorithm used. A second general approach to clustering, K-means clustering, involves optimizing an explicit criterion for grouping items into K categories (K is determined by the investigator; Banfield & Bassill, 1977; Duda & Hart, 1973; MacQueen, 1967). This criterion (usually called an objective function), can be viewed as a measure of category cohesiveness, and different methods are typically evaluated in terms of whether it is appropriate or not. Given a set of items, the output of such methods is a discrete (non-hierarchical) set of groups that best reflect whatever criterion has been employed.

Michalski and Stepp's (1983) CLUSTER/2 is a statistical clustering model that is relevant here primarily because one of the determinants of classification goodness is the simplicity of verbal description of the categories created (see also Ahn & Medin, 1992; Medin, Wattenmaker, & Michalski, 1987b). Additional criteria of classification goodness include how well the clusters fit observations, the disjointness of clusters, a "disjointness index," a "discrimination index," etc. As Michalski and Stepp state (p. 399) "The selection of elementary criteria, their ordering, and the specification of tolerances is made by a data analyst." This flexibility may be appropriate for statistical clustering, where a single model may have to deal with several different kinds of datasets (Fraboni & Cooper, 1989), but less so for cognitive modeling, where researchers need to be able to evaluate the plausibility of a model in terms of its free parameters relative to the degrees of freedom in the data.

We have briefly discussed only a small sample of the multitude of models in unsupervised categorization and learning, with a view to clarify the particular aspect of unsupervised categorization the simplicity model has been created to address: there are situations where people would spontaneously determine a preferred grouping for a set of items with no information either about the number of categories sought or the distributional properties of the objects categorized. Thus, internal to the operation of the simplicity model is both the computation of how a set of objects should be divided into categories and how many categories should be used. Most of the models of unsupervised categorization (and statistical clustering) we reviewed above could probably replicate the Fig. 2 predictions, but only with information about the number of groups to be used in each case (it is important to note that the fact that these models require information about the number of the groups is not a shortcoming but simply reflects modeling requirements different to the ones we address with the simplicity model). Also, there are many parametric models (particularly in machine learning and statistics) that could also replicate the Fig. 2 predictions without any information about the number of clusters sought. Specifically, the fact that all the clusters in Fig. 2 have a (broadly speaking) Gaussian form implies that Bayesian approaches (such as AutoClass and Ghahramani and Beal's approach) would be expected to discover the same clusters as the simplicity model. Thus, the experiments we provide with the Fig. 2 datasets do not specifically provide support for the simplicity model as opposed to these Bayesian models; future empirical work will be required before a conclusion is forthcoming with respect to when human classification is guided by parametric assumptions and when it is better described by non-parametric models (such as the simplicity model). Rather, our objective in this paper is to show that a non-parametric classification model can indeed be used to accurately describe human classification performance and that such a model can be derived on the basis of specific assumptions about psychological processes (namely the simplicity principle).

6. Experiment 1

Classification variability may be a problem in testing the simplicity model. In this experiment we used two ways to reduce variability. First, we used some datasets that are associated with very high compression (the *two clusters*, the *big, small cluster*, and, to a lesser extent, the *three clusters* datasets in Fig. 2). The simplicity model predicts that the classification structure should be very obvious, so that participants should be more consistent in their classification (the *little structure* dataset was used to contrast this prediction). Second, we presented the datasets in a format identical to that in Fig. 2: each dataset was printed on an A4 sheet of paper as a set of dots and shown to participants individually, an approach that has been adopted before in studies of unsupervised grouping (e.g., Compton & Logan, 1993, 1999). Participants were asked to illustrate their grouping preferences by drawing curves round groups of dots (see below). In this way classification variability would be reduced, as many classifications would become very unlikely (e.g., classifications that would involve complex curves).

6.1. Method

Ten paid participants, all members of the University of Oxford, were tested individually. Each participant was presented with A4 sheets on which the four datasets in Fig. 2 were printed separately. Participants were simply asked to divide the points in a way that seemed "natural and intuitive," by drawing a curve with a pencil round the points that the participant thought should be grouped together. Participants were told that they were allowed to make changes. The order in which the four datasets were presented was randomized for each participant.

6.2. Results and discussion

In this experiment we test the hypothesis that the classification favored by simplicity would be produced more frequently than alternative classifications. To achieve this, we need some measure of how frequently one would expect any classification to be produced by chance. Clearly, one cannot compute this frequency on the basis of all possible classifications for 10 objects: with 100,000 possibilities any observed frequency of one would be readily judged as "more frequent than chance." Thus, in order to determine whether some of the observed categorizations were more likely than others in each dataset, we adopted the following procedure. For each dataset we identified all the distinct categorizations produced by participants (referred to as the "distinct" solutions for a dataset in this and later discussions), as well as the number of times participants divided the dots in the way predicted by the simplicity principle (see Table 2). If there had been no preference for any particular categorization, we assumed that all distinct solutions would have been produced with a roughly equal frequency, given by the ratio (total number of groupings)/(number of distinct groupings). For example, if 10 participants produced overall five distinct classifications in one dataset, then the chance frequency of any one grouping for this dataset would be taken to be two.

Table 2
Performance with each of the datasets in Experiment 1

Dataset	Distinct solutions	Simplicity best
Two clusters	4	6
Big, small cluster	6	4
Three clusters	6	4
Little structure	10	1

Note: "Distinct solutions" refers to how many distinct classifications were produced for each of the datasets in Fig. 2; "simplicity best" denotes the number of times participants came up with the optimal classification according to the simplicity model. There were 10 participants.

A chi-square test was then used to examine the deviation of the frequency of any one solution from what would be expected by chance. For the *two clusters*, *big, small cluster*, and *three clusters* datasets the only groupings that were significantly more frequent than chance were those predicted by the simplicity model (for the three datasets, respectively: $\chi^2(1) = 6.53, 3.92, 3.92$; in all cases $p < .05$), while no other categorization was produced with a frequency that deviated from chance. For the little structure dataset the frequencies of all solutions observed were the same, suggesting that when the total compression possible with a set of items is too small no classification emerges as "obvious" or preferred. These results are consistent with the predictions of the simplicity model. The format of presentation of the datasets makes inappropriate other tests of the simplicity principle.

In this experiment the objects in the datasets could only be perceived in relation to the other objects (the position of a point in space is meaningful only in relation to the positions of the other points), and this made groupings involving spatially distinct points very unlikely. Moreover, it is possible that the results in this experiment reflect some aspects of spatial grouping phenomena, and would not be observed with stimuli not represented in spatial form. We proceed in the next experiment with a completely unconstrained grouping procedure, where the objects are presented as distinct, individual entities. This procedure would aim to mirror as closely as possible natural unsupervised classification situations.

7. Experiment 2

We repeated Experiment 1, using the same underlying category structure, but where the two dimensions associated with each stimulus were expressed in a different format: as parameters determining the lengths of parts of geometric objects, rather than as coordinates of points in space. In comparing human performance with the simplicity principle predictions, we assume that the physical space dimensions of the stimuli map onto some internal representation of these items. This is a standard assumption in supervised categorization research with simple, schematic stimuli.

7.1. Materials

The coordinates of the points in the Experiment 1 were used directly to construct simple star-shaped stimuli. These stars varied along two dimensions: their inner diameters corresponded to the vertical dimension of the datasets in the previous section, while their outer diameters corresponded to the horizontal dimension. Thus, each data point specified an inner and outer star and these were "blended" together, to give the impression of an individual object, as shown in Fig. 3a (Handel & Imai, 1972); successive levels of star size were readily discriminable, and inner diameters varied between 10 and 100 mm, while outer diameters varied between 108 and 198 mm. The step sizes in both dimensions were the same, so as to encourage participants to treat the dimensions as roughly equally salient (Handel, 1967; Handel & Preusser, 1969).

Stimuli were printed onto separate sheets of A4 paper in black ink and presented to participants in folders. To illustrate the relative variation between the stimuli, in Fig. 3b we present

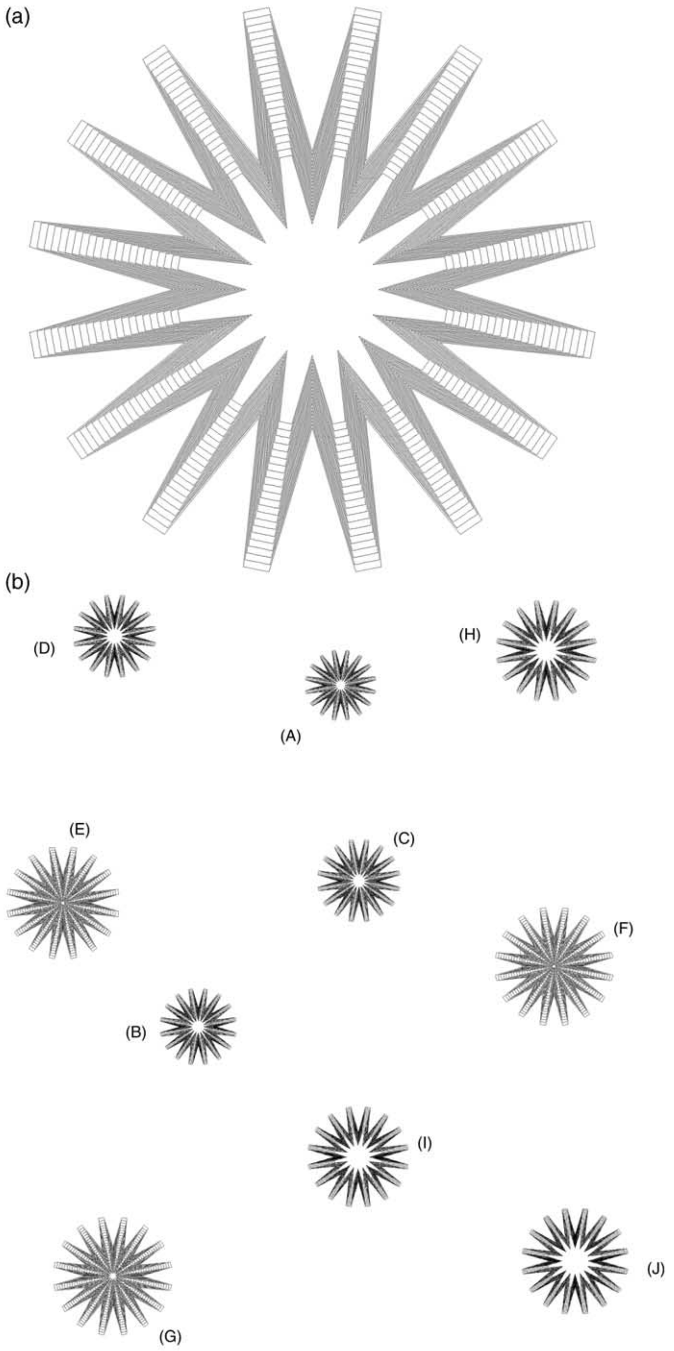

Fig. 3. (a) An example of the stimuli constructed from the datasets presented above. The inner and outer stars were "blended" together to give the impression of an individual entity. (b) The stars for the dataset *three clusters*. The classification predicted as optimal by the simplicity principle involves three clusters, that correspond to items $\{A, B, C, D\}$, $\{E, F, G\}$ and, finally $\{H, I, J\}$. In the experiments the stars were presented individually on A4 sheets of paper.

the whole stimulus set for the *three clusters* dataset. In practice, as the stimuli were individually placed in folders, participants found the task of sorting them into groups rather involving. No participant had trouble discriminating between the different stars.

7.2. Procedure

Twenty eight University of Oxford students, mostly Psychology undergraduates, were tested individually, receiving a small payment for taking part in the study; no participant had taken part in Experiment 1 and they were all experimentally naïve. Four sets of stars, each set corresponding to one of the datasets in Fig. 2, were presented to participants (i.e., each participant classified all four datasets; the order in which sets were presented was randomized between-participants). Participants were asked to partition the stars in a way that seemed “natural and intuitive.” They were also told that there was no limit to how many groups they could use, but that they should not use more than what they would think is necessary. Participants could see the stars in any order they liked, compare them, and also change their minds about which categorization was more suitable (see Handel & Preusser, 1969, for differences in free classification depending on whether the stimuli are presented simultaneously or sequentially).

7.3. Results and discussion

The simplicity model predictions were derived in terms of item coordinates in the Fig. 2 datasets, as in Experiment 1.

We first tested whether there is a preference for the simplicity predicted classification, against the null hypothesis that all the classifications produced were equally likely. Single sample chi-square tests were performed to investigate which solutions were produced more frequently than chance in each case. For the *two clusters* and *big, small cluster* datasets, the only categorizations that were significantly more frequent than chance were the ones associated with least codelength (for the two datasets, respectively: $\chi^2(1) = 28.3, 127.1$; in all cases $p < .001$). In the *three clusters* case, the best compression solution, which involved three clusters, just failed to reach significance ($\chi^2(1) = 3.32$, $.05 < p < .1$), while a solution where two of these clusters were merged into one (the four point cluster and the three point one below it in Fig. 2), was the only one that was selected significantly more often than chance ($\chi^2(1) = 27.5$, $p < .001$). There was no preference for any particular categorization for the fourth dataset, replicating the previous result that participants could not identify any classification structure with this dataset; this finding is consistent with the simplicity model in that the *little structure* dataset is associated with very low compression. The frequencies of the classifications predicted by simplicity for each dataset are shown in Table 3.

The predictions of the simplicity model for the *three clusters* dataset could not be supported on the basis of the chi-square analyses above. In light of our failure to find support for simplicity in the *three clusters* dataset we next present a more detailed analysis for this case. The Rand index is a way to assess the similarity between alternative cluster configurations on the same dataset (Rand, 1971). Two clusterings of a set of items are compared by considering for each pair of items whether the items both fall in the same or different groups in the clusterings.[4] Using the Rand index we can investigate whether the classification predicted by simplicity was

Table 3
Performance with each of the datasets in Experiment 2

Dataset	Distinct	Robust distinct	Simplicity best
Two clusters	18	3	8
Big, small cluster	12	3	15
Three clusters	12	5	5
Little structure	18	7	0

Note: In addition to the information provided in Table 2, for Experiment 2 we also calculated the "robust distinct" solutions, that is the number of distinct classifications produced with a frequency greater than one. There were 28 participants.

still more characteristic of overall performance, compared to the one produced most frequently. The Rand similarities of all distinct solutions produced to the best compression solution and the most popular one were compared. A significant paired-samples, two-tailed *t*-test confirmed that the average similarity of all distinct solutions produced to the best compression one (.86), was greater than the similarity of all distinct solutions to the most popular one (.70); $t(11) = 2.708$; $p = .02$. Thus, participants' solutions appear to be reflecting a bias to the categorization favored by the simplicity principle even where this solution is not the most frequently chosen. In this work, we will not further consider the Rand index analysis, but rather prefer the more straightforward chi-square analysis; the relevance of this procedure here is seen primarily as a way to acquire some better insight into the results with the *three clusters* dataset that appear problematic for the simplicity model.

As above, if the cognitive system is trying to optimize simplicity, then we would expect that the average compression of the classifications produced with each dataset would be as close as possible to the best possible compression. If participants were not classifying according to simplicity, then the average compression of their groupings would not bear a relation to the best possible in each case. In fact, the mean compressions of all classifications for each dataset were nearly identical (means computed from all approximately 100,000 possible classifications for each dataset). Supporting the simplicity model, we did find a very high, significant Pearson correlation (.951, $p = .049$, two-tailed) between the best possible compression and the average compression from the classifications participants came up with, for each dataset.

Table 3 shows classification variability for each dataset. It is interesting to note that the number of distinct solutions that were observed with frequencies greater than one is consistent with the simplicity model expectation that greater compression should be associated with less classification variability. However, the overall pattern of classification variability suggests that greater sample sizes would be needed before this hypothesis is properly examined.

In summary, the general pattern of results mirrors that found in Experiment 1, but using stimuli which can be perceived individually, and for which an explanation in terms of low-level perceptual grouping processes does not apply. The correlation of average compression from all classifications with a dataset and best possible classification for the dataset was unambiguously supported and provides the clearest and most valid indication of the psychological plausibility of the simplicity model in this experiment.

8. Experiment 3

In Experiments 1 and 2, participants have been given an open-ended categorization task. The realistic context we use in Experiment 3 contributes the following in a study of unsupervised categorization. First, it complements the experiments where the classification task was completely open-ended: psychologically, there are situations where we recognize groupings spontaneously, but also situations where we classify objects in groups as part of a specific task. Both situations reflect what we understand at this stage to be within the scope of the simplicity model. Second, at a practical level, the open-ended instructions of Experiments 1 and 2 may have made it more difficult for participants to identify the grouping problem with something they might be likely to engage in, in their everyday life. Indeed, in reasoning research there has been considerable discussion concerning the degree to which setting a problem in a practical, rather than abstract, context provides a more accurate means to investigate psychological performance (e.g., Cheng & Holyoak, 1985; Evans, Newstead, & Byrne 1991; Rumelhart, 1980).

8.1. Materials and procedure

Twenty eight University of Oxford students took part in the study for a small payment. They were all tested individually, and no participant was aware of the hypotheses tested, or had previous experience with Experiments 1 or 2.

The materials we use in this experiment are identical to those in Experiment 2, but the grouping problem is presented in a realistic situation. Thus, we told participants that the items to be grouped (the “stars,” as in Experiment 2), corresponded to drawings of real stars to be manufactured and shipped to customers. Stars are shipped in boxes with a central pole, which secures the stars in position. If the stars differ too much on their inner diameter, the pole will hold some of the stars only loosely, and they may move and break; but small boxes should be preferred for ease of handling, so that large variations in outer star diameter are also inappropriate. Participants also had to keep the number of different types of boxes small, to reduce manufacturing costs. On the whole, the different grouping requirements were made to broadly reflect categories that maximize within group similarity, while minimizing between group similarity (but the trade-offs between different types of constraint were obviously unspecified). Finally, to illustrate what the problem involved there was a paper model of a star and a box with a central pole, which we used after participants had been through the instructions. No participant reported not understanding the instructions or the task.

As in Experiment 2, after seeing the instructions participants received the datasets in a random order. The order in which the items in each dataset were arranged was also randomized and participants were allowed to see items in any way they liked (see the Experiment 2 procedure).

8.2. Results and discussion

The main prediction in this experiment, aside from those in the previous experiments, was that having the problem set in a specific context may help participants make more sense of what may otherwise appear to be a rather poorly-defined task.

Table 4
Performance with each of the datasets in Experiment 3

Dataset	Distinct	Robust distinct	Simplicity best
Two clusters	11	4	7
Big, small cluster	11	3	13
Three clusters	10	3	17
Little structure	19	7	0

Note: In Experiment 3 the realistic context and instructions apparently helped participants identify the simplest classifications. There were overall 28 participants.

The results from this experiment are analogous to the Experiment 2 results. Comparing solution frequency against "chance" performance, as in Experiments 1 and 2, we found that for each of the first three datasets, the frequency of the simplest categorization was above chance (respectively: $\chi^2(1) = 8.61$, $p < .01$; $\chi^2(1)$=86.15, $p < .001$; $\chi^2(1)$=80.01, $p < .001$; see Table 4). No other grouping was produced with a frequency greater than what would be expected by chance (as before, null hypotheses rejected at the .05 level). Thus, as in Experiment 1, the grouping favored by the simplicity model is most frequently chosen by participants for all three structured datasets. In the *little structure* dataset a solution different to that predicted by simplicity was produced more frequently than chance this time ($\chi^2(1) = 4.6$, $p = .03$). However, the compression associated with this classification was very close to the best possible compression for the *little structure* dataset (the difference between the two is about 90 bits; the information content of the dataset is 990 bits).

For each dataset we also computed the average compression of all categorizations participants produced and examined whether it correlates with the highest possible compression. A Pearson correlation of .97 ($p = .03$, two-tailed) between best possible compression and average compression for each dataset showed again that people's classification performance reflected a bias to select groupings that were simpler.

Fig. 4 compares the average overall compression in the four datasets for Experiments 2 and 3. It shows that context makes participants classify the items in a way more consistent with simplicity in the *three clusters* set (an independent samples, two-tailed *t*-test just failed to reach significance: $t(54) = 1.964$, $p = .055$) and with the *little structure* data: $t(54) = 3.605$, $p = .001$). In the first two datasets, performance is at ceiling in both experiments, and hence there is no difference.

Overall, the Experiment 3 results were slightly more consistent with the simplicity model compared to what we observed in Experiment 2. It is possible that the realistic context task provided a more sensitive measure of spontaneous classification performance, in that the grouping task could be readily identified with a real-life situation.

9. Experiment 4

So far the similarity relations between the items we used were derived on the basis of physical dimensions. In the final experiment we measure similarity by asking people to provide similarity

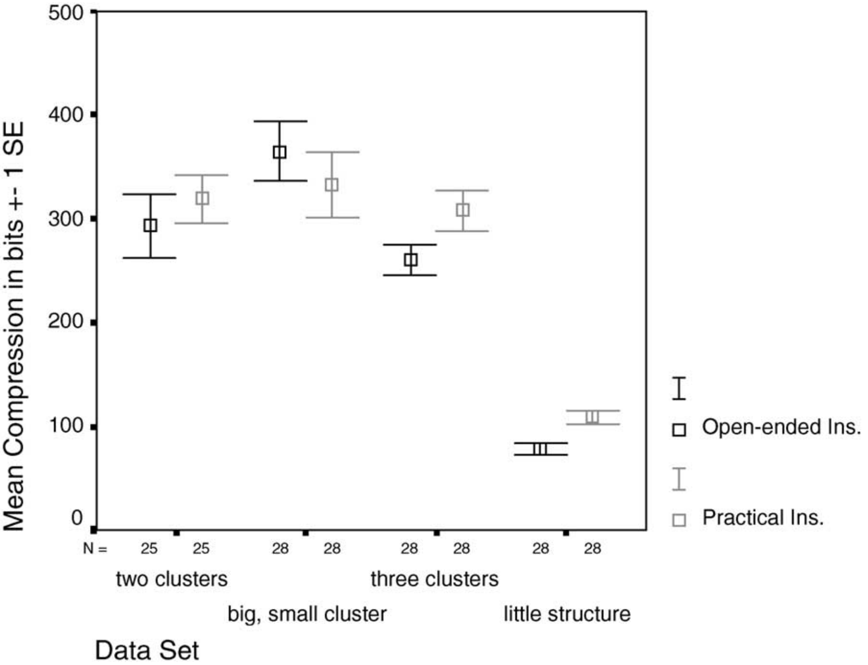

Fig. 4. Difference in the overall compression for each of the datasets, for Experiment 2 (open-ended/loosely-defined instructions) and Experiment 3 (practical/realistic instructions).

ratings. Participants were asked to numerically rate the similarity of each pair of stimuli, after they have been through the categorization task,[5] so as to obtain the similarity information that is most likely to have been reflected in the groupings produced. There is evidence that supervised categorization processes affect the perceived similarity structure of the stimuli used: the similarity relations between the stimuli are made more compatible with the required classification (see Archambault, O'Donnell, & Schyns, 1999; Goldstone, 1995; Goldstone, Steyvers, & Larimer, 1996; Harnad, 1987; Stevenage, 1998). While there are no comparable studies in unsupervised classification, it is possible that the particular way people spontaneously classify a set of items affects how they perceive their similarity structure. The simplicity model is predicting spontaneous classification on the basis of how similar to each other a set of objects are perceived. In order to obtain the appropriate prediction for the simplicity model, we thus had to use the most up to date similarity information for the categorized objects (and this would be obtained after the categorization task, allowing for possible influences of categorization on similarity).

9.1. Materials

To avoid interference between different datasets, and because the ratings part of the experiment was very time consuming, we used just one stimulus set (11 items).[6] As before, stimuli varied along two dimensions; the physical space representations used is shown in Fig. 5. The

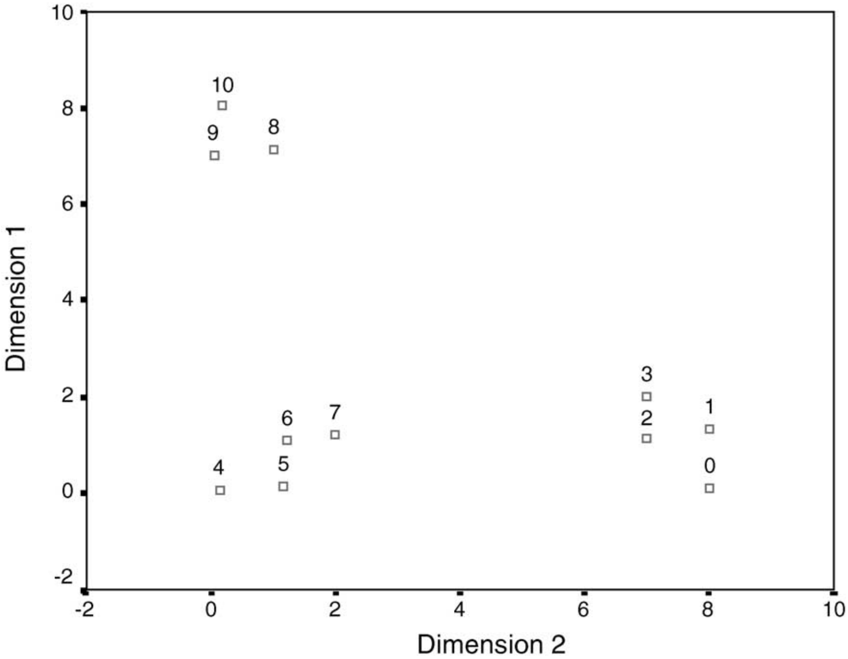

Fig. 5. The parameter space representation of the stimuli used in Experiment 4.

items can be divided into three distinct clusters. The two dimensions defined the size of a square and the size of the filled-circles texture inside the square, so that stimuli looked like the example in Fig. 6. The dimensions were intended to be perceived as independent and roughly equally salient. Successive levels on either the square-size dimension, or the texture dimension, were readily discriminable. The stimuli were presented in a folder, printed individually on A4 paper in black ink for the categorization task, and on a 15 in. Macintosh computer screen when participants were asked for similarity ratings.

9.2. Procedure

Participants were 29 University of Oxford students, who were paid for their participation. No participant had been tested in other experiments in this work and they were all tested individually. In the first part of the study, the classification part, procedure and instructions were nearly identical to the ones used in Experiment 2. After participants had classified the items, they performed the ratings task on a computer. The instructions noted that participants were about to see the items of the first part in pairs and that their task was to indicate the similarity between the items in each pair on a 1 to 9 scale, where a “1” would correspond to most similar items and a “9” to items that were most different. In particular, for each pair, the first item was presented for one and a half seconds, then there was a fixation point for 250 ms, the second item appeared for 1.5 s, a blank screen for 250 ms, and a 1–9 ratings scale. The order in which each item appeared in a pair was counterbalanced so that we had two ratings per

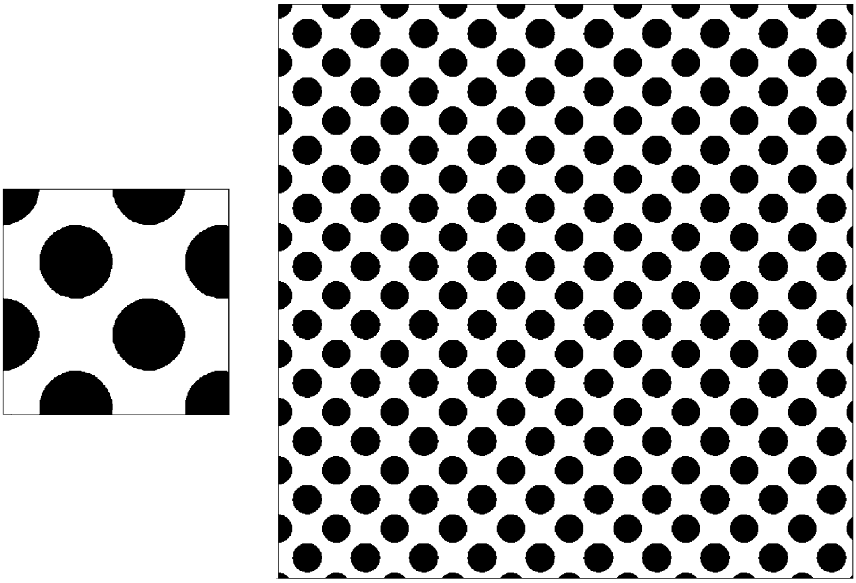

Fig. 6. An example of the stimuli used in Experiment 4, showing extreme variation in the size of the squares and the circles.

participant for each pair. We did not choose to present the items side by side since in that case we would have had to change the size of the stimuli (relative to the first part), and therefore the ratings would be in less direct correspondence to the items participants classified. Two randomized different orders were used for the ratings part of the experiment.

9.3. Results and discussion

The similarity ratings were averaged into a large similarity matrix for all the items. This matrix was made symmetrical across the diagonal by using the arithmetic mean and also self-similarities were set to 0 (corresponding to maximum similarity). The simplicity model predictions were computed on the basis of these ratings. The best compression categorization involved three groups, with items 0–3, 4–7, and 8–10 in each group (item labels correspond to Fig. 5), and is identical to that predicted using the physical space parameters in Fig. 5. The compression associated with this categorization was about 585 bits, out of 1,485 bits that would be required to code the dataset without groups (compare with the corresponding values for the previous *three clusters* case where we had a compression of 383 bits, out of 990).

In support of the simplicity model, the only categorization that came up with a frequency greater than what would be expected by chance was the simplest one ($\chi^2(1) = 84.8$, $p < .001$, using the same procedure as in previous experiments; the frequency of this categorization was 11 times, out of 29). Also, only two solutions were produced more than once (low solution variability). This demonstrates the applicability of the simplicity model in situations where similarity information is derived in a more sensitive and psychologically accurate way. It is interesting to note that ideally the simplicity model predictions would be applied to individual participant similarity information. In this work we were interested to support the model at a broad level, rather than attempt any detailed analysis, hence our approach to collect from each participant only two ratings per pair of stimuli and average these ratings. While with two ratings per stimulus pair we expected individual participant similarity information to be noisy, in future work we aim to collect more ratings from each participant so as to apply the simplicity model to individual participant similarity information.

10. General discussion

10.1. Summary

There are situations where people would spontaneously recognize that a set of objects can be organized in different groups, with no information either about the number of groups sought or the distributional properties of the objects (in other words, there are no prior expectations for the objects to be categorized). We have argued that this process of unconstrained classification is a process like perceptual organization. Utilizing the simplicity principle from perceptual organization (Chater, 1996; Hochberg & McAlister, 1953; Pomerantz & Kubovy, 1986) we proposed and empirically tested a simplicity model for unsupervised categorization.

We attempted to clarify the particular aspect of the categorization process that the simplicity model covers by contrasting it with other prominent approaches in categorization and machine learning. For example, the simplicity model is different from unsupervised categorization models where the number of categories sought is given, where specific assumptions about the distributional properties of items are made, or where the particular order of presentation of the items to be categorized is important. These differences restrict the comparability of the simplicity model with previous unsupervised categorization research. Of course, there are models of unsupervised clustering where the number of categories does not need to be specified externally, and in future work we will address the differences between the simplicity approach and such models.

It is useful to note that given the diversity of the categorization process it looks extremely unlikely that a single model will cover all observable cognitive categorization operations. For example, in the present paper we assumed that the objects to be categorized will be novel—this introduces a necessity for the model to be non-parametric (i.e., no expectations about the distributional properties of the items). But it is easy to envisage spontaneous grouping situations where the items to be categorized are drawn from a domain we are extensively familiar with. In such cases, clearly we do have specific expectations about how the features of the items will vary and it is entirely possible that parametric spontaneous grouping models will be more appropriate (e.g., Tenenbaum & Griffiths, in press). To conclude, the particular features of the simplicity model make it very suitable for one specific empirical situation, and we believe the model covers this situation well. This empirical situation is one of the many that arise in categorization, with other empirical situations requiring computational constraints different to these embodied in the simplicity model.

10.2. Simplicity

In formulating the simplicity model in this work the simplicity framework we followed most closely is that of Minimum Description Length (MDL; Rissanen, 1978). This principle asserts that the statistical model that allows the shortest encoding of the data should be preferred, where the relevant code consists of two parts: the first part encodes the 'model' of the data (here, the clustering); and the second part encodes the data in terms of the model (here this requires filling in missing constraints, and fixing incorrect constraints). The MDL principle, and closely-related notions, such as minimum message length (Wallace & Freeman, 1987), stochastic complexity (Rissanen, 1989), and Kolmogorov complexity (Li & Vitányi, 1997), represent an important growth area in contemporary statistics and machine learning. One key attraction of this type of approach is that it allows statistical inference to occur even where there is minimal prior knowledge of the data being studied—and this seems appropriate for many cognitive contexts, such as the ones described here. Interestingly, simplicity-based approaches to statistics can typically, though, be viewed as equivalent to more traditional Bayesian accounts, where the Bayesian prior assumptions are particularly broad (see Chater, 1996 for discussion).

The idea that simplicity may be an important principle in cognition has a long history, beginning with Mach (1959/1906) to contemporary research (Chater, 1997, 1999). Many aspects of cognition involve inductive inference—finding patterns, learning grammars, devising theories, creating rules and categories to explain (typically perceptual) data. The simplicity principle states that patterns, grammars, theories and the like are chosen according to which one provides the shortest encoding of the available data (which can be justified from a Bayesian perspective, and also by mathematical results showing that simple explanations provide the most reliable predictions concerning new data; Li & Vitányi, 1997, Chapter 5). Moreover, in applied disciplines concerned with inductive inference, variants of the simplicity principle have been successful (e.g., in statistics, see Rissanen, 1987, 1989; Wallace & Freeman, 1987; in computer science, see Quinlan & Rivest, 1989; Wallace & Boulton, 1968). So it seems that searching for simplicity works; perhaps some aspects of the success of the cognitive system can be explained by its preference for simplicity. Leaving normative considerations aside, there is also evidence for the importance of simplicity in psychological explanation from a range of domains, from perceptual organization (e.g., Buffart, Leeuwenberg, & Restle, 1981; Hochberg & McAlister, 1953; Leeuwenberg, 1969, 1971), aspects of language acquisition (e.g., Brent & Cartwright, 1996; Wolff, 1977; see also Tenenbaum & Xu, 2000), memory (Attneave, 1959; Garner, 1962, 1974), and theories of similarity (Hahn & Chater, 1997). Thus, the attempt to develop a simplicity-based account of how items are grouped may serve as a case study for the viability of the simplicity principle more generally.

An interesting issue is how simplicity-based models relate to *prediction*. Models of categorization frequently take one of the key goals of categorization to be to assist the agent in predicting new and unknown features of objects. For example, if a newly encountered animal is classified (e.g., on grounds of visual appearance) as a tiger, then we can predict quite a lot about its expected behavior (e.g., that it might be dangerous). Simplicity-based models of categorization also relate to prediction, but in a more fundamental way. An important mathematical result by Solomonoff (1978; see Chater, 1999 for discussion), shows that, for a very wide class of prediction problems (roughly, those where the data predicted is produced by some combination of a computable process and chance), excellent predictions are obtained by using the simplest model of the existing data and using this model to extrapolate to future data. This technical result is one justification for the success of practical prediction in statistics and artificial intelligence based on simplicity principles (e.g., Gao & Li, 1989; Quinlan & Rivest, 1989). It may also represent an important connection between the use of the simplicity principle in unsupervised categorization and the fact that successful prediction is fundamental to cognition (Chater, 1999).

10.3. Future directions

It is interesting to further explore the implications of the simplicity model for category coherence. In the present formulation of the simplicity model, spontaneous grouping has been assumed to be driven by similarity, which is reasonable since the focus so far has been novel, abstract items (for which there are no prior knowledge expectations). More generally, it is recognized that in modeling categorization processes similarity in not adequate in itself, even though models on supervised categorization based on similarity have been widely successful (e.g., Ashby & Perrin, 1988; Nosofsky, 1989). In principle, the simplicity approach to unsupervised classification is compatible both with similarity and general knowledge influences. According to the simplicity principle, the best explanation of a set of data (e.g., patterns of sensory input, or a set of objects) corresponds to the shortest description that encodes that data. Similarity- and theory-based categorizations can then be viewed as complementary ways of building short descriptions of available data. For example, the category of birds may be justified on the basis of similarity, to the extent that birds are highly similar, so that it makes sense to encode their common properties in a single category, rather than separately for each specific bird. But equally, the category of birds may be justified by the commonsense 'theoretical' generalizations defined over birds, concerning their biology, behavior and so on. One classification is preferred to another, if it provides a shorter overall description of the relevant data, irrespective of the degree to which the simple encoding is due to finding clusters of highly similar items, or describing robust theoretical regularities over the postulated categories. Investigating possible extensions of the simplicity model for non-novel items, and in a way that general knowledge effects could be taken into account, is an obvious direction for future research.

Notes

1. It is interesting to note that the categorization models originally criticized by Murphy & Medin (1985) as not providing an account of category coherence, most notably exemplar and prototype models, were models of supervised categorization; as such, category coherence is outside their scope.
2. From a technical point of view, note that the representation of the data in terms of inequalities between similarities will typically not be *completely* redundant. For example, if judgments are obtained from participants concerning whether the similarity between one pair of items is greater or smaller than the similarity between a second pair of items, then cycles such as $d(xa,\ xb) < d(xc, xd)$; $d(xc,\ xd) < d(xe, xf)$; $d(xe,\ xf) < d(xa, xb)$,

might occur (Tversky, 1977). But, to the extent that similarity behaves like a distance, we should automatically expect that some compression will be possible as a result of predictable regularity due to the metric axioms, even if the data has no cluster structure at all. This is one reason why some compression is possible even in the *little structure* dataset used in Experiments 1 to 3; the lack of structure in the stimuli is demonstrated, however, by the fact that the amount of compression is small and relatively independent of the particular cluster structure chosen.

3. Another practical problem in comparing category utility and the simplicity model is that category utility derives predictions only on the basis of featural representations. With stimuli constructed on the basis of two dimensions, as in Fig. 2, it is possible to interpret the dimensions as features, with the different values along the dimensions corresponding to different levels of the features. However, such an approach can be potentially very misleading, since the physical dimensions represent variation along certainly ordinal scales, maybe ratio as well. Different values for features, on the other hand, are simply nominal. Thus, it is fairly straightforward to construct datasets in two dimensions whereby category utility can be seen to make very counterintuitive predictions. But this would be a misapplication of the category utility principle, since it has been designed to operate on featural representations.
4. Fowlkes and Mallows (1983), observed that the Rand index is slightly biased to favor cluster configurations with more clusters; Hubert and Arabie (1985) suggested a correction that compensates for this bias. However, computational studies of the size of this bias performed by Milligan and Cooper (1986, p. 449) suggest it would be negligible compared to the differences observed in the present work.
5. Notice that this forces the pairwise comparisons to be transitive. That is, if *a* and *b* are more similar than *c* and *d*, and if *c* and *d* are more similar than *e* and *f*, then *a* and *b* must be more similar than *e* and *f*. If pairwise comparisons were measured directly, this need not be the case.
6. The stimulus set used actually consisted of 12 items; however, the 12th item was not the same in the free-sorting classification task and in the computer-based similarity ratings task. Thus, this item was eliminated from the analyses below. The classifications predicted by the simplicity principle with and without this item are identical, but for the fact that this 12th item is added to one of the groups (that is, its presence does not interact in any way with the other stimuli used, so as to lead to an overall different predicted category structure). For simplicity of exposition, we will assume this item was not present.

Acknowledgments

Part of this work was carried out while E. Pothos and N. Chater were at the Departmental of Experimental Psychology, University of Oxford. This research was supported by ERSC Grant Ref. R000222655 to Emmanuel Pothos and Nick Chater. Nick Chater has been additionally supported in part by European Commission Grant RTN-HPRN-CT-1999-00065. We would like to thank Doug Fisher, John Gennari, Frederic Gosselin, Ulrike Hahn, Francis Marriott, Brian Ripley, Philippe Schyns, and Steven Young, for their contribution in this project. We thank

Peter Hines for his insights and assistance with the formalization of some of the mathematical parts of this work.

Appendix A. Mathematical details of the simplicity model

A.1. Data

Let D be a dataset of r objects, such that a similarity function, d, obeying symmetry and minimality, can be defined between the objects.

We then assume that for all pairs of the r objects the inequality:

$$d(x, y) < d(a, b)$$

will be either true or false, where $a, b, x, y \in D$. We will call each of the inequalities possible between the elements of D "data inequalities."

Since for D there are $r(r-1)/2$ distinct similarities, there are overall $r(r^3-2r^2-r+2)/8$ data inequalities.

From information theory, each data inequality is equivalent to one bit (since determining each data inequality involves a binary choice), so that overall to specify all the data inequalities for D we require a binary word of length $r(r^3-2r^2-r+2)/8$.

A.2. Definition of a partition

A clustering is defined as a partition of D into subsets $X_0, X_1, \ldots, X_{n-1}$ such that

$$\bigcup_{i=0}^{n-1} X_i = D, \quad X_i \cap X_j = 0, \quad i \neq j$$

A clustering is defined to constraint by default data inequalities according to

$$d(a, b) < d(x, y), \quad a, b \in X_i, \quad x \in X_j, \quad y \in X_k, \quad i \neq j \neq k$$

All the data inequalities that are constrained in this way will be called constraints.

Let a particular partition of a dataset D of size r into n clusters be

$$\{X_i\}_{i=0}^{i=n-1}$$

Then the total number of constraints is given by

$$\sum_{i=0}^{n-1}\sum_{j=0}^{n-2}\sum_{k=j}^{n-1}\frac{|X_i|(|X_i|-1)|X_j||X_k|}{2}$$

Therefore, the number of data inequalities that are left unspecified by the partition are

$$\frac{r(r^3-2r^2-r+2)}{8}-\sum_{i=0}^{n-1}\sum_{j=0}^{n-2}\sum_{k=j}^{n-1}\frac{|X_i|(|X_i|-1)|X_j||X_k|}{2}$$

We call the unspecified data inequalities "free inequalities."

A.3. Codelength required for the partition

To specify the particular partition of the *r* objects into *n* clusters we need to consider all possible partitions of *r* objects into 1, ... , *r* clusters.

This is given by a binary word of length $\log_2(n+1) + \log_2(\text{part}(r, n))$, where

$$\text{part}(r,n) = \sum_{v=0}^{n}(-1)^v \frac{(n-v)^r}{(n-v)!v!}$$

We call this term "Cost to specify clusters."

A.4. Correcting erroneous constraints

In the set of constraints specified by a clustering some of them might be wrong.

That is, the clustering may specify $d(x, y) < d(a, b)$ when in fact it is the case that $d(x, y) > d(a, b)$.

Amending this involves two steps. First, specifying the number of *e* errors in the constraints requires a binary word of length $\log_2(u+1)$, where *u* is the total number of constraints. Second, selecting the number of *e* errors from the *u* constraints, requires a binary word of length $\log_2({}_uC_e)$. Thus, correcting the errors will require a binary word of length $\log_2(u+1) + \log_2({}_uC_e)$. We assume that when the errors are more than half the number of constraints for a partition then this partition is rejected.

We call this term "Cost to correct errors."

A.5. Simplicity model predictions

The simplicity model predicts that the psychologically preferred classification is the one that minimizes

$$\text{free inequalities} + (\text{cost to correct errors}) + (\text{cost to specify clusters}).$$

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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

Arabie, P., Hubert, L., & de Soete, G. (Eds.) (1996). *Clustering and Classification.* River Edge, NJ: World Scientific.

Archambault, A., O'Donnell, C., & Schyns, P.G. (1999). Blind to object changes: When learning one object at different levels of categorization modifies its perception. *Psychological Science, 10,* 249–255.

Ashby, G. F., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

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

Attneave, F. (1959). *Applications of information theory to psychology.* New York: Holt, Rinehart & Winston.

Banfield, C. F., & Bassill, S. (1977). A transfer algorithm for non-hierarchical classification. *Applied Statistics, 26,* 206–210.

Barlow, B. H. (1974). Inductive inference, coding, perception, and language. *Perception, 3*, 123–134.

Barsalou, L. W. (1985). Ideals, central tendency and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 629–654.

Bowdle, B. F., & Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology, 34*, 244–286.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts in conceptual development: Ecological and intellectual factors in categorization* (pp. 141–174). Cambridge, MA: Cambridge University Press.

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

Buffart, H. F. J. M., Leeuwenberg, E. L. J., & Restle, F. (1981). Coding theory of visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 241–274.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–591.

Chater, N. (1997). Simplicity and the mind. *The Psychologist*, 495–498.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, *52A*, 273–302.

Cheeseman, P., & Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In M. F. Usama, P. S. Gregory, S. Padhraic, & U. Ramasamy (Eds.), *Advances in knowledge discovery and data mining*. Menlo Park: The AAAI Press.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391–416.

Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of perceptual grouping. *Perception* & *Psychophysics*, *53*, 403–421.

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.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: The Free Press.

Duda, R. O., & Hart, P. E. (1973). *Pattern recognition and scene analysis*. New York: Wiley.

Evans, B. T. J., Newstead, S. E., & Byrne, R. J. M. (1991). *Human reasoning: The psychology of deduction*. Hove: Erlbaum.

Everitt, B. (1993). *Cluster analysis* (3rd ed.). London: Heinmann.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145–170.

Feller, W. (1970). *An introduction to probability theory and its applications*. New York: Wiley.

Fisher, D. (1987). Knowledge acquisition *via* incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, *4*, 147–179.

Fisher, D., & Langley, P. (1990). The structure and formation of natural categories. In B. Gordon (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 241–284). San Diego, CA: Academic Press.

Fisher, D., Pazzani, M., & Langley, P. (1991). *Concept formation: Knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufmann.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder). *Journal of the American Statistical Association*, *78*, 553–584.

Fraboni, M., & Cooper, D. (1989). Six clustering algorithms applied to the WAIS-R: The problem of dissimilar cluster analysis. *Journal of Clinical Psychology*, *45*, 932–935.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 234–257.

Gao, Q., & Li, M. (1989). An application of the minimum description length principle to on-line recognition of handprinted alphanumerals. *Proceedings of the eleventh international joint conference on artificial intelligence* (pp. 843–848). San Mateo, CA: Morgan Kaufmann.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: LEA.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition, 38*, 213–244.

Gennari, J. H. (1991). Concept formation and attention. *Proceedings of the thirteenth annual conference of the cognitive science society* (pp. 724–728). Hillsdale, NJ: Erlbaum.

Gennari, J., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–62.

Gentner, D., & Brem, S. K. (1999). Is snow really like a shovel? Distinguishing similarity from thematic relatedness. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the twenty-first annual conference of the cognitive science society* (pp. 179–184). Mahwah, NJ: Erlbaum.

Ghahramani, Z., & Beal, M. (2000). Variational inference for Bayesian mixture of factor analysers. In S. A. Solla, T. K. Leen, & K. R. Muller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 449–455). Cambridge, MA: MIT Press.

Gibson, E. J. (1991). *An odyssey in learning and perception*. Cambridge, MA: MIT Press.

Gluck, M. A., & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the seventh annual conference of the cognitive science society* (pp. 283–287). Hillsdale, NJ: Erlbaum.

Goldstone, R. L. (1993). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125–157.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178–200.

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science, 6*, 298–304.

Goldstone, R. L. (2000). Utilization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 86–112.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. *Proceedings of the eighteenth annual conference of the cognitive science society*. Hillsdale, NJ: Erlbaum.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and projects* (pp. 437–447). Indianapolis: Bobbs-Merrill.

Gosselin, F., & Schyns, P. G. (1997). Debunking the basic level. *Proceedings of the 19th meeting of the cognitive science society* (pp. 277–282). Hillsdale, NJ: Erlbaum.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics: A foundation for computer science*. Wokingham: Addison-Wesley.

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 43–92). Hove, UK: Psychology Press/MIT Press.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Oxford University Press.

Hampton, J. A. (1999). The role of similarity in natural categorization. In M. Ramscar, U. Hahn, E. Cambouropoulos, & H. Pain (Eds.), *Similarity and categorization*. Oxford: Oxford University Press.

Handel, S. (1967). Classification and similarity of multidimensional stimuli. *Perceptual* & *Motor Skills*, *24*, 1191–1203.

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., & Preusser, D. (1970). The free classification of hierarchically and categorically related stimuli. *Journal of Verbal Learning and Verbal Behavior*, *9*, 222–231.

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception* & *Psychophysics*, *12*, 108–116.

Harnad, S. (Ed.) (1987). *Categorical perception*. Cambridge: Cambridge University Press.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7–41). London: Psychology Press.

Heit, E., & Bott, L. (1999). Knowledge selection in category learning. In D. L. Medin (Ed.), *Psychology of learning and motivation*. San Diego, CA: Academic Press.

Hines, P., Pothos, E. M., & Chater, N. (submitted for publication). *A non-parametric approach to minimum description length clustering*.

Hintzman, D. L. (1986). Schema-abstraction in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Hochberg, J. E., & McAlister, E. (1953). A quantitative approach to figural goodness. *Journal of Experimental Psychology*, *46*, 361–364.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 322–330.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 418–439.

Horton, M. S., & Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development*, *51*, 708–719.

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

Imai, S., & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, *69*, 596–608.

Kaplan, A., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory* & *Cognition*, *27*, 699–712.

Katz, J. (1972). *Semantic theory*. New York: Harper & Row.

Katz, J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*, 170–210.

Krzanowski, W. J., & Marriott, F. H. C. (1995). *Multivariate analysis, Part 2: Classification, covariance structures and repeated measurements*. Arnold: London.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Leeuwenberg, E. (1969). Quantitative specification of information in sequential patterns. *Psychological Review*, *76*, 216–220.

Leeuwenberg, E. (1971). A perceptual coding language for perceptual and auditory patterns. *American Journal of Psychology*, *84*, 307–349.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folk biological taxonomies and inductions. *Cognitive Psychology*, *32*, 251–295.

Mach, E. (1959/1906). *The analysis of sensations and the relation of the physical to the psychical*. New York: Dover Publications.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth berkeley symposium in mathematical statistics and probability* (pp. 281–297). Berkeley: University of California Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception* (pp. 254–283). Cambridge, UK: Cambridge University Press.

McDermott, D. (1987). A critique of pure reason. *Computational Intelligence*, *3*, 151–160.

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

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

Medin, D. L., & Wattenmaker, W. D. (1997). Category cohesiveness, theories, and cognitive archeology. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge, UK: Cambridge University Press.

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

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987b). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

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

Michalski, R., & Stepp, R. E. (1983). Automated construction of classifications: Conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*, 396–410.

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.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Journal of Experimental Psychology: Perception and Psychophysics*, *38*, 415–432.

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 700–708.

Nosofsky, R. M. (1988). Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 54–65.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Journal of Experimental Psychology: Perception and Psychophysics*, *45*, 279–290.

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

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*, 94–140.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind and Language*, *6*, 1–38.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

Osherson, D. N. (1990). Judgment. In D. N. Osherson & E. E. Smith (Eds.), *Thinking: An invitation to cognitive science*. Cambridge, MA: MIT Press.

Pickering, M., & Chater, N. (1995). Why cognitive science is not formalized folk psychology. *Minds and Machines*, *5*, 309–337.

Pomerantz, J. R. (1981). Perceptual organization in information processing. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 141–180). Hillsdale, NJ: Erlbaum.

Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, volume II: Cognitive processes and performance* (pp. 1–45). New York: Wiley.

Posner, M. I., & Keele, S. W. (1968). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304–308.

Pothos, E. M., & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary or sufficient features? *British Journal of Psychology*, *91*, 439–450.

Quine, W. V. O. (1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 155–175). Ithaca, NY: Cornell University Press.

Quinlan, R. J., & Rivest, R. L. (1989). Inferring decision trees using the Minimum Description Length Principle. *Information and Computation*, *80*, 227–248.

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

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

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 347–363.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B*, *49*, 223–239.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry*. Singapore: World Scientific.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

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

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyles-Brian, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Erlbaum.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, *15*, 461–508.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition*, *67*, 147–179.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1997). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1–54.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

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

Smith, D. J., & Baron, J. (1981). Individual differences in the classification of stimuli by dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1132–1145.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, *24*, 422–432.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology*, *89*, 39–57.

Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 517–522). Hillsdale, NJ: Erlbaum.

Tenenbaum, J., & Griffiths, T. L. (in press). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

van der Helm, P. A., & Leeuwenberg, L. J. (1996). Goodness of visual regularities: A non-transformational approach. *Psychological Review*, *103*, 429–456.

van der Helm, P. A., & Leeuwenberg, L. J. (1999). A better approach to goodness: Reply to Wagemans (1999). *Psychological Review*, *106*, 622–630.

van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory* & *Cognition*, *10*, 396–404.

van Oeffelen, M. P., & Vos, P. G. (1983). An algorithm for pattern description on the level of relative proximity. *Pattern Recognition*, *16*, 341–348.

von Helmholtz, H. (1910/1962). *Treatise on physiological optics*. In J. P. Southall (Ed.) (Vol. 3). New York: Dover.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computing Journal*, *11*, 185–195.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, *49*, 240–251.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology*, *51B*, 235–270.

Wittgenstein, L. (1957). *Philosophical investigations* (3rd ed.). Oxford, UK: Blackwell.

Wolff, J. G. (1977). The discovery of segmentation in natural language. *British Journal of Psychology*, *67*, 377–390.

Zippel, B. (1969). Unrestricted classification behavior and learning of imposed classifications in closed, exhaustive stimulus sets. *Journal of Experimental Psychology*, *82*, 493–498.