Categorization creates functional features

Philippe G. Schyns
University of Glasgow
and
Luc Rodet
University of Grenoble

in press, *Journal of Experimental Psychology: Learning, Memory & Cognition*

Please address all correspondence regarding this manuscript to:
Philippe G. Schyns
Dept. of Psychology
University of Glasgow
Glasgow G12 8QB
United Kingdom
philippe@psy.gla.ac.uk

ABSTRACT

Many theories of object recognition and categorization claim that complex objects are represented in terms of characteristic features. The origin of these features has been neglected in theories of object categorization. Do they form a fixed and independent set that exists before experience with objects, or are features progressively extracted and developed as an organism categorizes its world?

This paper maintains that features can be flexibly learned, as a consequence of categorizing and representing objects. All three experiments reported in this paper used categories of unfamiliar computer-synthesized two-dimensional objects ("Martian cells"). The results showed that varying order of category learning induced the creation of different features that changed the perceptual appearance and the featural representation of identical category exemplars. Network simulations supported a flexible, rather than a fixed feature interpretation of the data.

Many theories of object recognition and categorization assume that objects are represented in memory as groups of components. To classify an object, one must first identify its components and then compare them to memory representations. For example, when a person sees a cup, he or she might first identify a container or a handle before categorizing the object properly. Of course, not all components of an object are necessary for its categorization, but many of them are probably identified during the recognition process. Componential accounts that embody this general approach include theories of object recognition and categorization (see, among others, Biederman, 1987; Marr & Nishihara, 1978; Rosch & Mervis, 1975; Smith & Medin, 1981; Treisman & Gelade, 1980).

Although most object categorization theories are componential, less attention has been paid to the origin of the components themselves. Do they form a fixed and independent vocabulary that exists prior to the experience with objects, or are features progressively learned and developed as an organism categorizes and represents its world? Most current models of category learning leave aside the issue of feature learning and feature development. Their feature set is fixed and unaffected by the classification and learning processes.

Classification and learning processes, however, operate on a stable featural analysis (a perceptual organization) of the ever changing retinal input. Even though our sophisticated visual apparatus probably comes equipped with *a priori* ways of analyzing and organizing retinal images, there are occasions when a relevant perceptual analysis is not readily available. For example, complete novices reading chest X-rays (e.g., Christensen, Murry, Holland, Reynolds, Landay & Moore, 1981), sexing chicken (Biederman & Shiffrar, 1987), and categorizing dermatosis (Norman, Brooks, Coblenz & Babcock, 1992) have little understanding of the relevant dimensional structure of these categories. Even when told what the signs of different diagnosis are, novices are not always able to see the features experts use to organize the input. If one takes a developmental perspective, it seems clear that infants and young children are not always able to analyze objects using all the stimulus dimensions that are used by adults (Smith, Carey, & Wiser, 1985; Smith & Kemler, 1978; Ward, 1983).

Thus, there is suggestive evidence that features are flexible--i.e., they adjust to the perceptual experience and the categorization history of the individual. Flexible features open the possibility that the same input is differently perceived and analyzed before being categorized. Hence, a complete theory of categorization and conceptual development should not only explain the ways in which object features are combined to form concepts, it should also explain the origin and the development of the features participating to the analysis of the input. The studies presented in this paper investigate further the claim that a significant part of learning a category involves learning the features entering its representation (Schyns, 1991; Schyns & Murphy, 1994; Thibaut, 1991; Wisniewski & Medin, 1994).

Fixed space category learning.

In Bruner, Goodnow and Austin's (1956; Bourne, 1982; Shepard, Hovland & Jenkins, 1961) classical concept learning experiments, subjects were shown simple colored geometric shapes and were instructed to learn the rules for their categorization. These rules were logical combinations of the features that were clearly demarcated in the stimuli. For example, subjects could learn that the rule "red and circle" categorized the objects, but there was little doubt about the featural analysis of the stimuli. A fixed set of *a priori* diagnostic features was used to design the stimuli and the space of possible concepts was effectively limited to possible combinations of these features.

Although concept learning and object recognition theories have come a long way since these early experiments, simple stimuli with clear-cut dimensions are still often used for reasons of experimental control (e.g., Gluck & Bower, 1988; Murphy & Smith, 1980; Wattenmaker, 1991; among many others). In a typical experiment, subjects are given a complete listing of the features composing the training items. Alternatively, stimuli are composed of a few clearly demarcated perceptual dimensions (e.g., color, texture, shape and size), so that there is an implicit agreement between the experimenter and the subject about possible features.

There are considerable practical advantages to fixing the features describing objects. For example, the issue of whether concepts are represented as exemplars, prototypes or classical definitions could not be resolved empirically unless the experimenter and the subjects agreed on what features describe the stimuli. Therefore, the exemplar model of Medin and Schaffer (1977) and its successors specify a number of stimulus dimensions that the items differ on. It is only these dimensions that form the basis for similarity comparisons that determine category learning. Similarly, Gluck and Bower (1988) encode items as individual pre-defined features or combinations of those features, and associations between such units and a category are the basis for learning (see also Anderson, 1991; Elio & Anderson, 1981; Estes, 1986; Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley & Glauthier, 1994; and many others). We should be careful, however in pointing out that many of these theorists do not explicitly claim that such fixed-feature learning is the only learning that occurs. Indeed, most simply do not address the question.

Some object recognition researchers have explicitly assumed a fixed feature set, and this is one possibility for a theoritical stance for categorization. For example, Biederman (1987) Recognition By Components theory claims that real-world objects are represented in memory with a combination of primitives taken from a fixed set of 36 “geons” (geometric elements). An important assumption of a fixed set stance is that primitives are the lowest, non-decomposable building blocks of object representation (see, e.g., Biederman, 1987). Using set notation, if $F^t = \{f_1, f_2, ..., f_n\}$ defines the fixed set of basic features available at time t, R^t the powerset[1] of F^t, defines the repertoire of all objects representations expressible with, for example, feature conjunctions. That is, $R^t = \{-, f_1, f_2, ..., f_n, f_1 \wedge f_2, ..., f_1 \wedge f_n, f_2 \wedge f_3, ..., f_2 \wedge f_n, ..., f_1 \wedge f_2 \wedge ... \wedge f_n\}$, where "^" represents the logical *and.* Although very simple, this characterization captures the idea that a set of fixed elements and construction rules can generate a large number of distinct object descriptions. For example, Biederman (1987) claims that different combinations of 36 geons span all possible basic-level object categorizations.

Componential theories of category learning should offer powerful principles for developing new representations, if only to account for the flexibility of conceptual development and the diversity of object categories the individual may encounter. Fixed space models would limit new concepts to new combinations of fixed features--i.e., to elements of R^t that are not actual concepts in memory. Consequently, if a categorization requires a feature that is not originally present in the fixed set (e.g., if F^t features must distinguish between two object descriptions $f_1 \wedge f_n$ and $f_1 \wedge f_n \wedge f_{n+1}$) the categorization cannot be learned because it violates the assumption that the set is fixed. Similarly, if a categorization requires the decomposition of a basic feature (e.g., if f_1 of F^t must be decomposed into f_{1*} and f_{1**} to distinguish between two objects $f_{1*} \wedge f_{1**} \wedge f_n$ and $f_{1**} \wedge f_n$), the categorical difference cannot be represented because it violates the assumption that basic features are not decomposable.

Of course, a clever theorist could always handcraft a fixed basis of features that is capable of solving a few, or even many instances of these two generic difficulties. These, however, do not challenge the competence of category learning theorists, but the

[1]Formally, the *closure* on the feature set is a more accurate description of this set. For simplicity of analysis, we approximate the conceptual repertoire with the assumption that no feature is ever repeated in a concept--i.e., each feature is tokenized. For example, a horizontal bar and a rotated bar would be represented by two different features. Feature tokenization reduces closure to the powerset of F^t, a set of size 2^n.

competence of category learning theories. In sum, we are claiming that a fixed feature stance does not have a way of accounting for the creation of new features, and we will argue and empirically demonstrate that such a process will have to be included in a complete account of category learning.

Flexible space category learning.

Starting from the functional role of features (i.e., to distinguish between object categories), we want to work backwards and suggest that the constraint of distinguishing between objects categories may induce the creation of new stimulus dimensions--i.e., the dimensions, or "functional features" solving the categorization. The *Functionality Principle* summarizes this view (Schyns & Murphy, 1994, p. 310): "If a fragment of a stimulus categorizes objects (distinguishes members from nonmembers), the fragment is instantiated as a unit in the representational code of object concepts."

In the formalism described above, at time $t+1$ of conceptual development[2], a new functional feature increases the dimensionality of the feature space ($F^{t+1} = \{f_1, f_2, ..., f_n, f_{n+1}\}$) and doubles the size of the conceptual repertoire ($R^{t+1}/R^t = 2^{n+1} / 2^n$). These estimates should not be taken literally, but they nonetheless capture the idea that the combinatorics of a componential approach applied to features flexibly grounded in the history of the categorizer offers more powerful representations to understand the complexity of conceptual development. Specifically, our proposal of flexible features relaxes the assumption of a fixed vocabulary, to allow the representation of new structures as new features, and it also relaxes the assumption that features are not decomposable, to authorize a decomposition of existing features representing coarse structures into sub-features representing finer structures. There will be limits to this decomposition, but conceptual expertise in science and everyday life often demonstrates that there are many

---

[2] Henceforth, conceptual development refers both to concept acquisition in childhood and adulthood, and to the changes in understanding that result from applying the newly developed features and concepts to the outside world.

levels of outside world structures to be discovered between those imposed by our F^t features and the limiting resolution of our perceptual apparatus throughout conceptual development. The goal of the present work, therefore, will be to document such feature construction and decomposition, and to outline some of the circumstances under which they will occur.

Although not addressed by fixed feature theories, there is a substantial body of evidence suggesting that the interactions between categorization and perception affect the features and dimensions representing objects (see Schyns, Goldstone & Thibaut, 1995, for a more thorough review). One of the simplest forms of influence is pre-differentiation, a procedure whereby pre-exposure to stimuli enhances sensitivity to some of their aspects in a later testing phase (e.g., Gibson, 1991; Gibson & Walk, 1956). In Gibson and Walk (1956), one group of rats was exposed to triangles and circles, but rats were not required to categorize these shapes. In a later task, these rats were better able to learn to discriminate between these triangles and circles than a control group which did not have this experience. Gibson (1991) interprets such effects of pre-exposure in terms of perceptual differentiation, a process whereby discriminative aspects of the stimuli are made more salient. Perceptual differentiation is typically described as an *unsupervised* form of learning; feedback on the categorizations is usually not required for sensitization, pre-exposure suffices.

Although pre-differentiation studies indicate that feedback is not mandatory to learn new stimulus aspects, other studies suggest that explicit categorizations exerts additional influences on the learning of relevant object structures. There are two main ways in which categorization exerts such influence (Schyns et al., 1995). First, it can affect the subjective distances between values within existing dimensions, or enhance the discriminability of entire preexisting dimensions. To illustrate, Goldstone (1994) trained human subjects on the categorizations of squares varying in size or brightness. After extensive training, subjects were given a same/different task in which squares were presented that varied on their size and brightness dimensions. It was found that subjects' same/different judgments were sensitized to the dimension (size or brightness) that was relevant during the training stage. Furthermore, within the categorization-relevant dimension, the greatest sensitization was found for dimensional values serving as the boundary between the learned categories. Because even the simplest same/different judgments involve cognitive factors such as attention, short-term memory and encoding, these results do not isolate a perceptual change *per se,* but they nonetheless demonstrate that categorization diagnosticity influences a task that is often assumed to tap relatively low-level processes (Schyns et al., 1995).

A second form of influence of category diagnosticity on low-level perception is the discovery of the features (or dimensions) of object categorization themselves. For example, complete novices who observe X-rays have no understanding of their relevant features. However, this is not simply an issue of not knowing the medical category going with each X-ray; medical students know the signs of different diagnoses, but they are not always able to see them in X-rays (see also Christensen et al., 1981; Kundel & Nodine, 1975; Lesgold et al., 1988; and Lesgold, 1984; Myles-Worsley, Johnston, & Simons, 1988; Norman, Brooks, Coblenz & Babcock, 1994). Although there are perceptual aspects of the X-rays that experts use, these are probably not clearly demarcated perceptual features that we have pre-specified representations of. Learning to classify X-rays and other expert categorizations (e.g., sexing chicken, Biederman & Shiffrar, 1987; categorizing determatosis, Norman, Brooks, Coblenz & Babcock, 1992) probably involves learning to identify new features as much as learning which feature characterizes which medical diagnosis. These perceptual changes directly result from the requirement of distinguishing new diagnostic structures in unfamiliar materials.

Schyns and Murphy (1991, 1994) systematically investigated such influence of categorization on feature discovery. In their experiments, subjects were asked to learn new object categories before being tested on their featural encodings of test items. The stimuli were continuous two- or three-dimensional shapes ("Martian rocks") composed of blobs. The Martian rocks had a complex blob structure so that naive subjects had no idea of what the relevant components of the rocks were. Categories were defined by an unfamiliar blobby shape that was common to all members; the remaining blobs were given random shapes. After learning different categories, subjects were instructed to segment the identical test rocks into their relevant parts. It was observed that different groups of subjects noticed different structures in the same Martian rocks, even though these rocks initially appeared as unstructured blobby shapes. Experiments with the Martian rocks suggested that the requirements of categorizing and representing new objects prompted the acquisition of their features. This type of influence of supervised categorizations on perception is the main object of our studies.

Feature creation vs. feature weighting.

Retinal images of complex objects such as Martian rocks and real world objects are stimuli in a high-dimensional space--formally, each retinal receptor constitutes one dimension of the initial input space. Studies in computational vision and neural networks have shown that categorization algorithms can hardly operate on the raw pixel values of an image; some qualitative description of the input must be first obtained. In other words, high-dimensional retinal images must first be recoded into (or projected onto) a low-dimensional basis of categorization features. Formally, this is a process of *dimensionality reduction* (Schyns et al., 1995).

It is this process of dimensionality reduction, and how category learning influences it, that has been neglected in modern category learning theories. However, a simple analogy illustrates the importance of dimensionality reduction for conceptions of category learning. Consider Principal Component Analysis (PCA), a dimensionality reduction method which projects high-dimensional data on low-dimensional axes. The axes are called the *Principal Components*, and they represent the main sources of variations of the data. To illustrate the impact of this analogy, imagine a data set composed of a face picture included in an ensemble of 20 other faces, and a second set composed of the same face included amongst 20 other object pictures. These two sets have different main sources of variations, and PCA will represent the variations with different basis of principal components. Hence, the face common to both object sets will "look" different to systems using these encodings based on different features. Although PCA is only an analogy (its features are those representing the main directions of variations of the data, not their categorizations, and the two do not necessarily correspond), it nonetheless illustrates an important point: If dimensionality reduction was an integral part of category learning mechanisms, individuals with different categorization histories could acquire different means for analyzing and perceptually organizing identical materials. Because these perceptual analysis are the materials available to categorization processes, the latter could not be understood without a thorough understanding of the former.

Of course, leading categorization models such as Kruschke's (1992) ALCOVE, Nosofsky's (1986) Generalized Context Model and Anderson's (1991) rational model also operate on low-dimensional featural recodings of the input (in fact, they must, and we will come back to this point in the General Discussion). However, the modeler, not the model, performs the reduction of dimensionality by choosing the appropriate features for the considered categorizations. Thus, the featural analysis and the perceptual organization of the input do not change with categorization experience. Instead, both are fixed by the modeller's feature basis and category learning mechanisms are limited to the selective weighting of the features that are diagnostic in the task (see, e.g., Anderson, 1991; Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, 1986; for implementations of these mechanisms)[3]. The main empirical goal of the present research is to distinguish

---

[3] We should be careful to point out that these models are models of fixed features experiments and were not designed with the goal of dimensionality reduction in mind. The models are used to illustrate the formal contrast between fixed and flexible approaches to category learning. We come back to the practical difficulties of constructing formal models of flexible feature learning in the General Discussion.

between this form of fixed space learning which selects and weights existing features from another type of category learning which changes the perception of the input by creating new features.

All our experiments used categories of novel shapes. As we wish to provide an “existence proof” of a feature learning mechanism, we wanted to use complex, unfamiliar, but controlled stimuli; stimuli that preserved the complexity of real-world objects instead of the familiar type of stimuli varying along a couple of clearly demarcated perceptual dimensions, or other such stimuli whose analysis was so obvious that they could produce a systematic bias against promoting feature learning and development. For example, we could not use categories of chairs, or similarly familiar object categories because subjects would already know their features.

To these aims, we synthesized categories of two-dimensional "Martian cells" with image processing softwares on a computer. We created the cells with the hope that they would be as dense as pictures of biological cells, in which new structures often are discovered. A Martian cell was created from a uniformly gray circle, by adding black components (cell bodies) at random locations within the circle. Some components were common to all category members, and other components were random fillers. Components were designed with the constraint that they would not resemble familiar shape features. A Gaussian filter was then applied to each cell to blend together the background circle and the bodies (see Figure 1 for a Martian cell exemplar).

The stimuli we use in our studies are complex and unfamiliar, as in expert categorization studies (chicken sexing, categorizing X-rays and dermatosis, and so forth). However, we should point out that our main objective is not to study expert categorization *per se*--Martian cells are only used to limit the knowledge that can be applied to stimulus analysis. Instead, our main goal is to demonstrate that categorization demands (and more generally experience with objects) can change the features that participate to the input analysis. Thus, we will not look at changes “after the fact,” but predict and explicitly test the nature of the perceptual changes that take place. To illustrate, consider Figure 1, which presents different analyses of the top Martian cell. To the observer who knows the rule of stimulus composition, the conjunction of features *a^b* (or *b^c*) and the configuration *ab* (or *bc*) are formally identical (see Figure 1, the bottom drawings). To the categorizer, however, the perceptual appearances of a conjunction and a configuration are very different: A conjunction is only identified after its components are identified, while a configuration is immediately perceived. A configural representation is not composed of subfeatures; it is a different holistic feature altogether. Our experiments rely on this important dissociation between conjunctive and configural features to demonstrate different perceptions (hence, different analysis) of identical aspects of stimuli.

A preliminary experiment tested that a simple pre-exposure to stimuli, without category feedback, was sufficient to change the features that were later applied to the description of identical test items. The first and second experiments were designed to demonstrate that categorization feedback can exert a profound influence on category learning mechanisms. Contrary to the fixed feature assumption that people who learn to categorize objectively identical stimuli analyze and represent them with identical features, experiments 1 and 2 tested that different features of analysis were learned as a result of categorizing the same objects.

PRELIMINARY EXPERIMENT

As explained earlier, pre-differentiation is one of the simplest form of perceptual learning in which an unsupervised pre-exposure to stimuli affects their later discriminability (Gibson, 1991; Gibson & Walk, 1956). In this preliminary experiment, we specificially wanted to test that pre-exposure to different objects was capable of changing the features that were later used to describe an identical set of test items. Specifically, we wanted to show that the same stimuli could receive mutually exclusive featural analysis, following different pre-exposures. Although this result would not strictly isolate mechanisms of feature creation, it would nonetheless indicate that even in the simplest form of perceptual learning (a learning without categorization feedback) the featural analysis of the input is not perceptually fixed, but flexibly adjustable to the experience of the organism.

Two groups of subjects were pre-exposed to different ensembles of "Martian cells." Each cell in the set was created from a particular combination of the *a*, *b* and *c* elementary components to which were added random cell bodies on a textured background (see Figure 1). For example, the *a* component was common to *A* cells, and *BC* cells were defined by the adjacent *b* and *c* components (see Figure 1). Cell bodies were randomly located in the cells, and the stimuli were designed to ensure that both groups would experience equally often the elementary *a, b,* and *c* blobs. Specifically, the *A-BC* group was exposed to 5 *A* cells and 5 *BC* cells; the *AB-C* group experienced 5 *AB* and 5 *C* cells (exemplars of *A* , *BC*, *AB* and *C* cells are shown on Figure 2). Each cell was presented only once, on a separate sheet of paper, and subjects could see each cell for as long as they wished. Subjects were instructed to carefully observe the cells because they would be asked questions about them. The stimuli were presented a maximum of 3 times, depending on subjects' self-assessment of their learning. Most subjects felt they knew the stimuli after two repetitions.

Remember that a pre-exposure to 10 stimuli is quite different from a typical category learning task. In pre-exposure, subjects are simply instructed to observe unknown stimuli, not to categorize them. In fact, subjects should not know whether to categorize the 10 Martian cells into 1, 2, 3, or 10 categories because no teacher ever gives the categorization feedback necessary to learn any classification rule. Pre-exposure is simply designed to sensitize subjects to the common structures of the stimuli.

INSERT FIGURE 1 & 2 AROUND HERE

Even though both groups saw on average the *a*, *b* and *c* components of stimulus design equally frequently, we nonetheless expected that pre-exposure would differentially affect subjects' sensitivity to different aspects of the test items. Specifically, we expected *A-BC* subjects to become spontaneously sensitized to two unitary configural features that we label $F = \{a, bc\}$, while the *AB-C* group should be sensitized to another, orthogonal feature vocabulary $F = \{ab, c\}$ (these parsings are illustrated by the bottom pictures of Figure 1). The two vocabularies are orthogonal because each represents a configural feature (*bc* or *ab*) that the other vocabulary would analyze as a conjunction. For example, sensitization to the *c* feature in the *AB-C* group should enable these subjects to decompose the *bc* feature that subjects of the *A-BC* group perceive as a unitary configuration (symmetrically, the *a* feature of *A-BC* should decompose the *ab* feature the other group should perceive unitarily). Thus, orthogonal feature vocabularies should elicit orthogonal analysis of the stimuli, which in turn should elicit mutually exclusive parsings of test items.

To test this hypothesis, we recorded subjects' free parsings of five test cells. Cells were presented one at a time on a computer screen, and subjects could freely draw outlines around the cells' parts using a computer mouse (see Schyns & Murphy, 1994). Remember that subjects were not asked to categorize the test items. Instead, they were simply instructed to draw outlines around the parts they *saw* during the pre-exposure. Test cells were named *ABC*, *AB*, *BC*, *A* and *C* to reflect the configurations they tested (the five test cells are shown on Figure 2).

We scored the groups on whether their segmentations of *ABC* was compatible with $F = \{a, bc\}$, or with $F = \{ab, c\}$. In the *A-BC* group, 83% of *ABC* segmentations were compatible with $F = \{a, bc\}$. In *AB-C*, all of these segmentations were compatible with $F = \{ab, c\}$. To further test the hypothesis that the groups created different feature vocabularies, we compared the delineations of the *AB* and *BC* test stimuli. Specifically, we wanted to know whether the *ab* and the *bc* units were truly seen as unitary features, instead of simple conjunctions of *a*, *b* and *c*. We hypothesized that *AB-C* (vs. *A-BC*) subjects should delineate *AB* (vs. *BC*) as a unitary feature and segment *BC* (vs. *AC*) because this group was sensitized to $F = \{ab, c\}$ (see Figure 1). We scored the groups on whether they divided *ab* or *bc*. In *A-BC* all divisions were divisions of *AB*; 89% of divisions were divisions of *BC* in *AB-C*. This result clearly indicates that the unitary feature of one group was composed of two features in the other group. This result suggests that the simple *a* and *c* parts may have a different status in each group. Specifically, *A-BC* (vs. *AB*-C) subjects should perceive *a* (vs. *c*) as a functional feature but dismiss *c* (vs. *a*) as "cytoplasmic noise." We also scored the groups on whether they circled *a* in *A* stimuli and *c* in *C* stimuli. In *A-BC*, most subjects circled *a* and most subjects circled *c* in *AB-C*.

In summary of the preliminary experiment, it appears that a simple unsupervised pre-exposure to different stimuli can sensitize subjects to orthogonal structures of identical stimuli. Specifically, each group perceived a unitary configuration that the other group analyzed as a feature conjunction. If such a simple procedure can elicit orthogonal analysis of identical materials, it is possible that the analysis of real-world objects for categorization is not performed with an identical set of fixed features.

EXPERIMENT 1

The preliminary experiment suggested that perceptual experience influences the features that describe objects. However, our theoretical argument mainly concerns the featural changes that result from categorization feedback and the requirement of discovering a feature basis that distinguishes categories. Although a weak case could be made that subjects discovered an implicit categorization during the pre-exposure stage, we still need to positively demonstrate that categorization feedback can drive the discovery of new features that affect the perceptual appearance of objects.

Specifically, we need to rule out the possibility that subjects in our experiments initially analyzed and perceived the stimuli with the elementary blobs *a*, *b* and *c* that we, as experimenters, used to design the Martian cells. The supplementary information of categorization feedback could help subjects to recombine these blobs into diagnostic categorization rules (e.g., Bruner et al., 1956; Elio & Anderson, 1981; Nosofsky, 1986; among many others). In this case, categorization feedback would *not* change the perceptual analysis of the stimuli. Instead, it would only affect a high-level, strategic decision rule based on *fixed* perceptual evidence. Experiments 1 and 2 were designed to rule out such a strategic, high-level interpretation of the feature construction process we propose. Our main goal was to demonstrate that category learning could flexibly change the units participating in the perceptual analysis of the stimuli.

Consider the standard approach of concept learning research applied to the very simple problem of distinguishing between two categories *X* and *XY*. First, a fixed feature vocabulary $F = \{x, y\}$ is specified. Secondly, two categories *X* and *XY* are constructed by the experimenter by combining features from this set. Thirdly, subjects are exposed to a task asking them to learn the two categories. Fixed feature approaches often assume that people then analyze input objects with features from the fixed set (see, e.g., Biederman, 1987; Biederman & Gerhardstein, 1995; Bourne, 1982; Murphy & Ross, 1994; Spalding & Ross, 1994). That is, people analyze *XY* objects as a conjunction of feature *x* and feature *y*. Consequently, they must perceive and use *y* to distinguish *XY* from *X*. Experiment 1 reproduces this classical approach, but concludes that even such a straightforward categorization problem may not necessarily have this obvious featural solution (see also Shin & Nosofsky, 1992).

We designed two Martian cell categories *X* and *XY* as just explained. Two groups of subjects were instructed to learn the categories, but in a different order. The *X->XY* group learned *X* first, but the *XY->X* group initially learned *XY* (see Figure 3). We hypothesized that subjects unaware of the relevant analysis of new Martian cells would not necessarily perceptually organize the cells with the objective *F* vocabulary that we, as experimenters, used to compose the stimuli. Consequently, we expected subjects to learn the relevant featural analysis of the stimuli as they learned their categories.

INSERT FIGURE 3 AROUND HERE

Here is how it would work. Subjects initially learning *X* should construct a unit *x* for its representation. When subjects experience the second category *XY*, *x* is the only feature that is initially perceived in the new exemplars, which are then perceived as *X* exemplars[4]. Such miscategorizations (subjects are instructed that the new exemplars form a new category) should force the discovery of another new structure (that we call *y*) to distinguish the new category. Note that this process might not be as obvious as discovering that *triangle* distinguishes the category *SQUARE+TRIANGLE* from *SQUARE*. Unlike the known shape *triangle* to which attention can be immediately shifted, *y* might require perceptual learning for its representation. Thus, the initial extraction of *x* from the *X* category in the *X->XY* condition would force the later construction of *y* to represent *XY* as a distinct feature conjunction in memory.

Feature construction could be very different in the group initially exposed to *XY*, because unlike subjects of the other group, this group has not learned the *x* feature to analyze *XY* exemplars. Subjects initally exposed to *XY* could then learn another configuration (that we call *xy*), without ever being aware of the feature conjunction the other group would perceive. Thus, a flexible feature construction process would make the rather counterintuitive prediction that learning identical categories can result in the construction of different features (specifically, $F^t_{XY\text{->}X} = \{xy, x\}$ and $F^t_{X\text{->}XY} = \{x, y\}$).

Differences in feature vocabularies could be revealed by an atypical test stimulus, such as an *X-Y* cell in which the features *x* and *y* are not adjacent to one another. The existence of *y* in $F_{X\text{->}XY}$ should result in more *XY* categorizations of *X-Y* cells, but the absence of *y* as a distinct feature in $F_{XY\text{->}X}$ should elicit more *X* categorizations of *X-Y*--even though the two groups would be equally adept at distinguishing *X* and *XY* cells. Together, these results would question the applicability of the classical approach claiming that the fixed vocabulary $F = \{x, y\}$ must be used to distinguish *X* from *XY*.

Methods.

Subjects. Twenty-one University of Grenoble students participated in the experiment.

---

[4] Spalding and Ross (1994) have called this effect "focusing" on a feature that is common to compared

Stimuli. Stimuli were Martian cells synthesized as explained before by combining category defining and random cell bodies on a gray background. We synthesized three categories of Martian cells for training. The conjunctive part *a^b* defined the *AB* category (*a* and *b* were adjacent on each exemplar). The subcomponent *a* and the subcomponent *b* defined the *A* and *B* categories, respectively. Five types of cells were used for testing: *AB*, *A*, *B*, *A-B* and distracters (see Figure 4). We describe these stimuli in more details in the presentation of the testing phase.

Procedure. *Initial Learning Phase*. In an initial learning phase, subjects were instructed that they would have to learn two separate categories of Martian cells. Each category was composed of 10 cells. Stimuli were presented one at a time for 1 sec. on a CRT screen. After viewing each category, subjects' learning was evaluated with 4 new cells (two new exemplars and two distracters). Subjects' task was to categorize each cell by pressing the appropriate key of a computer keyboard (possible responses were *learned_category*, *other*). A single mistake would restart the initial learning phase, and so perfect categorization was required to complete training.

We defined four groups of subjects according to order of presentation of the categories. In the *A->AB* group and the *B->AB* groups, the simple category (either *A* or *B*) was learned before the conjunctive *AB* category. This order was reversed in the *AB->A* and the *AB->B* groups. For ease of presentation, we will henceforth call the first two groups the *X->XY* learning condition and the other groups the *XY->X* learning condition, and refer to *x* as the component defining the simple category, *xy* as the components of the conjunctive category, and to *y* as the subcomponent of *XY* that did not in itself define a category.

Verification Phase. To insure that the memory of the second category did not affect the memory of the first category, we ran a categorization test after completion of the initial learning phase. Subjects were instructed to categorize 6 new cells (two simple, two

---

instances. We come back to this point in the General Discussion.

conjunctive, and two distracter cells), each presented for 2 s. on a computer monitor (possible responses were *Category_1*, *Category_2* or *none*, on the appropriate key of a computer keyboard). A single mistake would eliminate a subject from the experiment[5].

Testing Phase. In the testing phase, subjects were asked to categorize 20 new cells: 4 *X*, *4 XY*, 4 *X-Y* and 8 distracters. To ensure that subjects would categorize after (instead of during) their visual exploration of the cells, we simulated the effect of looking at cells through the aperture of a microscope. Figure 4 illustrates the procedure. For each cell, subjects saw 2 successively presented snapshots of different portions of the cell. For the *X* and *XY* cells, one snapshot showed the relevant component (*x* or *xy*) and the other snapshot revealed a distracter part (see Figure 4). For *X-Y* cells, *x* was presented in one snapshot, and *y* in the other. For distracter cells, both snapshots showed distracter parts. Each portion was presented for 2 s., and order of snapshot presentation, and of assignment of features to snapshots were randomized. Note that this procedure works against our hypothesis, because we explicitly expose all subjects to the *x* and the *y* features of *X-Y* cells. Note also that subjects did not know when (first or second snapshot) to expect the feature(s). Hence, it is difficult to argue that subjects would not see, and attend to, the structures present in the two snapshots. Subjects were instructed to wait until they saw the two snapshots to categorize each cell by pressing the appropriate keyboard key (possible responses were *Category_1*, *Category_2* or *none*).

INSERT FIGURE 4 AROUND HERE

Delineation Phase. To confirm which features (or feature combinations) were perceived in the testing phase, we simply instructed subjects to circle on five new cells the parts they *saw* during the experiment. Again, the instructions insisted on the parts

---

[5]Five subjects (2 in *X->XY* and 3 in *XY->X*) were eliminated from the experiment because they made one or more mistakes during the verification phase.

subjects saw, not on those they used to categorize the cells. The delineation cells were 1 *X*, 1 *Y* and 3 distracters presented one at a time, in a random order. We used the same feature delineation technique as explained earlier: Subjects used a computer mouse to circle freely on the new cells the boundaries of each component they thought they saw during the experiment, and there was no lower or upper bound on the number of components subjects could circle.

Results and Discussion.

The groups did not differ in the time required to learn their categories. On average, 1.4 learning blocks were required to learn both *X* and *XY* in *X->XY* and the same averages applied to the learning of the categories in *XY->X*. The two groups' categorization performances were equal (respectively, *X* = 97%, *XY* = 81% for the *X->XY* group and *X* = 94%, *XY* = 81% for the *XY->X* group). We can therefore be confident that subjects knew how to reliably distinguish their categories. These results would lead a fixed feature theorist to conclude that the distinction was effectively implemented with the *y* feature of the basis $F = \{x, y\}$. To verify this conclusion, we now turn to the categorizations of *X-Y* cells. Remember that these cells are *XY* exemplars testing the sequential presentation of the components *x* and *y* (see Figure 4). If $F = \{x, y\}$ was the vocabulary responsible for the perceptual analysis and the categorization of all Martian cell exemplars, *X-Y* cells should always be classified as *XY* members. Differences between the groups on these categorizations should indicate that they used different feature vocabularies.

We scored subjects on whether they categorized *X-Y* cells as *X*, or *XY* exemplars. In the *XY->X* group, *X* categorizations dominated at 84%. In the other group, *X* categorizations dropped at 44%. A t-test on the difference scores of *X* and *XY* categorizations of *X-Y* cells revealed a significant difference between the groups, $t(14) = 2.17$, $p < .05$. That is, *X->XY* subjects preferentially categorized *X-Y* stimuli as *XY* while the other group categorized the same cells as members of *X*, as if they did not perceive the feature *y*. But were subjects really "cognitively blind" to *y*? It was after all defining *XY* category exemplars in both conditions, and *X-Y* test exemplars explicitly isolated *x* and *y* in two different snapshots.

In the delineation phase, instructions explicitly asked subjects to circle the parts they *saw* during the experiment. If subjects did not see a part during the experiment, it is unlikely that they would reliably use it in a high-level, strategic categorization rule. Thus, the delineations should provide confirming evidence that only one group (*X->XY*) perceived *y* as a feature. Subjects were scored on whether they only circled *x* or both *x* and *y* in exemplars of *X* cells and *Y* cells. All but one of the eight subjects of *XY->X* circled only *x*, while five of eight subjects of the *X->XY* group circled both *x* and *y*. A Chi-square test of association revealed a significant association between group and the delineation of *x* or both *x* and *y*, $X^2(1) = 4.267$, $p < .05$. If the *X->XY* subjects often perceived both features, *XY->X* subjects always saw *x*, but almost never *y*.

The fixed feature approach discussed earlier requires that *y* discriminates *XY* from *X*. The data suggest, however, that one group (*XY->X*) reliably distinguished the two categories without apparently perceiving *y*. Thus, it must be that this group used a feature outside the fixed vocabulary $F = \{x, y\}$ to represent the distinction. A plausible hypothesis is that one group perceived and represented *XY* with a unique configural feature *xy*, while the other group analyzed and represented the same category with the conjunction of feature *x* and feature *y*. In sum, it is reasonable to assume that subjects flexibly extracted different structures (specifically, $F^t_{XY\text{->}X} = \{xy, x\}$ and $F^t_{X\text{->}XY} = \{x, y\}$) as a result of learning identical categories.

EXPERIMENT 2

Although damaging for a fixed feature account of category learning, the results of Experiment 1 only rule out a learning mechanism which selects the features of a fixed vocabulary which tend to characterize categories. Experiment 1 does not strictly rule out another interpretation of learning which selects *and* weights fixed components according to their diagnosticity. Starting with $F = \{x, y\}$, subjects learning *X* before *XY* could have learned to "shift more attention" to *y* because this feature was particularly diagnostic of *XY*. Not being *a priori* aware that *XY* had to be distinguished from *X*, subjects who learned *XY* before *X* could have assigned comparatively less attention to *y* --but still just enough to distinguish the categories! Consequently, the two groups could still perceptually analyze the stimuli with identical features, but the critical *y* would in one group "receive more attention" than in the other group. The orthogonal classifications of identical materials observed in Experiment 1 would then reflect differences in selective attention to the elements of an identical perceptual analysis; not a change in the features participating to this analysis.

Such interpretation, however, does not explain the delineation data. The delineation instructions asked subjects to circle the features they saw during the experiment, not those they preferentially used in their categorization rules. Only *X->XY* subjects circled both *x* and *y* , the other group only circled *x*. Delineation data, however, are not free of cognitive influence (Schyns and Murphy, 1994), and there is still a possibility (although remote) that subjects could have misinterpreted the instructions and circled the features they thought were diagnostic of the categories--not those they actually saw. Thus, a more conclusive demonstration that different features participated to the perceptual analysis of identical stimuli might require a procedure which better controls the diagnosticity of individual features.

In fact, we want to demonstrate that the groups have orthogonal perceptions of *XY* (i.e., a holistic *xy* vs. a composition of *x* and *y*), even when the components of *XY* are both diagnostic. To this end, we added the learning of a third category to a design similar to Experiment 1. A *X->Y->XY* group learned three categories in this order, and a *XY->X->Y* group learned the same categories in a different order. This insured that *x* and *y* were respectively diagnostic of *X* and *Y* in both groups. Hence, reliable classifications of *X* and *Y* test stimuli should, without any doubt, indicate that subjects *saw* and *attended* to the components *x* and *y*. Because these features also compose the critical *X-Y* stimuli, we should expect all subjects to categorize them as *XY*, if they all perceived *XY* as a feature conjunction. However, we predicted that only one group (*X->Y->XY*) would perform this categorization. In the other group, the holistic perception of *XY* should prompt *X* or *Y* classifications of *X-Y*. In sum, orthogonal classifications of *X-Y*, when its component features are diagnostic in both groups, should clearly indicate that different features were acquired to analyze and represent *XY*.

Methods

Subjects. 32 University of Montreal subjects participated to the experiment. They were randomly assigned to condition with the constraint that the number of subjects be equal in each group.

Stimuli. Stimuli were the categories of Martian cells used in Experiment 1. Test stimuli were also identical to those used in Experiment 1.

Procedure. The procedure was similar to Experiment 1. In an initial learning phase, subjects were instructed to carefully observe three separately presented categories of 10 cells. Experiment 1 revealed that only 1.4 learning blocks were required to learn each category. In this experiment, we augmented the presentation time of each cell to 2 s. but reduced their learning to only one block. A categorization test (with two category members and two distracters randomly presented) was run after experience with each category (possible responses were *learned_category*, *other*). A perfect categorization criterion was required to start learning the next category, and a single mistake would eliminate the subject from the experiment. In a verification stage, we ran a categorization task to test that the memory of a category was not affected by the subsequently learned categories. The test was composed of two exemplars of each category (*X*, *Y*, *XY*), presented each for 2 s. on a CRT. Subjects were instructed to categorize each exemplar (possible responses were *X*, *Y*, *XY*) by pressing the appropriate key of a computer keyboard. Perfect categorization was required to pursue the experiment[6].

The testing phase presented 16 new cells composed of 4 *XY*, 4 *X*, 4 *Y*, 4 *X-Y*. Again, we used the technique of presenting two consecutive snapshots of the cells (each for 2 s.) through the aperture of a microscope, to (1) ensure that subjects' attention would be drawn to the relevant components of the cells, instead of possibly missing one of the components and (2) to give subjects enough time to grasp the information presented in each snapshot (see Figure 4). As in Experiment 1, *X*, *Y* and *XY* cells were tested with one snapshot presenting the relevant feature while the other snapshot displayed an irrelevant portion of the cell. For *X-Y* cells, the two snapshots presented a different feature. Order of snapshot presentation was randomized in all test items. Subjects were instructed to categorize each cell by pressing the appropriate keyboard key. Possible responses were *Category_1*, *Category_2* and *Category_3*.

Results and Discussion.

The categorization results closely followed those reported in Experiment 1. Subjects classified equally well new exemplars of training categories (respectively, *X* = 98%, *Y* = 81%, *XY* = 92% for the *X->Y->XY* group, and *X* = 98%, *Y* = 88%, *XY* = 92% for the *XY->X->Y* group, see Table 1, under the heading **Empirical data**). This confirms that subjects reliably used their features to analyze and classify the learned items. In particular, note that the identical categorizations of *X* and *Y* in the groups insure that all subjects saw, attended to, and used the features *x* and *y* in their categorization rules. Despite this, the groups differed on their classifications of the *X-Y* cells. In *X->Y->XY*, 88% of these judgments were *XY* categorizations. In *XY->X->Y* , *XY* categorizations of the same cells dropped to 19% (with 54% of *X* and 27% of *Y* categorizations, see Table 1). We see these results as providing conclusive evidence that *XY* was perceived and

[6] Eight subjects evenly distributed between the two conditions were eliminated from the experiment because they made one or more mistakes during the verification phase.

represented as a conjunction of two features in one group, and as a single holistic features in the other group. That is, category learning resulted in the two groups extracting a different feature vocabulary from learning the same three categories. It is our belief that these vocabularies were $F^t_{X\text{->}Y\text{->}XY} = \{x, y\}$ and $F^t_{XY\text{->}X\text{->}Y} = \{xy, x, y\}$[7].

INSERT TABLE 1 AROUND HERE

At the time of writing, it does not seem that any existing fixed feature category learning model would predict the data of Experiment 2. The most likely candidate is probably Gluck and Bower's (1988) configural-cue model which allows the construction of associations between features (basic and configural) and category names, and where order of learning exerts a major influence in the strength of associations that develop. Although this model does not explain how features are learned and developed, it is nonetheless interesting to see whether the fixed feature algorithm best suited to model an order of learning effect would actually predict the orthogonal categorizations of Experiment 2.

Simulation 1: Fixed feature encodings.

The configural-cue model (Gluck & Bower, 1988) is a two-layer network (see Figure 5) where the input layer encodes the input's simple features (here, *x* and *y*) and their pair-wise configural conjunctions (here, only *xy*). The inclusion of configural cues on the input allows the model to solve categorization problems that require sensitivity to simple features *and* their conjunctions (Gluck, 1991). The output layer represents the category names.

---

[7] It is worth mentioning that about a third of the subjects in the *XY->X->Y* condition explicitly told the experimenter that they learned the decomposition of the *XY* stimuli into *x* and *y* during the testing phase, indicating that they were not represented as such before. The other subjects expressed great surprise when they were instructed after the experiment that *XY* stimuli were compositions of the two features characterizing the other categories.

To implement the idea of a fixed encoding with $F = \{x, y, xy\}$, a different input unit recorded the presence (vs. absence) of each feature of F with an activation value of 1 (vs. -1) (see Figure 5, the encoding with fixed features). The first unit encoded x, the second y, and the third the configuration xy. The activation of the three units was then [1 -1 -1] for X, and [-1 1 -1] for Y. The vector [1 1 1] encoded XY, to denote a perceptual encoding of XY with x, y and the configural xy feature (Gluck, 1991). The three different category labels were represented by turning "on" (+1) one of three output units while turning off (-1) the two others.

A standard error-correction mechanism was used to teach the associations between features and category labels (Widrow & Hoff, 1960). Specifically, in vector notation,

$$\Delta \mathbf{W} = \eta \ (\mathbf{t} - \mathbf{o}) \ \mathbf{i}^T \qquad (1)$$

where $\mathbf{t}$, the desired category name (the teacher), is subtracted to $\mathbf{o}$, the category name produced by multiplying the weight matrix $\mathbf{W}$ with the input layer $\mathbf{i}$. If $(\mathbf{t} - \mathbf{o})$, the current categorization error, is different from $\mathbf{0}$, learning changes the weight to correct the error. Hence, recently learned associations will generally have more strength in the associative memory than distal associations (Anderson, 1995). Using the stimulus encoding explained above, one system was trained on learning *X->Y->XY* in this order, and another network was trained on learning *XY->X->Y*. We tested different learning constants η ranging from 0.3 to 0.7, which all produced qualitatively similar results.

Only one exemplar was used per category, as this training regime produces the same result as using ten exemplars and a learning constant divided by the number of exemplars.

INSERT FIGURE 5 AROUND HERE

After learning, we submitted the two categorizers to the four types of test items used in Experiment 2. For X, Y and XY, the featural encoding was the same as explained before. The vector [1 1 -1] encoded *X-Y* to represent the fact that these stimuli contain the feature *x* and the feature *y*, but no configuration *xy* (see Figure 5). We analyzed the network's categorizations by computing the correlations between the produced output **o** and all learned category names **t**. The highest correlation determined the category membership of each test item. Table 1 summarizes the results.

In Table 1, **Network**, Simulation 1, the two columns with heading $F = \{x, y, xy\}$ indicate the fixed vocabulary that encoded the stimuli. The results revealed that the configural-cue network only predicts the subjects' behavior for the categorizations of *X*, *Y* and *XY* stimuli; opposite predictions arose for the critical *X-Y* stimuli--i.e., a *X* categorization of *X-Y* when subjects did categorize this stimulus as *XY*, and *vice-versa.* Thus, the configural-cue network does not account for the empirical data when the assumptions hold (1) that a fixed feature vocabulary $F = \{x, y, xy\}$ served to analyze and perceive the input, (2) that category learning only consists of weighting the diagnostic features of a fixed perceptual analysis[8], and (3) that the orthogonal categorizations reported result from a simple order of learning effect.

Simulation 2: Flexible feature encoding.

Our proposal for flexible features has two different, but related implications. First, we assume that features develop as the organism categorizes its world. Secondly, we claim that the learned features have perceptual status. That is, they impose a perceptual organization of the raw input for classification. In a second set of simulations, we incorporated these two hypothesis in a configural-cue network to study whether it would then predict the empirical data. As explained earlier, the configural-cue model does not learn its features. To incorporate the constraint of flexible vocabularies, two different feature vocabularies (those we believed subjects extracted) were used for the two orders of category learning--i.e., $F_{X\text{->}Y\text{->}XY} = \{x, y\}$ and $F_{XY\text{->}X\text{->}Y} = \{xy, x, y\}$. To incorporate the perceptual status of features, the featural analysis was adapted to the feature vocabularies (see Figure 5, the stimulus encoding with flexible features). Most of the encodings are self-explanatory, but explanations are necessary for *XY* and *X-Y*. To implement the perception of *XY* and *X-Y* as feature conjunctions in *X->Y->XY*, both *XY* and *X-Y* were encoded as [1 1]. To represent the perception of *XY* as a holistic feature in *XY->X->Y*, the [-1 -1 1] encoding was used. *X-Y* was in this condition encoded as [1 1 -1] to represent the perception of *x* and *y*. Note that these encodings are all predicted by our theory of flexible feature learning. The learning rule, learning constant and other simulation details were otherwise identical to those of the previous simulations.

The results are summarized in the two rightmost columns of Table 1, under the heading Simulation 2. A comparison with the two leftmost columns reveals that the model now predicts all aspects of our subjects' categorizations (*X*, *Y*, *XY, and X-Y*). Note in particular that Table 1 shows that an identical $F = \{x, y, xy\}$ was used in the *XY-*

---

[8] A separate simulation with the feature vocabulary fixed to $F = \{x, y\}$ did not account for the data either. In this case, the categorizations of *X->Y->XY* were all accounted for (*X-Y* included), but those of *XY->X-*

>X->Y condition of simulations 1 and 2. However, the network's performance was only compatible with subjects' categorizations when its perception of *XY* was holistic. Thus, it is only when different feature vocabularies changed the perceptual analysis of the same input that the configural-cue network modeled our empirical data.

In sum, Experiment 2 was designed to insure that the components of a fixed feature vocabulary would each be diagnostic of a different category. The similar categorizations of *X*, *Y* and *XY* test items, together with the orthogonal classifications of *X-Y* suggested that the groups used a different feature basis to analyze, perceive and categorize identical stimuli. This interpretation was corroborated by the simulations of a fixed feature model which only predicted our results when its features were flexible and conform to our hypothesis about their role in perceiving and analyzing the input. Consequently, our results suggest that feature creation is here a more plausible explanation than feature weighting.

GENERAL DISCUSSION

Diagnostic features categorize objects--i.e., they distinguish category members from members of contrast categories (Tversky, 1977). The functionality principle (Schyns & Murphy, 1994) suggests that new functional features can sometimes be learned as a consequence of categorizing and representing new objects. Thus, an individual's feature vocabulary can represent the portions of objects that implemented important category similarities or differences in his/her history of categorization.

Contrary to the idea that features are fixed and do not develop with categorization experience, the research presented here suggested that features were flexible, were developed with categorization experience, and did influence the perception of subsequent category exemplars. Specifically, the preliminary experiment showed that different unsupervised pre-exposures to new objects could change the features that were later used to analyze identical test items. Experiment 1 and 2 further demonstrated that the influence

---

>Y were not. The network classified in this case *X-Y* as *XY*, *not* as *X* or *Y*, as did subjects.

of category learning goes beyond a strategic selection, combination and weighting of fixed features: It sometimes triggers the creation of new features which were shown to change the analysis and the perceptual appearance of the stimuli.

Although the “existence proof” of feature creation has only been established with artificial and unfamiliar Martian cell categories (and Martian rocks, see Schyns & Murphy, 1994), we wish to remain agnostic about the extension of feature creation principles to familiar object categories (e.g., dogs, chairs, cars and so forth). Future research may reveal that feature creation only applies to the learning of very specialized (or subordinate) categories such as X-rays, dermatosis and so forth. Alternatively, it may turn out that feature creation mechanisms pervade the very early stages of conceptual development (when the first categories and their structuring features must be learned), but that mature categorizers (who tend to know the relevant perceptual analysis of most objects) only create new features when they learn expert categorizations. Whether feature creation mechanisms have a broad, or a limited impact on the development of familiar object concepts is now an empirical issue of developmental psychology.

Implications for feature spaces, similarity and comparison-based learning. There is a similarity between Schyns and Murphy’s (1991; 1994) conception of a functional feature space and Wisniewski and Medin’s (1991; 1994) idea that theories and sophisticated knowledge about objects sometimes synthesize new descriptive features. Both proposals are analogous in their suggestion that concept learning goes beyond the mere selection and weighting of fixed features: It sometimes augments the description space with new features. The proposals differ on the origin and the nature of these flexible features. Wisniewski and Medin’s new features originate in the theories and sophisticated knowledge people apply to interpret the outside world (see also Murphy & Medin, 1985). For example, in one of their experiments, subjects viewed a series of children’s drawings and were told either that they were drawn by creative children or urban children. In one case, a series of dots on the shirt of a figure was interpreted as "buttons," revealing a "detailed" drawing of a creative child by one subject, but was interpreted by another subject as a "tie" revealing "city clothing" drawn by an urban child. That is, subjects' knowledge about the source of the drawing drove the inference of new descriptive features.

The important difference between this and our approach is that we have emphasized the perceptual changes that accompany feature creation. In Wisniewski and Medin's experiments, when subjects interpret a series of dots as "buttons," or as "a tie," there is little reason to suggest that the inferred features alter the immediate appearance of the drawings. Rather, they probably alter the properties that will be inferred about the drawings. There is a difference in immediacy between conceiving and seeing that an object might be expressed in terms of a particular feature (Schyns et al., 1995). The experiments reported in this paper suggest that the immediate appearance (the perceptual analysis) of objects can be altered by the history of categorization. Thus, we see Wisniewski and Medin's and our proposal for flexible perceptual features as complementary accounts of flexible feature spaces. Together, theory-derived and perceptually-derived flexible features offer powerful principles to augment the vocabulary of object descriptions with new features. These principles should expand the scope of similarity and categorization theories beyond their current borders (see Thibaut & Schyns, 1995).

Research on flexible feature spaces has important implications for research on similarity. By definition, a functional feature is created to represent (at least) one new category (or object) similarity or distinction. Thus, the feature vocabulary specifies the ways in which two objects may be similar or different. Functional recoding therefore transforms the raw input using features which, by definition of functionality, are important to the individual's representational needs. This emphasis on the individual's history of categorization questions the existence of universals of similarity and the notion of an objective interpretation of similarity. As the feature vocabulary evolves dynamically in response to the classifications facing the individual, the individual's history of categorization could be more important to understand similarity judgments than is the raw appearance of the compared objects. In fact, our experiments showed that minor differences in history of categorization resulted in contrasted perceptions of identical materials.

Most formal models of similarity implement differences in similarity judgments with different (nonlinear) weightings of identical features (see, among others, Nosofsky, 1984; Smith, 1989; Shepard, 1987; Tversky, 1977). Changes in perceptual similarity is therefore quantitative and it results from a higher weighting of the features that are more diagnostic or important in the task. Flexible feature sets suggest another, more qualitative conception of perceptual similarity and perceptual change. The addition of new perceptual and other features to the vocabulary changes the qualitative description of the input. Of course, both conceptions of perceptual change and perceptual similarity (quantitative and qualitative) are necessary for a complete understanding of similarity judgments.

Medin, Goldstone and Gentner (1993) also discuss how similarity comparisons alter the qualitative featural encoding of an object. Although many of their experiments do not isolate perceptual changes *per se*, a few of their experimental situations are closer in spirit to our approach. In one particular experiment, subjects were asked to list the features two objects *X* and *Y* had in common. It was shown that the features listed for *X* were different when *X* was compared to *Y* than when *X* was compared to *Z*. Importantly, subjects sometimes used mutually exclusive features to describe the *X* objects in the two conditions--e.g., object *X* was perceived as having three vs. four prongs, or as being two- vs. three-dimensional (see also Hochberg & Brooks, 1960).

Our experiments and those of Schyns and Murphy (1991, 1994) also revealed such mutually exclusive perceptions, in the general case of category learning. For example, in experiments 1 and 2, different subjects perceived the same *XY* category as a features conjunction, or as a configuration depending on their categorization experience.

The interaction between the raw input and categorization experience determined these perceptions--not the object's attributes alone. Our account, however, is grounded on the history of the categorizer. Hence, we can predict some of the circumstances under which the featural description of an object will use this subset of functional features, or will require the creation of new features.

Spalding and Ross (1994) also studied the importance of the history of categorization on the subsequent description of category exemplars. Spalding and Ross argue (1994, p. 1252) and present evidence that what is learned as a result of comparing exemplars of a category is "... selective, in that the aspects common to the compared instances are considered more important than aspects that are not common to the compared instances. That is, learners *focus* [our emphasis] on these common features ... as important aspects of the category." Our experiments also revealed that what was learned early in categorization history could serve as "focus" in subsequent category learning. For example, after creating feature *x,* subjects in the *X->XY* condition of Experiment 1 (and the *X->Y->XY* condition of Experiment 2) initially focused on this feature when they analyzed new *XY* exemplars. In fact, a case could be made that a focusing on *x* was responsible for the perceptual analysis of *XY* exemplars as feature conjunctions. Thus, our approach share with Spalding and Ross (1994) the idea that it is important to understand how very early categorizations affect the featural descriptions of exemplars that are learned later (see also Ross, in press). However, we have different conceptions of the nature of representational changes. While we argued for, and demonstrated the necessity of, flexible perceptual features to explain some of the changes that accompany category learning, Ross explained representational differences in terms of attentional weighting of fixed features. As already argued, attentional weighting does not account for our data.

Implications for formal concept learning models. Attentional weighting of fixed features is the basic processing metaphor underlying most concept learning models.

These models have successfully demonstrated that category representations can be learned from exemplars described with a small, pre-specified feature set (e.g., Gluck & Bower, 1988; Krushke, 1992; Rumelhart, Williams & Hinton, 1986; Widrow & Hoff, 1960). Although these mechanisms do not discover new perceptual features in high-dimensional data, it could be argued that such learning would occur if standard mechanisms were scaled up to operate in high-dimensional spaces.

The *curse of dimensionality* (Bellman, 1961) and the *bias/variance dilemma* (Geman, Bienenstock & Doursat, 1994) impose severe limitations to this idea. Error-correction models (e.g., Kruschke, 1992) are nonparametric statistical models (Geman et al., 1994). They are nonparametric because they are not biased to particular classes of solutions (e.g., a linear or quadratic decision surface to separate the categories). Instead, the architectures are unbiased so as to flexibly learn new structures from data. Geman et al. (1994, p. 9-10) demonstrated that the miscategorization error of error-correction networks can be algebraically decomposed into a bias and a variance term. Networks make a bias error when they are “too fixed” (when they are dedicated to a class of solutions that is inappropriate for the considered categorizations). Flexibility (low bias), however, comes at the cost of high variance, the second term of the error (a flexible system is too sensitive to the data and therefore learns many idiosyncrasies of the exemplars before learning the invariants of a category). Consequently, experience with implausibly large data sets is necessary for the network to “forget” idiosyncrasies and learn robust features.

The bias/variance dilemma has important consequences for plausible models of feature creation and categorization. A “brute force” approach assuming little or no perceptual bias on the input representation (e.g., pixels or line segments) is not a promising research avenue because it does not sufficiently constrain the parametric model and consequently imposes enormous data sets to discover category-relevant perceptual invariants[9]. Instead, the dilemma suggests that different aspects of perceptual learning will be best modeled with very specific biases; biases that do not trivialize the perceptual changes that need to take place in the system, but biases that are sufficiently constraining to allow the discovery of new and general features from a reasonably small data sets (see also Anderson & Rosenfeld, 1988; Geman et al., 1992; Shepard, 1989, among others). The nature of these biases and their interaction with categorization influences is a promising--if very difficult--avenue of future research in category learning and conceptual development.

The quest for the quanta of cognition. There is always the question of whether the functional features we propose really are the minimal units of a theory of object categorization. This is a very difficult question. Good hygiene of theory construction recommends that the analysis of the goals and purposes of a recognition task precedes the study of its representations and algorithms (Marr, 1982). Thus, the categorizations a feature vocabulary must solve determine (at least partially) the nature of the primitives entering the vocabulary. To illustrate, in Fisher's (1986) influential model theory of letter recognition (cited in Czerwinski, Lightfoot & Shiffrin, 1992), the identification of three primitives (two diagonal bars and a horizontal bar) precedes the recognition of the capital letter "A". Clearly, diagonal and horizontal bars were selected as primitives with the task of letter categorization in mind; the same primitives would be particularly inadequate to categorize classes of ellipses.

In componential theories, primitives provide the initial discretization--the perceptual organization--of the input for recognition. Our functional theory is also componential, and we also assume that people initially discretize the input with task-dependent features. However, it could be argued that functional features are higher-level

---

[9] This is mostly due to the fact that the many facets of a high-dimensional input space need to be explored before one is sure that the extracted features are the “right ones”--i.e., those that will all generalization to new category exemplars. This curse of dimensionality does not affect as much fixed feature models which usually operate in smaller spaces for which sufficient exemplars can sometimes be generated.

constructs, and that the “real” features that analyze the input are lower-level constructs. In a sense, this must be trivially true, because the cones and rods of the retina provide the initial encoding of any visual signal. However, we argued earlier that an effective recognition system will not operate directly on these high-dimensional encoding, but on a low-dimensional recoding. Our research suggest that these low-dimensional perceptual recodings are driven (at least partially) by categorization constraints. It is an important avenue of future research to understand how categorization and perception interact to produce low-dimensional basis of categorization features (see, e.g., Oliva & Schyns, 1995; Schyns & Oliva, 1994; Schyns et al., 1995).

In summary, many componential accounts of recognition assume that the primitives of stimulus design form the low-dimensional primitives of input analysis. We would like to argue that these primitives are as functional as are our features: A component line segment in letter recognition may be as low- (or as high-) level a primitive as a Martian cell feature. The *raison d'être* of the two features can only be understood in the context of a particular task. In one case, the feature results from the experimenter's understanding of letter recognition; in the other case, the feature results from subjects' understanding of complex object categories. The quest for the quanta of cognition may redefine the traditional boundaries between perception and categorization, and between structural and functional theories.

CONCLUSIONS

The origin of feature has been largely neglected in categorization theories. One possible stance is that features are fixed and independent of experience with objects categories. Alternatively, features could be flexible and gradually develop with the functional requirements of categorizing and representing objects. The experiments reported here tested that the individual’s history of categorization could trigger the development of new features that changed the perceptual appearance (the perceptual analysis) of stimuli. Results demonstrated that new features were learned resulting in mutually exclusive featural analysis and categorizations of identical materials. Hence, complete theories of conceptual development will not only have to explain how features are grouped to form concepts; they will also have to include principles of feature creation to explain what these features are.

REFERENCES

Anderson, J. A. (1995). *Neurocomputing.* MIT Press: Cambridge, MA.

Anderson, J. A., & Rosenfeld, E. (1988). *Neurocomputing, Foundations of Research*. MIT Press: Cambridge, MA.

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

Biederman, I. (1987). Recognition-by-components: A theory of image understanding. *Psychological Review*, **94**, 115-147.

Biederman, I. & Gerhardstein, P.C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception &Performances*, **21**, 1506-1514.

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. Journal of Experimental Psychology: Learning , *Memory and Cognition*, **13**, 640-645.

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

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

Christensen, E. E., Murry, R. C., Holland, K., Reynolds, J., Landay, M. J., & Moore, J. G. (1981). The effect of search time on perception. *Radiology*, **138**, 361-365.

Czerwinski, M., Lightfoot, N., & Shiffrin, R. M. (1992). Automatization and training in visual search. *American Journal of Psychology*, **105**, 271-315.

Elio, R. & Anderson, J. R. (1981). The effect of category generalizations and instance similarity on schema abstraction. Journal of Experimental Psychology: *Human Learning & Memory*, **7**, 397-417.

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

Fisher, D. L. (1986). *Hierarchical models of visual search: Serial and parallel processing*. Paper presented at the meeting of the Society for Mathematical Psychology, Cambridge, MA.

Geman, S., Bienestock, E., & Doursat, R. (1994). Neural networks and the bias/variance dilemma. *Neural Computation*, **4**, 1-58.

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

Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology*, **49**, 239-242.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, **2**, 50-55.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, **27**, 166-195.

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

Harnad, S. (1990). The symbol-grounding problem. *Physica D*, **42**, 335-346.

Hochberg, J., & Brooks, V. (1960). The psychophysics of form: reversible-perceptive drawings of spatial objects. *American Journal of Psychology*, **73**, 337-354.

Hoffman, D., & Richards, W. (1985). Parts in recognition. *Cognition,* **18**, 65-96.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search, *Radiology*, **116**, 527-532.

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

Lesgold, A. M. (1984). Acquiring expertise. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower* (pp. 31-60). San Francisco: W. H. Freeman.

Lesgold, A. M., Glaser, R., Rubinson, H., Klopfer, D., Feltovitch, P. & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser, M. J. Farr (Eds.), *The nature of expertise*. Hilldsale, NJ: Erlbaum.

Lockhead, G. R. (1966). Effects of dimensional redundancy on visual discrimination. *Journal of Experimental Psychology*, **72**, 95-104.

Markman, E. (1989). *Categorization and naming in children.* Cambridge: MIT Press.

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

Marr, D., & Nishihara, K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B*, **200**, 269-294.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

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

Murphy, G. L. & Smith, E. E. (1980). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, **21**, 1-20.

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 553-557.

Norman, G. R., Brooks, L. R., Coblentz, C. L., & Babcock, C. J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Memory & Cogintion*, **20**, 344-355.

Nosofsky, R. M. (1984). Choice, similarity, and the context of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 104-114.

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

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

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and an extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, **22**, 352-369.

Oliva, A., & Schyns, P. G. (1995). Mandatory scale perception promotes flexible scene categorization. *Proceedings of the XVII Meeting of the Cognitive Science Society*.

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

Ross, B. H. (in press). Category representations and the effects of interacting with instances. *Journal of Experimental Psycholgy: Learning, Memory & Cognition.*

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

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1995). *The development of features in object concepts*. Indiana University Cognitive Science Research Report 133, Indiana University, Bloomington, IN.

Schyns, P.G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychological Science*, **5**, 195-200.

Schyns, P. G., & Murphy, G. L. (1991). The ontogeny of units in object categories. *Proceeding of the XIII Meeting of the Cognitive Science Society*, 197-202, Lawrence Erlbaum: Hilldsale, NJ.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. In Medin (Ed.), *The Psychology of Learning and Motivation,* **31**,301-349.

Shepard, T. W. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

Shepard, R. N. (1989). Internal representations of universal regularities: A challenge for connectionism. In L. Nadel, L. A. Cooper, P. Culicover, and R. M. Harnish (Eds.), *Neural Connections, Mental Computations.* Bradford: Cambridge, London.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: a case study of the development of the concepts of size, weight and density. *Cognition*, **21**, 177-237.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, **96**, 125-144.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, **10**, 502-532.

Spalding, T. L., & Ross, B. H. (1994). Compariison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition,* **20**, 1251-1263.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectations and predictions. *Psychological Review*, **88**, 135-170.

Thibaut, J. P. (1991). *Récurrence et variations des attributs dans la formation de concepts*. Unpublished doctoral thesis. Department of Psychology. University of Liège, Belgium.

Thibaut, J. P., & Schyns, P. G. (in press). The development of feature spaces for similarity and categorization. *Psychologica Belgica.*

Treisman, A. , & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

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

Ward, T. B. (1983). Response tempo and separable-integral responding: evidence for an integral-to-separable processing sequencing in visual perception. *Journal of Experimental Psychology : Human Perception and Performance*, **9**, 103-112.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 908-923.

Widrow, G. & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of radio engineers, western electronic show and convention, convention records,* **4**, 96-194.

Wisniewski, E. J., & Medin, D. L. (1991). Harpoons and long sticks: The interaction of theory and similarity in rule induction. In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), *Concept formation: Knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufman.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, **18**, 221-283.

FOOTNOTES

Title note: This work was partly supported by an NSERC grant to the first author. The authors wish to thank Robert Goldstone, John Hummel, Serge Larochelle, Barbara Malt Gregory Murphy, Robert Nosofsky, Linda Smith, and Jean-Pierre Thibaut for many stimulating discussions and comments. A special thank to Annie Archambaut and Aude Oliva for helping us carrying out the experiments. Correspondence regarding this article should be sent to Philippe G. Schyns, Dept. of Psychology, University of Glasgow, Glasgow G12 8QB, United Kingdom. philippe@psy.gla.ac.uk

Table1: Table 1 presents the empirical data and the network model predictions of Experiment 2. Under the heading **Empirical data**, the categorizations of each type of test item is reported for each experimental group. Under the heading **Network**, simulation 1 shows the prediction of Gluck and Bower's (1988) configural-cue model using a fixed feature encoding of the exemplars. Simulation 2 predicts the empirical data with the same model using a flexible feature encoding of the input. In both simulations, the learning constant of the error-correction algorithm was set to 0.5.

	Empirical data		**Network**			
			Simulation 1		Simulation 2	
			$F = \{x, y, xy\}$	$F = \{x, y, xy\}$	$F = \{x, y\}$	$F = \{x, y, xy\}$
	X->Y->XY	*XY->X->Y*	*X->Y->XY*	*XY->X->Y*	*X->Y->XY*	*XY->X->Y*
X	98%	98%	.9	.96	.68	.96
Y	81%	88%	.96	.97	.9	.97
X Y	92%	92%	.97	.9	.98	.9
X - Y	88% ***X Y***	19% *XY*	-.13 *XY*	.75 ***X Y***	.98 ***X Y***	-.9 *XY*
		54% ***X***	.75 ***X***	.22 *X*	-.45 *X*	.61 ***X***
		27% *Y*	.22 *Y*	-.13 *Y*	-.35 *Y*	.37 *Y*

FIGURE CAPTIONS

Figure 1. The top picture shows an example of the Martian cells used in our experiments. The abstract shape underneath is the basis of our experiments; the other components of the cells are what we call "random cell bodies." In the experiments, random cell bodies were different on each exemplar of Martian cell. The numbered arrows indicate the possible parse points predicted by applying the popular perceptual segmentation explained in Hoffman and Richards (1985). The bottom left pictures indicate the segmentation of the middle picture into the *a* and *bc* features (obtained by connecting the 2 and 7 parse points). The bottom right pictures show the segmentation of the middle shape into the *ab* and *c* features (obtained by connecting the 3 and 6 parse points).

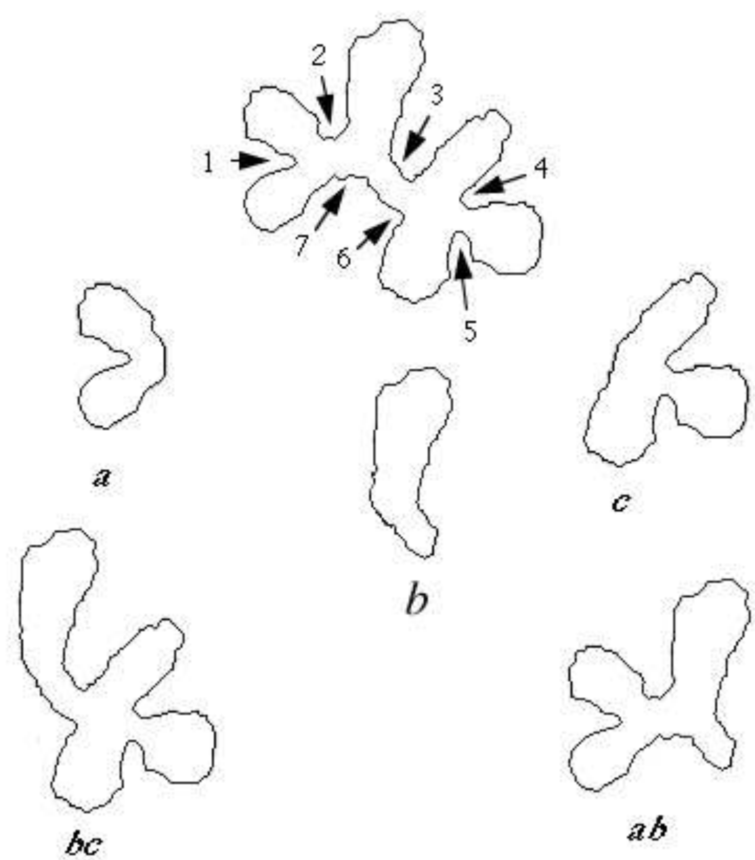
Figure 2. Figure 2 illustrates the Martian cells used in the testing phase of the preliminary experiment. The simple *A* and *C* cells are respectively defined by the *a* and *c* components. Composed *AB*, *BC*, and *ABC* cells are defined by various combinations of the *a*, *b* and *c* components.

Figure 3. Figure 3 summarizes the learning stage of Experiment 1. *XY->X* subjects saw *XY* stimuli before *X* stimuli, and the inverse was true of *X-XY* subjects. The three category defining components are illustrated at the bottom of the figure. The four exemplars of cells presented in Figure 3 show the category defining components plus the random cell bodies.

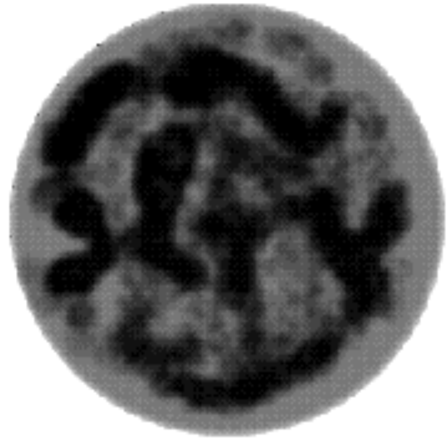
Figure 4. This figure illustrates the procedure used in the testing phase of experiments 1 and 2. Subjects saw two successively presented portions of Martian cells (as if each cell was observed through two snapshots of a microscope). *X*, *Y* and *XY* exemplars each consisted of a snapshot containing the relevant feature and a snapshot containing an irrelevant shape. Note that *X-Y* test items formally are members of *XY* (each portion contains a feature), but that the features are separately presented in different portions. In the testing phase, order of assignment of feature to portion was randomized. For example a *X-Y* item could either start with the presentation of *x*, or the presentation of *y*.

Figure 5. This figure summarizes Gluck's (1991) configural-cue model used to model the subjects' categorizations in Experiment 2. Each input unit encodes the presence (1) vs. absence (-1) of a different feature (*x*, *y* and *xy*) in the input stimulus. The fixed and flexible modes of encoding are represented by different activation values for each of the *X*, *Y*, *XY* and *X-Y* exemplars. In the fixed feature encoding, the same feature set *F = {x, y, xy}* was used to encode the stimuli in *X->Y->XY* and *XY->X->Y*. In the flexible encoded, a different vocabulary was used for each order of category learning.

Figure 6. This figure summarizes the *X->XY* and the *X->Y* initial learning conditions of Experiment 3. An exemplar of the *XYZ* category experienced in the second learning phase is also provided. Note the difference in features between the *XY* and the *XYZ* category.


1
2
3
4
5
6
7
a
b
c
bc
ab

ABC

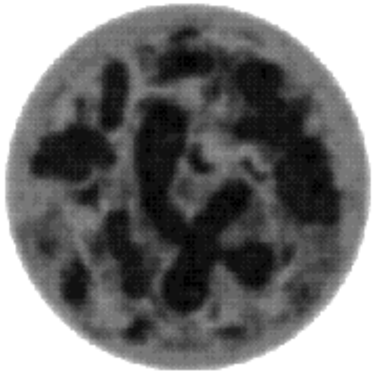


AB

C

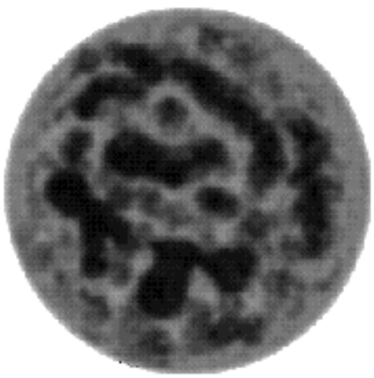
A

BC



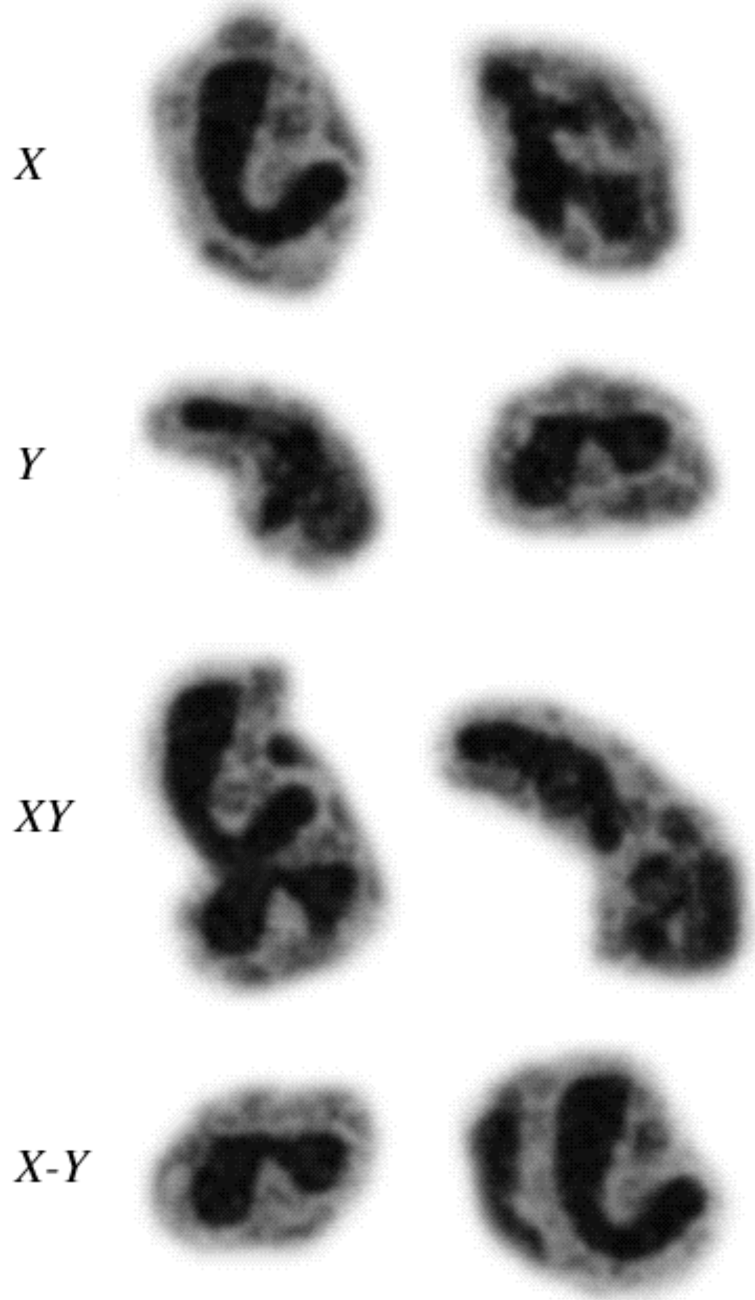
XY

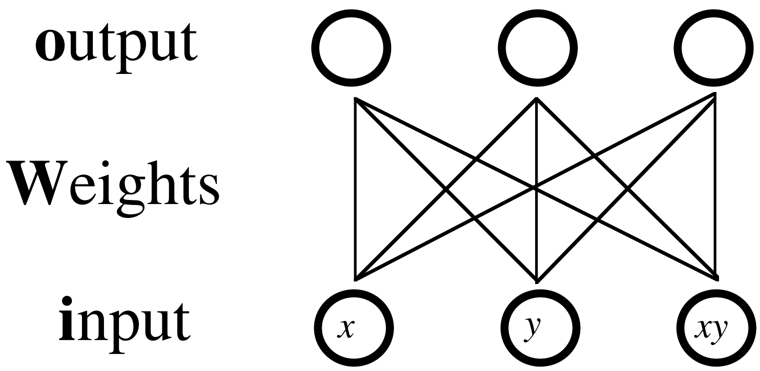
X



X

XY


X
Y
XY
X-Y



Stimulus encoding

Fixed features

$F = \{x, y, xy\}$

X	1	-1	-1
Y	-1	1	-1
XY	1	1	1
$X\text{-}Y$	1	1	-1

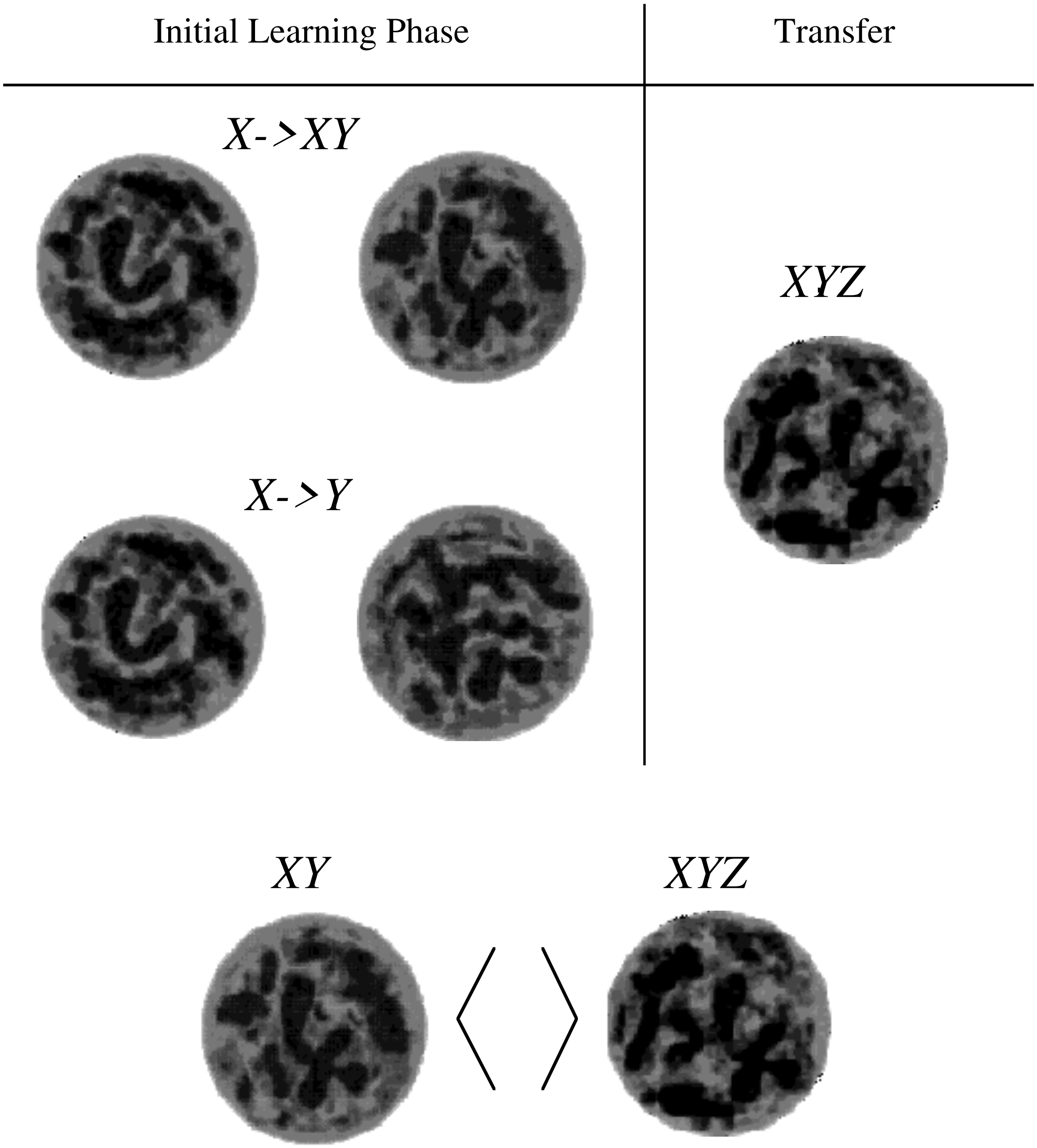
Flexible perceptual features

$F_{XY\text{->}X\text{->}Y} = \{x, y, xy\}$

X	1	-1	-1
Y	-1	-1	-1
XY	-1	-1	1
$X\text{-}Y$	1	1	-1

$F_{X\text{->}Y\text{->}XY} = \{x, y\}$

X	1	-1
Y	-1	1
XY	1	1
$X\text{-}Y$	1	1


Initial Learning Phase
Transfer
X->XY
X->Y
XYZ
XY
XYZ