

Representational Shifts during Category Acquisition: A Preference for Features that Provide  
Information for Multiple Categories

Michael Fink  
Interdisciplinary Center for Neural Computation  
The Hebrew University of Jerusalem  
Jerusalem 91904, Israel  
fink@huji.ac.il

Gershon Ben-Shakhar  
Department of Psychology  
The Hebrew University of Jerusalem  
Jerusalem 91905, Israel  
mskpugb@pluto.msc.huji.ac.il

Shimon Ullman  
Faculty of Mathematics and Computer Science  
Weizmann Institute of Science  
Rehovot 76100, Israel  
shimon.ullman@weizmann.ac.il

May 29, 2005

## Abstract

The present study was aimed at examining the factors inducing feature creation in the perceptual and semantic systems. The proposed research was guided by the hypothesis that due to the substantial load of categorization tasks, a preference exists for encoding features that are informative for the recognition of multiple categories. This claim extends theories stipulating that features are encoded for each categorization task individually and challenges theories that focus solely on encoding coincidences in input patterns. The research hypotheses were tested in a series of experiments, requiring participants to learn several new categories in controlled conditions. Experiment 1 demonstrated that while acquiring a sequence of four novel perceptual categories, features that are informative for recognizing several categories are created. This was observed despite an absence of suspicious coincidences in the input patterns. Moreover, Experiment 1 provided evidence that after acquiring novel features, the perceptual system actively reconstructs the representations of categories learned in the past so that they comply with the newly encoded features. Experiment 2 demonstrated that learning an additional, fifth, perceptual category can be significantly facilitated, if it is congruent with the previously encoded feature structure. Finally, similar results were observed in Experiments 3 and 4, that used semantic stimuli (job candidate characteristics) instead of the perceptual stimuli (color cubes configurations) used in the first two experiments. It is therefore suggested that encoding features that are informative for multiple categories may be a general principle, underlying the feature creation processes in both the perceptual and the semantic systems.

## Introduction

The dynamic environment in which humans live requires fast and accurate recognition of numerous categories. These requirements must often be met after only few examples of a novel category have been encountered. Yet, humans operate quite well under these conditions, recognizing a substantial number of categories (Biederman, 1987) with unparalleled speed and accuracy (Thorpe, Fize, & Marlot, 1996). It remains a puzzle how can such a categorization load be efficiently managed and how might the hindering effects of over fitting be avoided when learning from a small sample (Goodman, 1972).

Theories of categorization are typically based on measuring the similarity of an incoming stimuli to an internal representation of the category, be it a prototype or a set of exemplars (Medin & Smith, 1984; Barsalou, 1985; Nosofsky, 1988). The tradition set by Tversky (1977) emphasized the role of having the appropriate features for measuring similarity, yet remains vague as to what is the criterion for encoding the appropriate features. This research postulates that the categorization load might be managed, only by encoding a representation that is highly tuned to the specific categorization requirements posed by the environment. Specifically, it is proposed that large scale categorization systems must actively encode features that are informative for recognizing a large number of categories. The proposed research hypothesis follows the feature creation theory postulated by Schyns, Goldstone, and Thibaut (1998), in rejecting fixed feature sets, like the geon theory suggested by Biederman (1987). In fact it seems that the perceptual system readily initiates novel feature creation even when required to perform simple similarity judgments (Spencer-Smith & Goldstone, 1997). Two theoretical frameworks have suggested a computational criterion as to when a feature set should be dynamically extended. The first theoretical framework, proposed by Barlow (1989), emphasizes the importance of features that encode suspicious coincidences between input patterns (say  $x_i$  and  $x_j$ ). Formally, this theory encodes events where:

$$\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \gg 1 .$$

Indeed, it is by now a well established fact that features sensitive to input statistics might be created in the absence of any categorization feedback (Rosenthal, Fusi, & Hochstein, 2001; Edelman, Hiles, Yang, & Intrator, 2001; Fiser & Aslin, 2002). In contrast to this input driven process, the second theoretical framework stipulates that the fundamental criterion of feature creation is maximum mutual information provided by a feature  $F^*$  in determining whether a target category  $C$  is present ( $C = 1$ ) or absent ( $C = -1$ ) (see (Ullman, Vidal-Naquet, & Sali, 2002)). When considering a set  $\mathcal{F}$  composed of binary features, this criterion could be formally presented as:

$$F^* = \operatorname{argmax}_{F \in \mathcal{F}} (I(F; C)) = \operatorname{argmax}_{F \in \mathcal{F}} - \sum_{\substack{F \in \{-1, +1\} \\ C \in \{-1, +1\}}} p(F, C) \log \frac{p(F, C)}{p(F)p(C)} = \operatorname{argmax}_{F \in \mathcal{F}} (H(C) - H(C|F)) \quad (1)$$

where  $H(C)$ , is the uncertainty whether category  $C$  is present and  $H(C|F)$  is the uncertainty about  $C$  given the presence or absence of feature  $F$ . Several empirical findings provide indirect evidence for this theoretical framework. Goldstone (1994) has demonstrated that different categorization tasks induce selective feature sensitization. Similarly, Archambault, O'Donnell, and Schyns (1999) show that general vs. specific categorization tasks might influence the perceived properties of the same distal object. More direct evidence for feature creation induced by a categorization task, was provided by Schyns and Murphy (1994) and by Schyns and Rodet (1997). Common to these different lines of research is the fact that at each training stage, the contribution of a feature to encoding a single target category is estimated. As previously described, categorization systems must be capable of recognizing many thousands of categories (Biederman, 1987), often consolidating many categories in a common time period. It therefore seems that examining the setting where feature creation is induced by an individual category does not capture the essence of the load constraints in which perceptual and semantic systems must operate. Therefore, the proposed research is aimed at augmenting the two theories stated above by examining whether novel features might be encoded in the absence of suspicious coincidences in the input statistics and despite the fact that they do not maximize the information provided for any individual target category.

This research emphasizes the fact that typically numerous categorization tasks ( $n > 1$ ) are posed by natural environments. Thus, stating that in order to maintain speed, accuracy and efficient generalization in large scale catego-

rization systems, a common set of category informative features must be actively preferred. Thus, Eq. (1) is modified by summing  $C$  over all multiple category assignments  $C \in \{-1, +1\}^n$  rather than for the presence or absence of a single category  $C \in \{-1, +1\}$ :

$$F^* = \operatorname{argmax}_{F \in \mathcal{F}} - \sum_{\substack{F \in \{-1, +1\} \\ C \in \{-1, +1\}^n}} p(F, C) \log \frac{p(F, C)}{p(F)p(C)}. \quad (2)$$

It was predicted that this principal would have the three following manifestations:

1. preferable encoding of novel features that are informative for recognizing many categories
2. encoding novel features might entail reconstructing former category representations
3. once encoded, novel features can facilitate acquisition of future categories

Four experiments were designed to test these three predictions. Experiment 1 examined whether the first two predictions indeed characterize the perceptual system and Experiment 2 validated the third prediction. Both experiments utilized controlled sets of perceptual stimuli (configurations of colored cubes). In order to assess whether the computational principal proposed in Eq. (2) might be manifested in higher levels of the perceptual-conceptual continuum (Goldstone & Barsalou, 1998), Experiments 3 and 4 implemented the same empirical setting as in Experiments 1 and 2, utilizing semantic stimuli (job candidates' descriptions).

## Experiment 1: Encoding Perceptual Features Informative for Multiple Categories

Experiment 1 was designed to examine whether the perceptual system preferably encodes features that are informative for recognizing many categories and whether this process might actively reconstruct the representations of former categories acquired in the past.

### Method

The experimental setting was based on a detection task consisting of eight input elements  $x_{i,i=1,\dots,8}$ , jointly notated as  $\mathbf{x}$ , and four output elements  $t_{i,i=1,\dots,4}$ , notated as  $\mathbf{t}$ . Both input and output elements were binary, being either in an *on* (+1) or an *off* (-1) state. Four target categories  $C_{n,n=1,\dots,4}$  were defined as a function from the input vector set  $\{\mathbf{x}\}$  to the output category vector set  $\{\mathbf{t}\}$ . Each category was associated with a single output element and was fully characterized by four specific input elements being in an *on* position detailed in Table 1. As shown in Table 1, categories  $C_1$  and  $C_2$  required two common input elements ( $x_3$  and  $x_4$ ) to be in an *on* state. This common *pair-feature*, was termed  $v_2$ . In fact, each of the four categories shared a common *pair-feature* with two other categories (see Fig. 1). These four *pair-features* were formally defined as:

- $v_1 \equiv (x_1 = 1) \cap (x_2 = 1)$
- $v_2 \equiv (x_3 = 1) \cap (x_4 = 1)$
- $v_3 \equiv (x_5 = 1) \cap (x_6 = 1)$
- $v_4 \equiv (x_7 = 1) \cap (x_8 = 1)$

It will now be shown that the proposed category structure enables a direct validation of whether the perceptual system preferably encodes features that are informative for recognizing many categories. As stated above the target

Input Elements	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
Category $C_1$	1	1	1	1	*	*	*	*
Category $C_2$	*	*	1	1	1	1	*	*
Category $C_3$	*	*	*	*	1	1	1	1
Category $C_4$	1	1	*	*	*	*	1	1
Category $C_5$	*	*	1	1	*	*	1	1

Table 1: Definitions of the four categories learned in Experiment 1, and of the fifth category learned in Experiment 2. Input elements indicated by 1 must be in an *on* state for the category to be present, while \* denotes the category’s indifference to certain input elements.

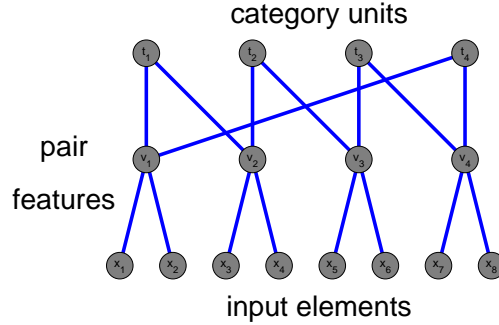


Figure 1: *Pair-features*  $v_1$  through  $v_4$  can provide an efficient solution for the categorization task.

categories were defined as fourth order conjunctions over the input elements. Therefore, the candidate features set, functioning as an intermediate representation, was defined to include all input element conjunctions of the 1st, 2nd and 3rd degrees (a total of  $\binom{8}{1} + \binom{8}{2} + \binom{8}{3} = 92$  features). The mutual information between each of the candidate features and the target categories was evaluated using Eq. (2). Although each of the *pair-features* contains little information (0.008 or 0.264 bits) for any individual category, the *pair-features* are highly informative (0.581 bits) for the four target categories collectively (see Fig. 2). Thus, the first research prediction is manifested if a salient representation of the *pair-features*:  $v_1, v_2, v_3$  and  $v_4$  would emerge while acquiring the four categories.

The proposed setting must control the alternative factors influencing feature creation previously described in the introduction section. This goal was achieved by finding an alternative representation that is comparable in all aspects to the *pair-feature* representation but is not as informative for the categorization tasks. Such an alternative representation is generated by arbitrarily segmenting the eight input elements into four *incongruent-pairs*, that appear each in just one category. For example, one such arbitrary segmentation of *incongruent-pairs* is:

- $h_1 \equiv (x_2 = 1) \cap (x_8 = 1)$
- $h_2 \equiv (x_4 = 1) \cap (x_6 = 1)$
- $h_3 \equiv (x_3 = 1) \cap (x_5 = 1)$
- $h_4 \equiv (x_1 = 1) \cap (x_7 = 1)$

The proposed methodology establishes the existence of the *pair-features* through a systematic comparison to the a-priori similar *incongruent-pairs*. The *pair-features* and *incongruent-pairs* representations are viewed as a-priori similar since they are both arbitrary parsings of the input elements into four pairs.

### Participants

Twenty undergraduate students from the Hebrew University of Jerusalem participated in Experiment 1 for payment or course credit. Participants, aged 20 to 30 years, were screened for normal color vision.

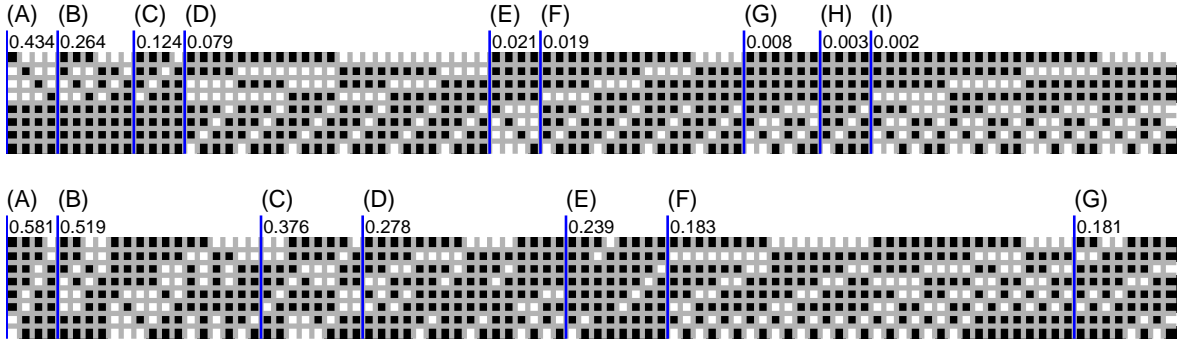


Figure 2: The candidate feature set includes all 1st, 2ed and 3ed order conjunctions over the eight input elements. Each column represents a single feature by highlighting the conjunction elements (white). Features are grouped by information content on category  $C_1$  (top) and by information content on the four categories collectively (bottom). Although the pair features are not highly informative for category  $C_1$  individually (appearing in groups (B) and (G) in the upper pane), they appear as the maximally informative features for the four categories collectively (group (A) in the lower pane).

### Materials

To test whether the perceptual system complies with the first research prediction the input and output vectors sets, ( $\{\mathbf{x}\}$  and  $\{\mathbf{t}\}$ ), were implemented into a visual detection task. The eight-dimensional binary inputs were translated into images composed of eight enumerated *color cubes*  $x_{i,i=1,\dots,8}$ . For each cube, one color was selected to function as the *on* state ( $x_i = +1$ ) and another color as the *off* state ( $x_i = -1$ )<sup>1</sup>. For each individual participant, the set of 16 colors was randomly allocated to the eight cubes (see Fig. 3).

In the color cube implementation, each category requires four specific neighboring cubes to be in an *on* position. Categories based on neighboring cube configurations were chosen, because they are easier to acquire. Exemplars of each category were generated by using color combinations of the remaining four non-relevant cubes (Fig. 4). Beneath the *color cube* images, an array of five target buttons was presented (Fig. 5). Each of the four peripheral buttons was randomly associated with one of the four target categories. The role of the central button will be detailed in the following procedure section.

The intersections of the categories' relevant cubes defined the set of *pair-features* (see Fig. 6). It was previously predicted that an internal representation of these *pair-features* should evolve if features are not selected by their information content for each categorization task individually, but rather by their information content to all categorization tasks collectively.

### Training Procedure

Experiment 1 was composed of four training stages. At each stage, participants learned one additional category in a trial-and-error paradigm (Fig. 7). In every trial a random subset of the eight input elements  $\mathbf{x}$  was activated. The resulting cube configuration image was displayed until the participants activated the category output buttons  $\mathbf{t}$ . If the wrong category elements were activated an error signal was presented. For example, if a certain trial presents the input elements  $\mathbf{x} = [-1, -1, +1, +1, +1, +1, -1, +1]$ , any reply other than  $\mathbf{t} = [-1, +1, -1, -1]$  would trigger the error signal. In theory, a certain cube configuration might be labeled by one of the  $2^4$  (16) different combinations of activating the four target category buttons  $\mathbf{t}$ . In order to simplify the training requirements the full 16 label setting was converted into a setting where only one button press is required at each trial. This goal was achieved by maintaining that only a subset of the target buttons were active at each trial and by reserving the central *default button*, for images not associated with any of the active category buttons. Therefore, in the first stage, dedicated to learning category

<sup>1</sup>The more dominant color was typically selected as the *on* state (e.g. Red-*on* vs. Cyan-*off*). In addition, cube edges were colored in the opponent color to emphasize the binary role of each input element.

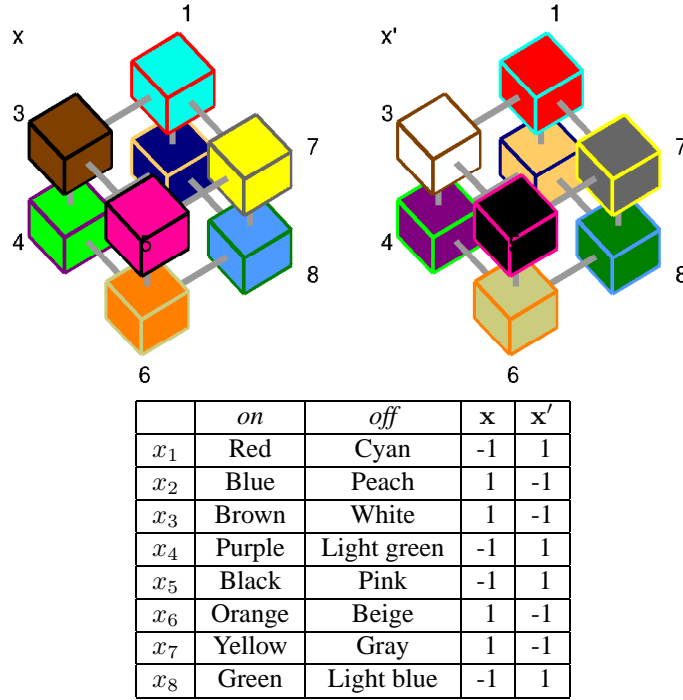


Figure 3: An example of two stimuli  $\mathbf{x}$  and  $\mathbf{x}'$  composed of eight binary *color cubes*. Each *color cube* appeared in two distinct colors indicating whether the cube input element was in an *on* or *off* state.

$C_1$ , both the central *default button*, and the button associated with category  $C_1$ , are constantly active. However, in the second stage, dedicated to learning category  $C_2$ , the central *default button* is constantly active but in certain trials only one of the buttons associated with categories  $C_1$  and  $C_2$  is active, thus maintaining that at each trial one and only one button was the accurate response<sup>2</sup>. If a wrong button was pressed, an error tone was triggered. In each stage, trials were generated until reaching a predefined criterion. This criterion required 100 consecutive successes, thus indicating that the participant reliably succeeded in associating the new category images with the designated target button. It should be emphasized that the random activation probability over the input set  $\{\mathbf{x}\}$  is constant throughout all four training stages. Thus, participants were constantly viewing stimuli generated with the same probability, while incrementally learning to recognize which exemplars were members of categories  $C_1$  through  $C_4$ . When these four training stages were concluded, a test was employed to validate whether the hypothesized *pair-features* have emerged.

In the remaining training procedure section, elaborate descriptions of several experimental controls are provided for completeness. The reader focusing on the conceptual core of the proposed experiment can skip to the following testing procedure section. As stated above, the random selection of input patterns for training, was based on a fixed sampling probability over the vector set  $\{\mathbf{x}\}$  that does not bias the *pairs-features* over the *incongruent-pairs*. For example, a uniform activation of all 256 possible input patterns  $\mathbf{x}$  maintains a similar frequency of appearance ( $\frac{1}{4}$ ) for the *pairs-features* and the *incongruent-pairs*. It could therefore be assured that if the hypothesized *pair-features* emerge, it is not as a result of suspicious coincidences or any other advantage in frequency of appearance over the *incongruent-pairs*. In preliminary experiments it had been established that the ratio of positive and negative examples of a certain category, is an important factor in the number of trials needed for training. therefore, a non-uniform distribution of input element vectors  $\{\mathbf{x}\}$  was chosen. This non-uniform distribution provides that the appearance rate of each category is  $\frac{5}{16}$ , rather than  $\frac{1}{16}$  resulting from the uniform probability. Appendix I provides a detailed description of the trial generating procedure employed in Experiment 1. It should be noted that the mutual information measurements presented earlier are derived using the same input generating distribution observed by the participants. The proposed setting also guarantees that the emergent representation is not a result of the *pair-features* discriminative value for any individual category. This is a result of comparing each *pair-feature* to an *incongruent-pair* which provides

<sup>2</sup>caution was taken so that participants could not induce which category was present by the pattern of active buttons

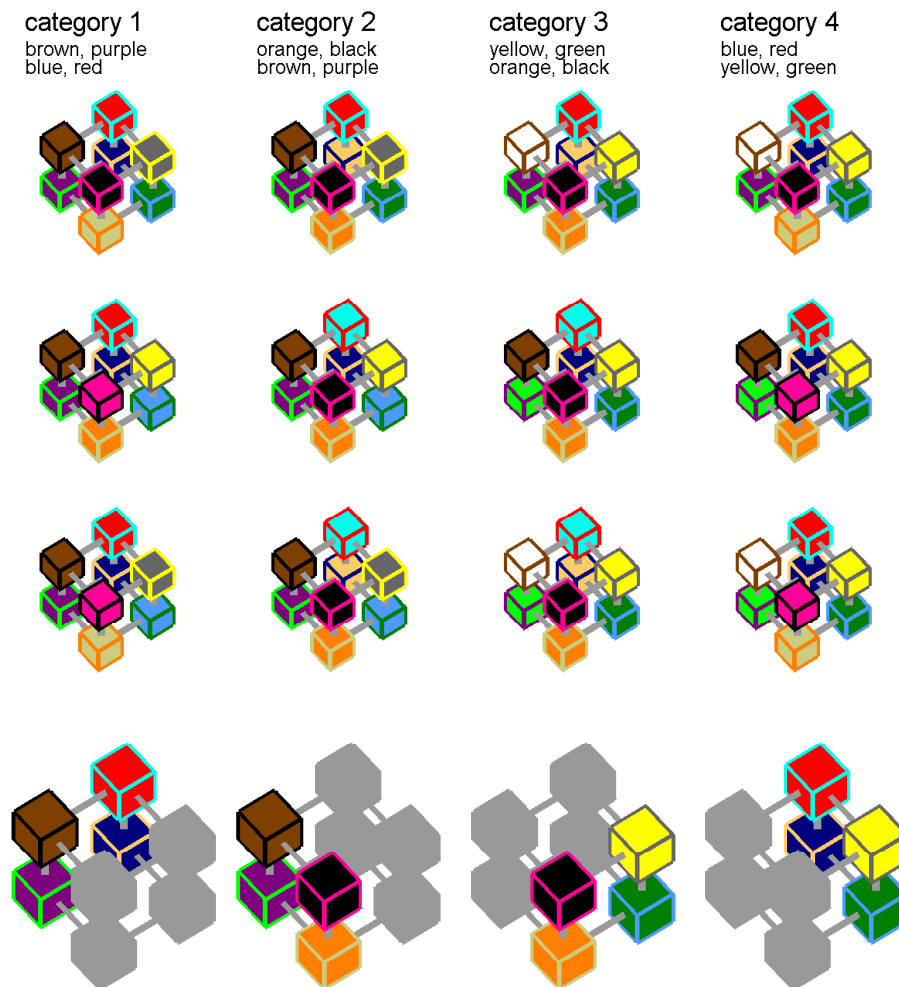


Figure 4: Prototypes (bottom row) and three exemplars of each of the four categories defined in Experiment 1.

an equal amount of information to any one specific category. For example, the information provided by *pair-feature*  $v_1$  to category  $C_4$ , was equal to the information provided by *incongruent-pair*  $h_1$  to the same category (both appearing in group (B) at the upper pane of Fig. 2). An additional factor that must be controlled is the perceptual salience of the *pair-features* and the *incongruent-pairs*. Although the input elements are arbitrarily parsed into *pair-features* or *incongruent-pairs*, an incidentally strong perceptual salience of specific pairs, might introduce a non desirable bias into the feature creation process. One might also claim that such an a-priori bias might exist for vertical vs. horizontal parsing of the categories into pairs. In order to take into account these two considerations, at every trial the entire three dimensional cube configuration was rotated in a way that guarantees similar spatial salience to the *pair-features* and the *incongruent-pairs*. For example, the cube configuration displayed in Fig. 5 is a rotated exemplar of the category  $C_4$  prototype displayed in Fig. 4. Thus the correct response button in this trial is the lower button associated with category  $C_4$  (see Fig. 7). A description of this rotation procedure is available at Appendix II.

### Testing Procedure

In the testing stage, each of the target buttons was highlighted in a sequential manner while requiring the participants to verbally report which *color cubes* composed the associated category. Participants were instructed to quickly and accurately report only the *color cubes* relevant for the highlighted category. The proposed test relies on the spreading of activation model to describe the activation process of the representational units (Collins & Loftus, 1975). Assume that the *pair-feature* representation depicted in Fig. 1 has emerged. It is therefore anticipated that activating

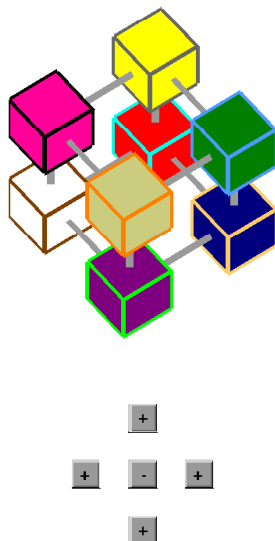


Figure 5: Participants were presented with a configuration of eight *color cubes* and set of target buttons. Each of the four categories was associated with one of the four peripheral target buttons.

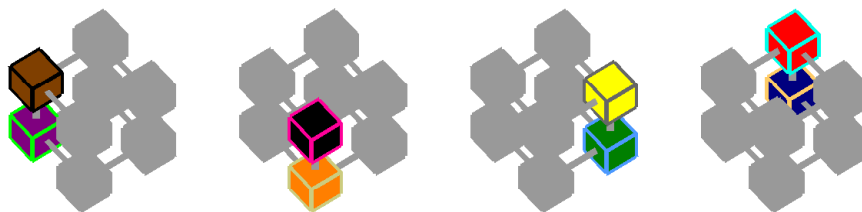


Figure 6: The color cube *pair-features* convey the maximum information on the four target categories.

category unit  $t_1$  would activate internal unit  $v_1$ , this in turn should activate input elements  $x_1$  and  $x_2$ . Similarly, it is anticipated that by stimulating category unit  $t_1$ , unit  $v_2$  should be activated as well, thus leading to an indirect activation of input elements  $x_3$  and  $x_4$ . In this spreading of activation scenario it was expected that the activation of units  $x_1$  and  $x_2$  should be highly correlated, since both result from the level of activation of the hidden unit  $v_1$ . Similarly the activation of units  $x_3$  and  $x_4$  should be highly correlated, since both result from the level of activation of the hidden unit  $v_2$ . Thus, the proposed test assumes that if an internal representation of the *pair-features* has emerged, then activating category unit  $t_1$  should lead to a correlated activity of two pairs of input elements, the first being  $x_1$  and  $x_2$  and the second being  $x_3$  and  $x_4$ . These correlation patterns would be manifested as a verbal report composed of two *pair-features*. Similar reporting patterns should also be exhibited in case of activating category output units  $t_2$  through  $t_4$ . These specific correlation patterns should not be exhibited if a representation based on any of the *incongruent-pairs* has emerged. It should be noted that although participants were explicitly required to verbally report each category, the sequence of reported *color cubes* is a purely implicit measurement. It is therefore assumed that the reporting sequence is not intentionally or unintentionally biased by the participants, and that these reports essentially reflect the internal structure that has emerged in the learning process.

### Results

Participants completed the four training stages described above in 2-3 hours. Approximately 1000 trials were required for passing all four training stages. All twenty participants succeeded in reporting the four colors relevant to each category. The frequency of reporting according to the *pair-feature* pattern was observed to be significantly

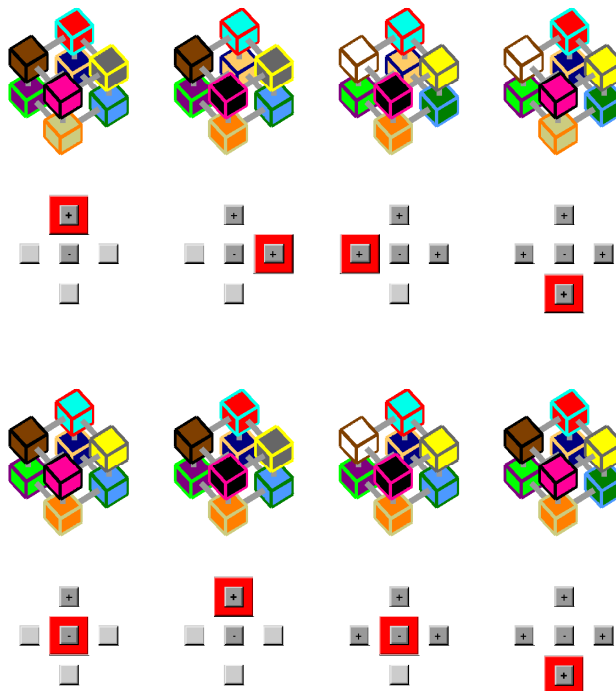


Figure 7: At each stage, participants learned one additional category. Depicted here from left to right, are examples of two trials (one on the top and one on the bottom), from each of the experimental stages 1 through 4. The category assignment to target buttons is: 1-top, 2-right, 3-left and 4-bottom. Active buttons (indicated by a "+" and dark gray shading) limited the choice of possible answers, thus assuring that only one button is the correct response at each trial. Correct responses are highlighted here for convenience by a red rectangle (not presented in the experimental setting).

higher than that of a comparable *incongruent pair* pattern in all four categories (binomial test,  $p \leq 0.05$ ,  $n=20$ ). Fig. 8 depicts the number of reports congruent with the *pair-feature* structure. Thus, Experiment 1 validates the first research prediction, stating that the *pair-features*, informative for recognizing many categories are preferentially encoded. However, Experiment 1, also addresses the second prediction, stating that encoding novel features might lead to a reconstruction of former category representations. This claim is based on observing the evolved representation of category  $C_1$ . When participants pass the first training stage, they are equipped with a representation that enables perfect recognition of all category  $C_1$  exemplars. At this point the *pair-features* have no salient representation because they are by definition identical to the *incongruent-pairs* until learning at least one additional category. However, when analyzing the reporting results of category  $C_1$ , registered *after* concluding the learning procedure of all four categories, it was evident that the initial representation of category  $C_1$  had been actively reconstructed. This reconstruction maintained that the representation of Category  $C_1$  complies with the *pair-features* acquired only later in the training process. Thus, the first two research predictions have been addressed while controlling the alternative factors that might have explained the emergent *pair-feature* structure.

## Experiment 2: Facilitated Perceptual Category Acquisition

Experiment 2 examined the third research prediction, namely that the perceptual system may utilize the informative feature set to facilitate learning of additional future categories.

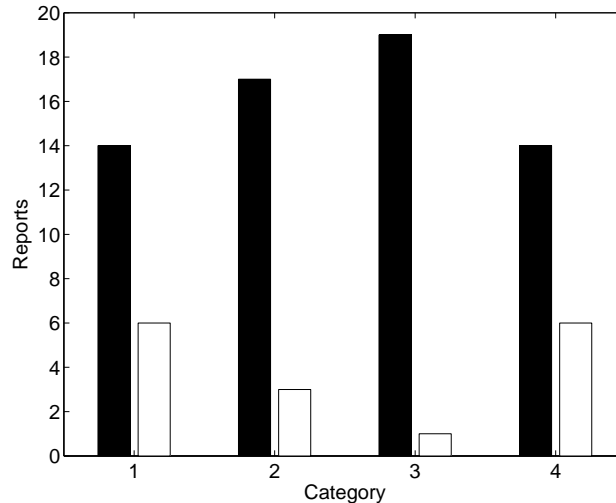


Figure 8: The number of reports congruent with the *pair-feature* structure (black) for categories  $C_1$  through  $C_4$ .

### Method

Experiment 2 included an additional training stage to the four training stages of Experiment 1. In this additional stage a examples of a new category,  $C_5$ , were presented. This new category is defined as  $C_5 \equiv (v_2 = 1) \cap (v_4 = 1)$  (see Table 1). Unlike the training procedure of Categories  $C_1$  through  $C_4$ , that continued until a criterion of perfect categorization performance had been reached, the fifth stage was restricted to a prefixed number of trials. It was predicted that when training a fifth category that is congruent with the *pair-feature* structure a significant facilitation would be observed. Comparing the performance of participants that have already learned the four initial categories to the performance of naive participants is meaningless, due to history confounds. Participants after two hours of training acquire a familiarity with the experimental setting that is not necessarily relevant to their feature representation, but that might lead to a performance advantage over naive participants. Therefore, Experiment 2 maintained a between subject paradigm by providing a control group that shares the same experimental history as the experimental group. This control group learned a different fifth category,  $\bar{C}_5 \equiv (h_2 = 1) \cap (h_4 = 1)$ , that is incongruent with the hypothesized *pair-features* (see Fig. 9).

### Participants

Ten undergraduate students from the Hebrew University of Jerusalem participated in Experiment 2 for payment or course credit. Participants, aged 20 to 30 years, were screened for normal color vision, and then randomly allocated to the experimental and control groups.

### Materials

The stimuli and setting of Experiment 2 were similar to those described in Experiment 1.

### Training Procedure

Participants were first required to complete training stages 1 through 4 as described in Experiment 1. The stimuli presentation procedure of the fifth stage was identical to the first four training stages except for the fact that trials are limited to a prefixed duration of six seconds. Both groups were required to learn the fifth category using a limited set of 48 training images. The fifth category, in both experimental conditions, was not composed of neighboring input elements and was thus expected to require more training trials than categories  $C_1$  through  $C_4$ . Providing 48 training trials was aimed at capturing an intermediate snapshot of the training process where the differential training rate of the experimental and control groups might be manifested. Rather than sampling 48 images from the input element

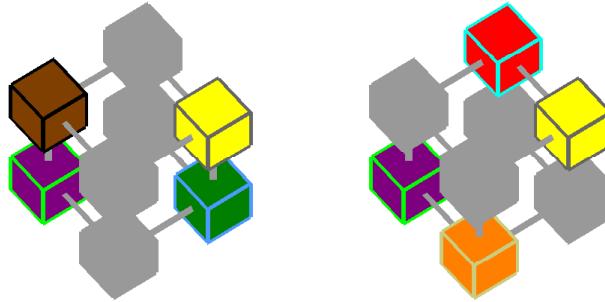


Figure 9: The input elements defining the fifth category in the congruent condition  $C_5 \equiv (v_2 = 1) \cap (v_4 = 1)$  (left) and in the incongruent control  $\bar{C}_5 \equiv (h_2 = 1) \cap (h_4 = 1)$  (right).

distribution, as in the first (unbounded) four training stages, the fifth stage displayed, in a random order, a predefined set of 24 negative examples and 24 positive examples of either the congruent category  $C_5$  or the incongruent category  $\bar{C}_5$ . The detailed composition of these 48 training trials is described in Appendix III.

#### Testing Procedure

Following the 48 training trials, participants were required to verbally report the *color cubes* composing the new category. If the *pair-features* have no functional influence on future category learning, it would be expected that the learning rate of the fifth category should be equal in both experimental groups. On the other hand, it was hypothesized that the previously encoded *pair-features* might facilitate future learning of a new congruent category. It was therefore anticipated that under such constrained training conditions, the congruent group would display a significant advantage in learning the fifth category.

#### Results

Both the experimental and the control groups reported that the training procedure of the fifth category was difficult. This effect is probably due to limited time provided for the 48 training trials and due to the fact that the fifth category was not composed of neighboring input elements. It therefore seems that the restricted training procedure had successfully avoided ceiling effects. When comparing participants' performance in the experimental and control groups it was observed that the learning rate was significantly higher in the congruent condition (Fisher Exact Probability Test,  $p \leq 0.05$ ,  $n=10$ ). Members of the congruent group reported on average 2.2 correct *color cubes*, i.e. participants learned most of the new category's characteristics. Members of the control group, reported on average only 0.8 out of the four *color cubes* present in category  $\bar{C}_5$ . Thus, the emerged *pair-feature* representation was indeed a functional tool in facilitating the acquisition of a novel perceptual category.

### Experiment 3: Encoding Semantic Features Informative for Multiple Categories

The goal of Experiment 3 was to examine whether the results of Experiment 1 were specific to the perceptual system or whether evidence for encoding features that are informative for recognizing many categories might also be observed in higher semantic categorization tasks.

#### Method

Due to the fact that Experiment 3 was aimed at replicating the structure of Experiment 1, the applied method was similar to that described above. The few modifications from the setting of Experiment 1, resulted from methodological

		<i>on</i>	<i>off</i>
$x_1$	Languages	Spanish	French
$x_2$	Marital Status	Married	Single
$x_3$	College	Private	Public
$x_4$	Age	Thirties	Twenties
$x_5$	Gender	Male	Female
$x_6$	Profession	Lawyer	Accountant
$x_7$	Residency	New York	New Jersey
$x_8$	Department	Local	International

Figure 10: The eight binary semantic characteristics implemented in Experiment 3

The figure displays two candidate profiles and their corresponding feature selection interface. Each candidate's profile is shown as a grid of feature-value pairs. Below each profile is a set of five target buttons (plus and minus signs) for feature selection.

**Candidate 1: JACQUELINE LOMAX**

- Residency: New Jersey
- Department Preference: Local
- Gender: Female
- Occupancy: Accountant
- College: Private
- Languages: Spanish
- Age: Thirties
- Marital Status: Single

**Candidate 2: TAYLOR DOUGHERTY**

- Gender: Female
- Residency: New York
- Department Preference: Local
- Marital Status: Married
- College: Private
- Languages: Spanish
- Occupancy: Lawyer
- Age: Thirties

Below each candidate's profile is a set of five target buttons (plus and minus signs) for feature selection.

Figure 11: Two trials presented in the fourth training stage of Experiment 3. To avoid spatial biases the field locations were randomly permuted in every trial.

hindsight gained in the previous experiments.

### Participants

Twelve undergraduate students from the Hebrew University of Jerusalem participated in Experiment 3 for payment or course credit. Participants, aged 20 to 30 years, were screened for normal vision.

### Materials

In order to test whether the semantic system complies with the first research prediction, the category structure described in Table 1 was implemented in a job candidate assignment task. In this experimental task participants were required to assign job applicants to four business firms. Eight binary characteristics  $x_{i,i=1,\dots,8}$ , encoded the input description of each candidate. For each characteristic, one value was selected to function as the *on* state  $x_i = 1$  and another as the *off* state  $x_i = -1$  (e.g. Department Preference: Local +1 and International -1). Thus, each of the four categories (business firms) required four specific characteristics to be in an *on* state. For example, Firm 3 required all candidates to be male, lawyers, living in New York and with a preference for working in the local department. As in Experiment 1, beneath each applicant sheet, an array of five target buttons was displayed (see Fig. 10). In order to control feature saliency, the allocation of characteristics was randomly selected for each participant. In addition the spatial position of each field was randomly permuted in each trial (Fig. 11).

The intersections of the categories' relevant characteristics defines the set of *pair-features*:

1. Spanish-speaker *and* Married
2. Private-collage *and* Thirties
3. Male *and* Lawyer
4. New York *and* Local-department

Since the input and category probabilities were identical to those described in Experiment 1, the mutual information measurements from Fig. 2 remain valid. It is predicted that an internal representation of the semantic *pair-features* should evolve, if features are not selected by their information content for each categorization task individually, but rather by their information content to all categorization tasks collectively.

### Procedure

The training procedure of Experiment 3 was similar to that described in Experiment 1. When the four training stages were concluded, participant were required to verbally report the characteristics relevant for each of the four firms.

### Results

Participants completed the four training stages described above in 4-6 hours. The extended training duration did not result from requiring more trials, but rather from the fact that each trial demanded approximately twice the time required in the perceptual training stages of Experiment 1. Analogously to the test applied in Experiment 1, the *pair-feature* report pattern was systematically compared to an a-priori equally complex set of *incongruent-pair* reports. Each report was encoded by observing whether the first two characteristics composed a *pair-feature* or an *incongruent pair*. It was observed that the frequency of reporting the congruent pattern was significantly higher than the *incongruent-pair* reports (binomial test,  $p \leq 0.05$ ,  $n=12$ ) in all of the four firm reports. Fig. 12 depicts the number of reports congruent with the *pair-feature* structure. Thus, the pattern of results provides evidence that in the semantic system, features are preferably encoded, if they provide information for multiple categories. The fact that the report of Category  $C_1$  reflects the *pair-feature* structure, validates the second prediction regarding the reconstruction of former category representations.

In addition to registering the report sequence (as in Experiment 1), the participants of Experiment 3 were recorded while verbally reporting the requirements of each firm. The reports of each firm were manually annotated by marking the starting time and the ending time of the four characteristics. An annotated audiogram of category  $C_3$  is presented in Fig. 13. By performing this annotation it becomes possible to measure the duration of the three gaps between the four reported characteristics and to assess whether the input elements composing the *pair-feature* are indeed temporally fused. It should be noted that only recordings of participants that reported Firm  $C_1$  as two *pair-features* were used in this analysis and that corrupted recordings and recordings where participants did not report only the relevant characteristics (e.g. incorporating words like AND into the report) were not used. The three gaps were scored according to the ascending order of their duration (the shortest gap was scored as 1, the intermediate gap was scored as 2 and the longest gap was scored as 3 see Fig. 13). It was observed that the second gap score ( $M = 2.86$ ,  $SD = 0.38$ ) was significantly larger ( $t(6) = 4.86$ ,  $p \leq 0.05$ , one tailed) than the first gap score ( $M = 1.83$ ,  $SD = 0.69$ ) and in addition significantly larger ( $t(6) = 10.49$ ,  $p \leq 0.05$ , one tailed) than the third gap score as well ( $M = 1.43$ ,  $SD = 0.53$ ). Moreover, the second gap score was significantly larger than the first and third gaps in all the reports of categories  $C_2$  through  $C_4$ . One might claim that this pattern of report results from a preexistent bias and does not reflect the *pair-feature* structure, therefore, the gap duration was also compared to a control group, where participants were required to only learn and report Firm  $C_1$ . This between subject test, compared the score difference between the second and first gaps (e.g. the report in Fig. 13 would be scored as  $3 - 2 = 1$ ). A significant difference ( $t(13) = 2.21$ ,  $p \leq 0.05$ , one tailed) was observed between the gap scores of the experiment group ( $M = 1.00$ ,  $SD = 1.00$ ) and the control group ( $M = 0.00$ ,  $SD = 1.16$ ). Similarly, when comparing the score difference between the second and third gaps (e.g. the report in Fig. 13 would be scored as  $3 - 1 = 2$ ). A significant difference ( $t(13) = 7.17$ ,  $p \leq 0.05$ , one tailed) was observed between the gap scores of the experiment group ( $M = 1.43$ ,  $SD = 0.53$ ) and the control group ( $M = -0.38$ ,  $SD = 1.19$ ).

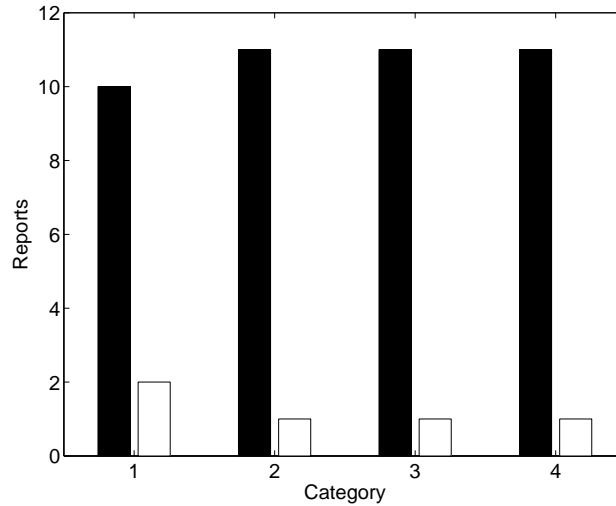


Figure 12: The number of reports congruent with the *pair-feature* structure (black) for categories  $C_1$  through  $C_4$ .

#### *Experiment 4: Facilitated Semantic Category Acquisition*

Experiment 4 examined whether the semantic system may utilize the informative feature set to facilitate learning of additional future categories.

##### *Method*

Experiment 4 replicated Experiment 2, utilizing the job candidate stimuli described in Experiment 3. Thus, the additional training stage required learning a fifth firm  $C_5 \equiv (v_2 = 1) \cap (v_4 = 1)$  from just a few training examples. As in Experiment 2 the facilitation of acquiring category  $C_5$  was assessed in comparison to a control fifth firm, defined as  $\bar{C}_5 \equiv (h_2 = 1) \cap (h_4 = 1)$ .

##### *Participants*

Twelve undergraduate students from the Hebrew University of Jerusalem participated in Experiment 4 for payment or course credit. Participants, aged 20 to 30 years, were screened for normal vision, and then randomly allocated to the experimental and control groups.

##### *Materials*

The stimuli and setting of Experiment 4 were similar to those described in Experiment 3.

##### *Procedure*

Participants were first required to complete training stages 1 through 4 as in Experiment 3. The stimuli presentation process of the fifth stage was similar to that described in the first four training stages except for the fact that the trial duration was fixed to sixteen seconds. Both the experimental and the control groups were required to learn the fifth category using a limited set of 48 training images (detailed in Appendix III). Following the 48 training trials, participants were first tested on their capability to correctly categorize the 48 training examples, by repeating the training procedure in the absence of any feedback signal. Only then did participants verbally report the candidate characteristics composing the new firm's requirements.

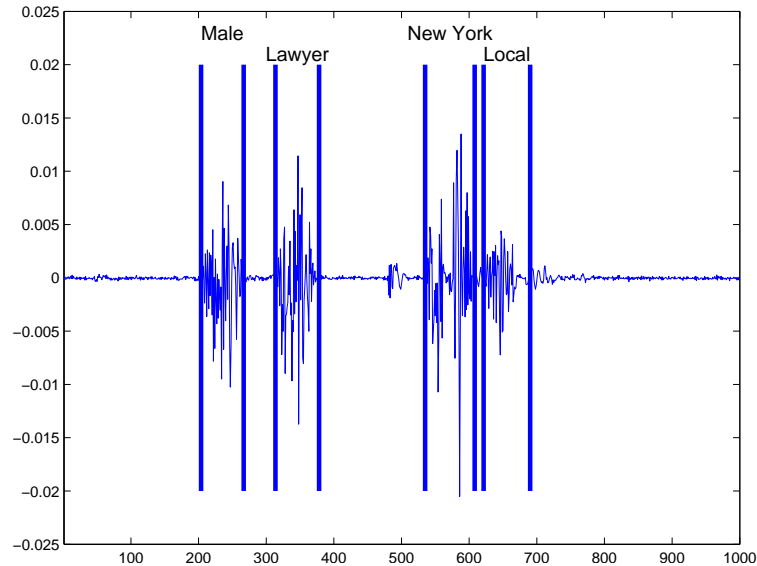


Figure 13: A 10-second audiogram recorded from one of the participants while reporting category  $C_3$ . In addition to the reporting sequence complying with the *pair-feature* structure, these pairs seem to be temporally clustered: the shortest gap appearing between NY and Local, the second shortest gap between Male and Lawyer while the longest gap appears between Lawyer and NY.

### Results

As in Experiment 2, both the experimental and the control groups reported that the training procedure of the fifth category was difficult. The participants' performance on the 48 testing trials (presented without feedback), was scored using an equal error rate measure:

$$\frac{1}{2} \left( \frac{\#true\ positives}{\#positives} + \frac{\#true\ negatives}{\#negatives} \right) .$$

This score averages the proportion of the correctly classified examples of the fifth category with the proportion of the correctly classified remaining fillers so that chance level is 0.50. It was observed that while the control group performs slightly above chance level ( $M = 0.59$ ,  $SD = 0.11$ ) the accuracy of the experimental group ( $M = 0.81$ ,  $SD = 0.13$ ) was substantially higher ( $t(10) = 2.38$ ,  $p \leq 0.05$ , one tailed).

Rather than simply counting the number of correctly provided characteristics as in Experiment 2, the reporting results were scored slightly differently by subtracting the number of any incorrect reported characteristics from the number of all correctly reported characteristics (thus, the score might range between -4 if four incorrect characteristics were reported and +4 if all four relevant characteristics are provided). Here too, a significant difference ( $t(10) = 4.65$ ,  $p \leq 0.05$ , one tailed) was observed between the experimental group scores ( $M = 2.33$ ,  $SD = 1.51$ ) and the control group scores ( $M = -0.33$ ,  $SD = 1.21$ ). It could therefore be concluded that the emerged *pair-feature* representation was a functional tool in facilitating the acquisition of a novel semantic category.

## Summary and Discussion

This research suggests that the human capacity to efficiently learn and recognize numerous categories relies on a rich set of features, learned during a prolonged history of categorization tasks. It was hypothesized that categorization systems must actively attempt to encode features that are informative for many of the categories in the environment. Three predictions were derived from the proposed theory. The first stated that a preference for encoding features that are informative for recognizing many categories, would be observed. The second prediction suggested that encoding

novel features might entail reconstructing former category representations. Finally, the third prediction claimed that once encoded, novel features can facilitate acquisition of future categories.

The experimental setting included four new categories that had to be learned, each based on a conjunction of four input elements. Pairs of input elements, termed *pair-features*, were suggested as the preferred internal representation due to their information content for the target categories. Although no direct feedback was provided for the *pair-feature* structure, it was experimentally found that participants' reporting patterns corresponded to an internal structure based on these predicted pairs. The existence of the *pair-features* cannot be attributed to their perceptual salience or frequency of appearance, for these factors were controlled by a systematic comparison to an alternative representation of *incongruent-pairs*. It was also observed that the *pair-feature* structure actively reconstructed the previously acquired representation of the first category. In addition, the experiments demonstrated that the emergent *pair-features* can facilitate learning of an additional fifth category. Thus, it can be summarized that the three predictions have been validated both in the perceptual system (Experiments 1 and 2) and using semantic stimuli (Experiments 3 and 4).

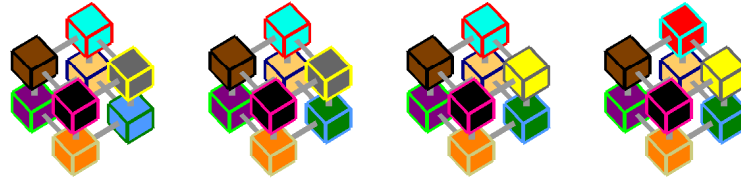
Several questions regarding the possible generalization of the proposed theory might be raised. First, the simple learning tasks stripped the category acquisition and feature creation processes to the bare minimum. Two conjunctions of binary inputs deterministically defined each category. The number of categories learned in Experiments 1 and 3 was selected so that in Experiments 2 and 4 it would be possible to train a novel combination of *pair-features*. The effects of more complex settings, like continuous input dimensions, complex combination rules (e.g. spatial-temporal / non-linear combination rules) and other learning schemes have not been tested. Testing the three predictions in these more elaborate (and natural) settings is the goal of future research. Finally, it will be necessary to assess the computational limits of the proposed theory. Direct maximization of the information between a large set of candidate features and a set of numerous categories is computationally unfeasible, and thus must be efficiently approximated. The challenge remains as to how the general principal of multi-category information maximization might be efficiently implemented in the perceptual and semantic systems.

## Appendix I: Experiment 1 Trial Composition

In preliminary experiments it had been established that a very low rate of positive examples entails a non-feasible training duration. If the random process generating the state of the input elements  $x$  samples uniformly from all  $2^8$  (256) possible configurations, the positive example rate would be  $\frac{16}{256}$ . This means that after viewing an example of a certain category, participants would observe an average of 15 negative examples before seeing another positive one. Learning in these conditions is extremely difficult. Therefore an alternative random generating process for the input elements was favored. This process selected with probability  $\frac{1}{4}$  whether four, five, six or seven input elements would be active in the current trial (see Fig. 14). Then with equal probability any of the images complying with this constraint might be selected. Thus, the probability of seeing a specific input pattern  $x$  with six active elements (e.g. the category  $C_2$  exemplar:  $x = [1, -1, 1, 1, 1, 1, -1, 1]$ ), is  $\frac{1}{4} \times \frac{1}{\binom{8}{6}} = \frac{1}{112}$ . Using these statistics the positive frequency rate increases to  $\frac{5}{16}$ , thus enabling training in a reasonable time for an experimental framework. It should be emphasized that this input generation procedure maintains equal probability rates for the *pair-features* and *incongruent-pairs*.

## Appendix II: Stimuli Rotation

One of the factors that must be controlled is the perceptual salience of the *pair-features* and the *incongruent-pairs*. Although the input elements are arbitrarily parsed into *pair-features* or *incongruent-pairs*, an incidentally strong perceptual salience of specific pairs, might introduce significant noise into the feature creation process. In addition, a-priori biases might exist for vertical vs. horizontal parsing of the categories into pairs. In order to take into account these two considerations, at every trial the entire three dimensional cube configuration was rotated in 90 degrees to a random selection of either the-X, the-Y or the-Z axis. This 90 degree rotation was displayed as a two second (18-frame) animation, appearing at the end of each trial (assisting the participants not to lose spatial orientation). As a result of this perceptual salience control, each cube configuration could appear in 24 different orientations (see Fig. 15).



active inputs	4	5	6	7
<i>on</i> colors	brown	brown	brown	brown
	purple	purple	purple	purple
	black	black	black	black
	orange	orange	orange	orange
		green	green	green
			yellow	yellow
				red
$p(\mathbf{x}) =$	$\frac{1}{4} \times \frac{1}{\binom{8}{4}} = \frac{1}{280}$	$\frac{1}{4} \times \frac{1}{\binom{8}{5}} = \frac{1}{224}$	$\frac{1}{4} \times \frac{1}{\binom{8}{6}} = \frac{1}{112}$	$\frac{1}{4} \times \frac{1}{\binom{8}{7}} = \frac{1}{32}$

Figure 14: *Pairs-features* and *incongruent-pairs* statistics are controlled by uniformly sampling from the input patterns  $\mathbf{x}$  having four, five, six or seven input elements in an *on* state.

### Appendix III: Experiment 2 Trial Composition

The fifth training stage of Experiment 2 displayed in random order a fixed set of 24 negative examples and 24 positive examples of either the congruent category  $C_5$  or the incongruent category  $\bar{C}_5$ . The 24 negative examples included 4 examples of each of the categories  $C_1$  through  $C_4$  and 8 non-category related fillers (Table 2). The 24 positive examples of  $C_5$  are described in Table 3. The composition of the 48 training trials of the incongruent control group results from replacing the 24 positive examples of category  $C_5$  with 24 positive examples of category  $\bar{C}_5$  (see Table 4).

### References

- Archambault, A., O'Donnell, C., & Schyns, P. G. (1999). Blind to object changes: When learning the same object at different levels of categorization modifies its perception. *Psychological Science*, *10*(3), 249–255.
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.
- Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 629–654.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.
- Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic priming. *Psychological Review*, *82*, 407–428.
- Edelman, S., Hiles, B. P., Yang, H., & Intrator, N. (2001). Probabilistic principles in unsupervised learning of visual structure: human data and a model. *Proceedings of the Neural Information Processing System conference (NIPS) 2001*.
- Fiser, J., & Aslin, R. N. (2002). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200.
- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and cognition. *Cognition*, *65*, 231–262.

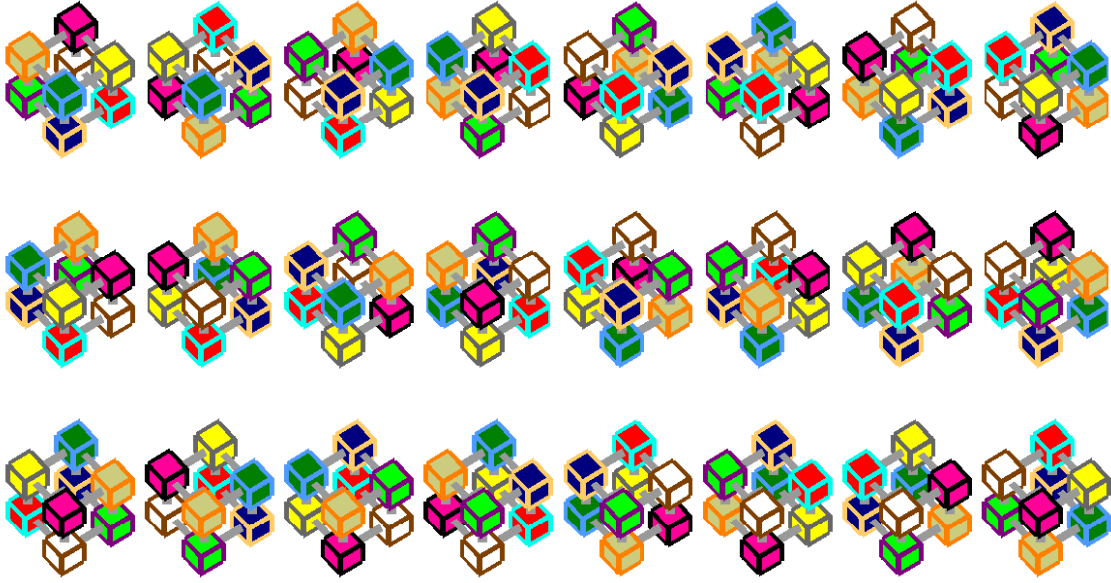


Figure 15: Each cube configuration is rotated to avoid spatial biases. Depicted here are the 24 possible displays of the category  $C_1$  exemplar  $[1, 1, 1, 1, -1, -1, -1, -1]$ .

	appearances	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
$C_1$	1	1	1	1	1	-1	-1	-1	1
	1	1	1	1	1	-1	-1	1	-1
	1	1	1	1	1	-1	1	-1	-1
	1	1	1	1	1	1	-1	-1	-1
$C_2$	1	-1	-1	1	1	1	1	-1	1
	1	-1	-1	1	1	1	1	1	-1
	1	-1	1	1	1	1	1	-1	-1
	1	1	-1	1	1	1	1	-1	-1
$C_3$	1	-1	-1	-1	1	1	1	1	1
	1	-1	-1	1	-1	1	1	1	1
	1	-1	1	-1	-1	1	1	1	1
	1	1	-1	-1	-1	1	1	1	1
$C_4$	1	1	1	-1	-1	-1	1	1	1
	1	1	1	-1	-1	1	-1	1	1
	1	1	1	-1	1	-1	-1	1	1
	1	1	1	1	-1	-1	-1	1	1
Default	1	-1	1	1	1	-1	1	-1	1
	1	1	-1	1	1	1	-1	1	-1
	1	1	1	-1	1	-1	1	-1	1
	1	1	1	1	-1	1	-1	1	-1
	1	-1	1	-1	1	-1	1	1	1
	1	1	-1	1	-1	1	-1	1	1
	1	-1	1	-1	1	1	1	-1	1
	1	1	-1	1	-1	1	1	1	-1

Table 2: The 24 negative examples displayed the fifth stage of Experiment 2 including 4 examples of each of the categories  $C_1$  through  $C_4$  and 8 non-category related fillers.

	appearances	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
$C_5$	6	-1	-1	1	1	-1	1	1	1
	6	-1	-1	1	1	1	-1	1	1
	6	-1	1	1	1	-1	-1	1	1
	6	1	-1	1	1	-1	-1	1	1

Table 3: The 24 positive examples of the congruent category  $C_5$  defined as a conjunction of  $v_2$  ( $x_3$  and  $x_4$ ) and  $v_4$  ( $x_7$  and  $x_8$ ).

	appearances	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
$C_5$	6	1	-1	-1	1	1	-1	1	1
	6	1	-1	-1	1	1	1	1	-1
	6	1	-1	1	1	1	-1	1	-1
	6	1	1	-1	1	1	-1	1	-1

Table 4: The 24 positive examples of the incongruent category  $\bar{C}_5$  defined as a conjunction of  $h_2$  ( $x_4$  and  $x_6$ ) and  $h_4$  ( $x_1$  and  $x_7$ ).

- Goodman, N. (1972). *Problems and projects*. Indianapolis: Bobbs Merrill.
- Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113–138.
- Nosofsky, R. M. (1988). The dynamics of similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(1), 54–65.
- Rosenthal, O., Fusi, S., & Hochstein, S. (2001). Forming classes by stimulus frequency: Behavior and theory. *Proceedings of the National Academy of Science*, 98, 4265–4270.
- Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). Development of features in object concepts. *Behavioral and Brain Sciences*, 21, 1–54.
- Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The Psychology of Learning and Motivation*, 31, 305–349.
- Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23(3), 681–696.
- Spencer-Smith, J., & Goldstone, R. (1997). The dynamics of similarity. *Bulletin of the Japanese Cognitive Science Society*, 4, 38–56.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–372.
- Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7), 682–687.