Title
How and Why Does Category Learning Cause Categorical Perception?

Journal
International Journal of Comparative Psychology, 30(0)

ISSN
0889-3675

Authors
Pérez-Gay Juárez, Fernanda
Thériault, Christian
Gregory, Madeline
et al.

Publication Date
2017

DOI
10.46867/ijcp.2017.30.01.01

INTERNATIONAL JOURNAL OF COMPARATIVE PSYCHOLOGY

2017, 30
Catherine Plowright
Special Issue Editor
Peer-reviewed

How and Why Does Category Learning Cause Categorical Perception?

Fernanda Pérez-Gay[1,2], Christian Thériault[1], Madeline Gregory[2], Hisham Sabri[1], Dan Rivas[1], and Stevan Harnad[1,3]

[1]Université du Québec à Montréal
[2]McGill University
[3]University of Southampton, UK

Learning to categorize requires distinguishing category members from non-members by detecting the features that covary with membership. Human participants were trained to sort visual textures into two categories by trial and error with corrective feedback. Difficulty levels were increased by decreasing the proportion of covariant features. Pairwise similarity judgments were tested before and after category learning. Three effects were observed: (1) The lower the proportion of covariant features, the more trials it took to learn the category and the fewer the participants who succeeded in learning it; after training, (2) perceived pairwise similarity decreased between categories and, to a lesser extent, (3) increased within categories, at all levels of difficulty, but only for successful learners. This perceived between-category separation and within-category compression is called categorical perception (CP). A very simple neural network model for category learning using uniform binary (0/1) features showed similar CP effects. We hypothesize that CP occurs because learning to selectively detect covariant features and ignore non-covariant features reduces the dimensionality of perceived similarity space. In addition to (1) – (3), the nets showed (4) a strong negative correlation between the proportion of covariant features and the size of the CP effect. This correlation was not evident in the human participants, probably because, unlike the formal binary features of the input to the nets, which were all uniform, the visual features of the human inputs varied in difficulty.

What is Categorization?

Categorizing includes sorting things into groups based on similarity, pigeon-holing things, and naming things (Goldstone, 1994a; Nosofsky, 1986; Zarate & Smith, 1990). But to define a category such as "categorization" we cannot just give examples of it. We have to say what its covariant features are: what are the features that are common to all instances of "categorization"? Some have suggested that there is something wrong with this "classical view" of categorization, according to which all category members share features that determine whether or not they are members (Mervis & Rosch, 1981; Smith & Medin, 1981). But the alternative to assuming that there must be features that distinguish members from non-members is either to assume that categories are not categorical but just a matter of degree (McCloskey & Glucksberg, 1978) – i.e., that everything is a member of every category to some extent, but more a member of some categories than others – or to assume that categorizing has no objective (covariant) basis at all (Fodor, 1998).

This paper assumes that the classical view of categories and categorization is correct. The only thing that determines whether something is or is not a member of a category is the presence or absence of the features that distinguish the members of that category from the non-members – i.e., the features that covary with membership in the category. This puts a big load on the notion of a "feature". Features are mostly sensory properties of things, such as size, color, shape, loudness or odor. But each feature is itself a potential category too, with members and non-members. That is why it is important to define "categorization" first:

Please send correspondence to Fernanda Pérez-Gay, Département de Psychologie, Université du Québec à Montréal, Pavillon André Pinard (SU), Room 1230, 100 rue Sherbrooke, Montréal, QC, Canada H2X 3P2. (Email: fernanda.perezgay@mail.mcgill.ca)
https://doi.org/10.46867/ijcp.2017.30.01.01

Categorizing is an action. Categorizing is something that organisms -- and not just human ones (Jitsumori & Delius, 2008; Thompson & Oden, 2000) -- do with "things": to categorize is to do the right thing with the right kind of thing (Harnad, 2005). A category is a kind of thing. "Things" are whatever input an organism receives through its senses from objects (including events and actions and states and properties) in the world; and the "right thing to do" is whatever the organism needs to do with the members of the category and not-do with the non-members of the category (Fajen & Turvey, 2003) in order to survive, succeed and reproduce; for example, eat what's "edible" and don't eat what's "inedible"; approach "prey," avoid "predators;" and mate with members of your own species and not with trees or stones.

Naming. For human beings, too, to categorize is to do the right thing with the right kind of thing; but apart from the basic survival- and subsistence-doings that we share with other species, our species is capable of a kind of doing that is almost certainly unique to us: naming things (verbally or gesturally). Names are arbitrary, and we are the ones who agree, by convention, to call some kinds (categories) of things this and other kinds of things that. But we don't name categories just for the pleasure of naming. The unique power of language (Blondin-Massé, Harnad, Picard, & St-Louis, 2013; Cangelosi, Greco, & Harnad, 2002; Pagin & Westerståhl, 2010; Wisniewski & Gentner, 1991) is to enable us to combine category names to produce subject/predicate propositions that describe or define further categories whose features are a combination of the categories that we have already learned and named: "bachelor = unmarried man"; "apple = red, round fruit"; "prime number = divisible only by itself and 1"; etc.

The propositional power of learning new categories through language alone depends, in turn, on our previously having already learned enough categories without language, directly through our senses, and to have given them names, so as to enable us to then go on to name and define other categories through words alone. How many categories -- and which ones -- we need to have learned and named in order to be able to do this (Vincent-Lamarre, Blondin Massé, Lopes, Lord, Marcotte, & Harnad, 2016) is not the subject of this paper. This paper is about what must precede language: how we learn categories through direct sensorimotor doings rather than through verbal doings based on re-combinations of categories that we have already learned and named.

Innate and learned categories. Not all categories need to be learned. Some are inborn: The frog is born knowing that the right thing to do with an object of a certain size flying across its visual field is to flick its tongue out and try to catch and eat it (Zhaoping, 2016); the duckling is born knowing that the right thing to do with a moving thing of a certain size and speed is to follow it (Bateson, 2000); and the newborn mammal knows that the right thing to do is to seek things that look and smell like nipples and to suckle (Porter & Winberg, 1999). But even for nonhuman animals, many of their categories must be learned (Gershman & Daw, 2017; Marino, 2017; Smith, 2017; Smith, Zakrzewski, Johnson, Valleau, & Church, 2016; Zentall, Wasserman, Lazareva, Thompson, & Ratterman, 2008) and for human beings virtually all the nouns, verbs, adjectives and adverbs in our dictionaries are the names of categories that we had to learn rather than being born already knowing them (Bloom, 2000; Horst & Simmering, 2015).

Supervised and unsupervised category learning. What is it to "know" a category, whether innately or through learning? It is to be able to categorize (i.e., do the right thing with its members, and not its non-members). How do organisms know what to do with what? Whether the category is innate or learned, their brains must be able to detect the features that distinguish the members from the non-members, the features that covary with category membership (Gao, Cai, Li, Zhang, & Li, 2016; Smith & Rangarajan, 2016). The frog, the duckling and the human neonate each have inborn feature-detectors for their innate categories, selectively tuned to the covariant features. But when the organism does not have feature-detectors that have already been

"prepared" by millions of years of evolutionary trial and error by random variation and selective retention based on survival and reproduction, the organism has to learn by trial and error during its own lifetime what is the right thing to do with the right kind of thing. This kind of learning is called "supervised learning" (LeCun, Bengio, & Hinton, 2015) or "reinforcement learning" (Sutton & Barto, 1998), because it is based on the learner's first trying to do the right thing with an input and then receiving corrective feedback (reinforcement, supervision) from the consequences of that action – feedback that is positive if the organism has done the right thing and negative it if has done the wrong thing[1]. Under these concepts, "supervised learning" refers to learning model architectures in which the actual response and the correct response are simultaneous components of the model's input and connection strengths are adjusted to minimize the discrepancy, as in error back-propagation. "Reinforcement learning" refers to performance tasks in which the response occurs first, then the consequences (negative or positive) follow, and then the connection strengths are adjusted. By these means, the brain (or any learning mechanism) must learn to detect which features of the input covary with membership – the features that are shared by the members of the category and that distinguish them from the non-members.

Categorical Perception (CP)

This paper will report some findings from experiments on human participants who are trained in the laboratory to learn new categories. We are particularly interested here in a phenomenon called "categorical perception" (CP) that sometimes (but not always) accompanies categorization, both innate and learned (Harnad, 1987, 2003). CP occurs when members of different categories look more different from one another and members of the same category look more alike. On the face of it, this sounds trivial: Of course members of the same category look more like one another than like members of a different category! Otherwise why would they be in the same category? This sounds like the very first notion of categorization mentioned in the opening paragraph of this paper: "sorting things into groups based on similarity." But categorization depends on doing the right thing with the right kind of thing; and it is not clear that the obvious similarities and differences in the appearances of things (to any of our senses) are always enough to inform us as to what to do with what -- at least not for some categories, and not immediately. For categories whose covarying features are harder to find (rather than evident upon repeated exposure through unsupervised learning), learning to categorize may be more challenging and time-consuming. And the resulting feature-detector may alter what we perceive as more similar to what, so as to make the category readily recognizable.

The Whorf hypothesis. According to the "Whorf Hypothesis" (Hussein, 2012; Kay & Kempton, 1984), it is learning to put things in different categories by giving them different names that makes them look more different to us, rather than vice-versa: Categorical Perception (CP), the expansion or separation of the perceived differences between members of different categories and the compression of the perceived differences among members of the same category is attributed to "language" by the Whorf Hypothesis: The reason different shades of green all look green rather than blue is that we call them "green" rather than "blue." In languages that do not have a word to distinguish green from blue, but only the equivalent of "bleen" to name them both, they both look bleen, rather than green or blue, according to the Whorf Hypothesis.

Color categories. Perceived colors (hues) correspond to differences in wave-length of light (Figure 1). Small equal-sized differences in wave-length that cross the boundary between the range we call "green"

[1] Another form of learning is unsupervised learning (Fisher, Pazzani, & Langley, 2014), through repeated passive exposure, without response or feedback. Some covariation can already be learned this way, and this may facilitate later category learning. Unsupervised learning is further discussed in the section on our neural net modeling.

and the range we call "blue" look more different and are easier to discriminate than equal-sized differences within the range of green or within the range of blue (Bird, Berens, Horner, & Franklin, 2014; Hanley, 2015). If this CP separation/compression had been the result of language and naming, it would have been a bona-fide Whorfian effect. But cross-cultural and cross-linguistic comparisons of color naming and color discrimination have found that not only do most languages give distinct names to the very same distinct sub-regions of the visible wave-length continuum, but that even the speakers of those languages that do not (as in "bleen" languages) nevertheless show the same separation/compression CP effects within and across the color boundaries as the speakers of those languages that do (Ozturk, Shayan, Liszkowski, & Majid, 2013). So in the case of color perception, the categorization indeed follows the perceived similarity gradient rather than causes it. Yet the perceived similarities and differences are not really in the color input, which is just light varying continuously along a uniform gradient of wave-length. The CP effect is the result of the brain's innate feature-detectors for color (involving cones that are selectively tuned for the red, green and blue regions of the frequency spectrum, paired red/green and blue/yellow opponent processes, and other specialized neural feature-detectors for perceiving color categories; Jacobs, 2013). So it is not language that has produced the perceived separation/compression among colors. It is Darwinian evolution.

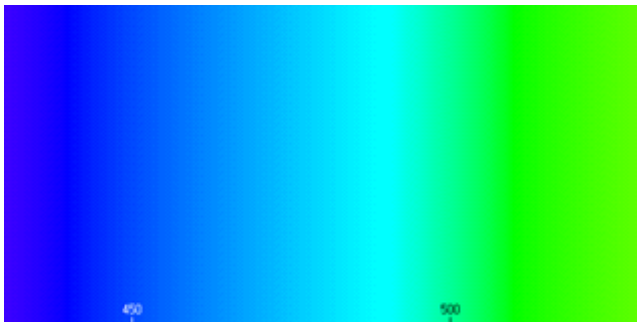

Figure 1. **Blue-Green ("Bleen") continuum (in nanometers) of light wavelength**. By Xession (Own work) [CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons.

CP induced by learning. There is nevertheless growing evidence in other sensory domains that "weak" Whorfian CP effects – not as dramatic as the rainbow, not always both separation and compression, but going in that direction – have been reported under some conditions as a result of category learning alone, with simple stimuli, colors and faces (Clifford et al., 2012; De Baene, Ons, Wagemans, & Vogels, 2008; Goldstone, 1994b; Goldstone, Lippa, & Shiffrin, 2001; Goldstone & Steyvers, 2001; Harnad, 1987; Kang 2014; Livingston, Andrews, & Harnad, 1998; Notman & Snowden, 2005; Pérez-Gay et al., 2016; Pevtzow & Harnad, 1997; Pothos & Reppa, 2014; Sigala & Logothetis, 2013; Simanova et al., 2016; Wallraven, Bülthoff, Waterkamp, van Dam, & Gaißert, 2013). Studies of shape category learning, however, have been more mixed (Folstein et al., 2012, 2014; Gillebert, Op de Beeck, Panis, & Wagemans, 2008; Jiang et al., 2007; Newell & Bülthoff, 2002; Ozgen & Davies, 2002; Van Gulick & Gauthier, 2014).

We report here some recent behavioral findings on CP separation/compression effects induced by learning to categorize unfamiliar visual textures of increasing levels of difficulty (Experiments 1, 2 and 3). Textures were used because they are novel stimuli and their features are distributed and holistic, rather than local and shape-based. This was intended to make it harder for participants to learn to categorize them by using an explicit, verbal rule based on familiar, local features. A simplified version of the human participants' category-learning task was then administered to a category-learning neural network model (Experiment 4) to test whether it too generated CP separation/compression effects, and if so, how and why.

Experiment 1

Method

Participants. Forty-two right-handed participants (22 males, 20 females), aged between 18 and 35 years were recruited online through Kijiji and the UQAM and McGill Classified Ads Website. Participants were either native English-speakers or native French-speakers, free of significant neurological and/or psychiatric conditions. Each participant was assigned randomly to one of four levels of difficulty.

This experiment, as well as the following two, were approved as part of the same project by the CIEREH, Comité institutionnel d'éthique de la recherche avec des êtres humains of UQAM – Université du Québec à Montréal. All participants signed the informed consent form approved by the Committee before participating in the experiment.

Stimuli. To design a categorization task with unfamiliar stimuli and features that were not local but distributed (hence less readily verbalizable), we computer-generated a large set of 270 x 270 pixel black and white square-shaped textures. The textures were in turn composed of twelve distinct 6 x 6 black and white matrices, each consisting of 18 black and 18 white pixels arranged in twelve different patterns (Figure 2a). These twelve matrices were then split (arbitrarily) into 6 pairs to serve as binary, mutually exclusive features. For simplicity, we will henceforth refer to the matrices as "features".

Each individual texture was thus built out of 900 features, 30 along the width dimension and 30 along the height dimension, their spatial positions randomly distributed. The resulting 180x180 grid was then amplified 1.5 times to result in our final 270x270 textures. For each category, the texture would include only one or the other of each of the six binary features. This generated a very large sample of textures to be used as training sets for category learning (examples in Figure 2b).

The task had four levels of apriori (i.e., presumed) difficulty based on what proportion of the features co-varied with category membership (but each participant was trained at only one difficulty level). At the easiest level, all 6 binary features covaried with category membership: the zero value of each binary pair occurred in every member of the K category (KALAMITES) and the one value of each pair occurred in every member of the L category (LAKAMITES). Our assumption was that stimuli in which all the features covaried with category membership would be the easiest to learn to categorize, and that difficulty would increase as the proportion of covarying (relevant) features decreased and the proportion of non-covarying (irrelevant) features increased. The four levels of difficulty we tested ranged from 6/6 covariants (easiest), to 5/6, 4/6 and 3/6 (hardest). The non-covarying features varied randomly at each level, independent of category membership. Pilot testing suggested that the smallest ratios of covariants (2/6 or 1/6) made the category unlearnable within a single training session, so we tested only from 6/6 to 3/6. Each stimulus set consisted of 180 different texture images (each presented two to three times with corrective feedback following each response, across a total of 400 training trials). The experiment was created and the stimuli were generated using the PsychoPy2 open source software (Peirce, 2007, 2009).

Although we tried to make all six features as equal as possible, minimizing disparities, only in the formal neural net model described in Experiment 4 could full feature equality be ensured. In Experiments 1 and 2, the proportion of covariant features decreased at each difficulty level (k/N, k being the number of features that co-vary with category membershinp, and N being the total number of features that conform each stimulus), but only one arbitrary combination of k features was tested at each level, not every possible combination of k features: For example, all Ss trained at level 3/6 viewed stimuli with the very same three (arbitrarily chosen) covariant features (Figure 2).

One of the objectives of Experiment 1 was to derive from the apriori difficulty levels an estimate of the aposteriori (i.e., observed) difficulty levels based on the Ss' actual performance, as reflected by the percentage of successful learners and the number of trials it took them to reach the learning criterion. On this performance basis we could then choose our stimuli and difficulty levels for Experiment 2, as described below. The outcomes are important for our discussion, as the role of feature inhomogeneity for the human visual system compared to the feature homogeneity of the neural net model proved to be highly relevant factor in the interpretation of our overall findings.

Procedure. The experiment was conducted in a sound isolated chamber with dim lighting and no other sources of electromagnetic interference. Ss were seated in a comfortable armchair in front of a glass window through which they saw the computer screen presenting the stimuli. They had a keyboard on a table between themselves and the window to click the K and L keys.

Task. A standard supervised learning procedure was used, consisting of trial-and-error training with corrective feedback following each trial. The training session lasted about forty minutes (pauses included). Ss had to learn to categorize each texture as

either a “KALAMITE” or a “LAKAMITE” by pressing K or L. The training set included 180 different textures generated as described above. Ss saw a total of four hundred textures (each texture appeared 2-3 different times during the task).

Each trial consisted of a fixation cross (500 ms) followed by one of the stimuli, shown at the center of the computer screen against a white background (1.25 s). Ss were instructed to click K or L on the keyboard to indicate the category. Ss had to respond within 2s of the onset of the stimulus; if they did not, the computer prompted them to respond faster. Categorization was followed by immediate feedback (lasting 750 ms) indicating whether the response had been correct or incorrect.

The 400 training trials were divided into four blocks of 100 stimuli each. Following each block, there was a pause in which Ss responded to a questionnaire about whether they thought they had detected the difference between the KALAMITES and LAKAMITES. If they replied “yes”, they were asked to describe in words what the difference was. If they replied “no”, they were asked to describe the provisional strategy they were using to try to categorize the stimuli. The instructions and questionnaires were in English or French depending on the Ss’ native language. We recorded responses as well as reaction times during the task.

Results

Training trials and indicators of task difficulty. Forty-two participants (aged 19-34, 22 male, 20 female) completed the visual category-learning task, each randomly assigned to each of the four difficulty levels. Overall, 28 of the 42 Ss succeeded in learning the category as determined by attaining our apriori monotonic learning criterion (reaching and sustaining at least 80% correct). Four Ss were classified as “Borderlines”: they reached but did not sustain the 80% level. The remaining 10 Ss did not reach the learning criterion throughout the training session and were classified as non-learners (Figure 3).

To test the apriori difficulty of our stimulus sets (based on the assumption that the lower the proportion of features covarying with category membership, the greater the difficulty), we compared performance at each level of difficulty in terms of (1) the proportion of Ss who successfully reached the 80% learning criterion and (2) the number of trials it took them to do so (Table 1 and Figure 4). A one-way ANOVA showed a significant effect of difficulty level (proportion of covariates) on the number of trials it took to reach the learning criterion, $F(3,25) = 3.066$, $p = 0.046$. The participants’ performance levels did not correspond to our apriori expected levels of difficulty, however. We had expected the hardest level to be the one with the lowest proportion of covarying features (3/6), but this level turned out to require the smallest number of trials to reach the learning criterion. Also, the level with 5/6 covarying features turned out to be (nonsignificantly) easier than the level with 6/6 covarying features, $t(20) = 0.341$, $p = 0.283$. This indicates that the sets did differ in difficulty, but that the degrees of difficulty did not correspond to the proportion of covariants.

Table 1
Outcome Profile for Each Apriori Difficulty Level in Experiment 1

Apriori difficulty level	Covarying features (k/N)	Learners	Borderlines	Non-Learners	Trials to learn: mean (SE)	Aposteriori difficulty level
1	6/6	7	1	3	149 (21)	3
2	5/6	8	0	3	180 (38)	2
3	4/6	5	3	2	278 (61)	4
4	3/6	8	0	2	138 (20)	1

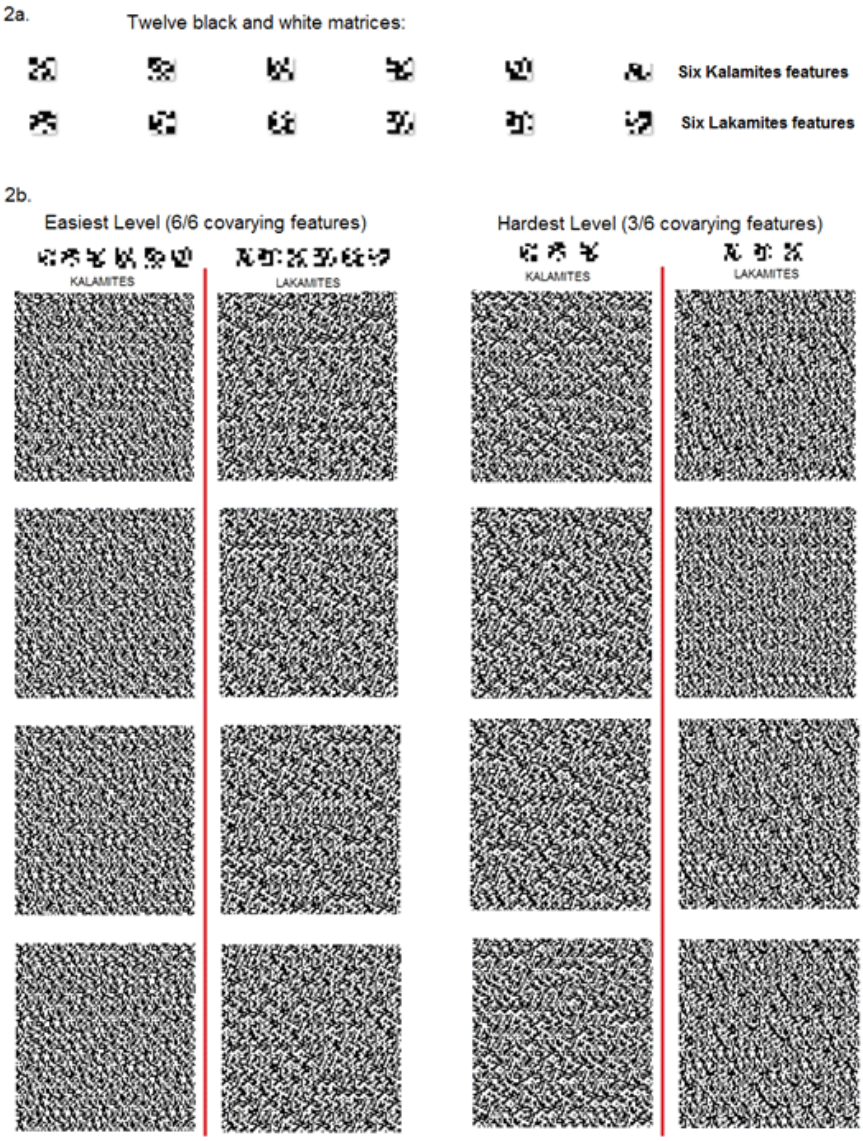

Figure 2. **The twelve matrices (henceforth "features") of which the textures were composed and a sample of category members from the a-priori easiest (6/6) and hardest (3/6) level of difficulty in Experiments 1 and 2**. Above (2a): the six specific pairs of binary features used to generate the two texture categories: "Kalamites" (Ks) and "Lakamites" (Ls). Below (2b) Left: sample of 4 Kalamites and 4 Lakamites at the easiest level (6/6, in which all six features covaried with category membership) Right: 4 Kalamites and 4 Lakamites at the hardest level (3/6, in which only three of the six features covaried with membership; the non-covarying pairs varied randomly).

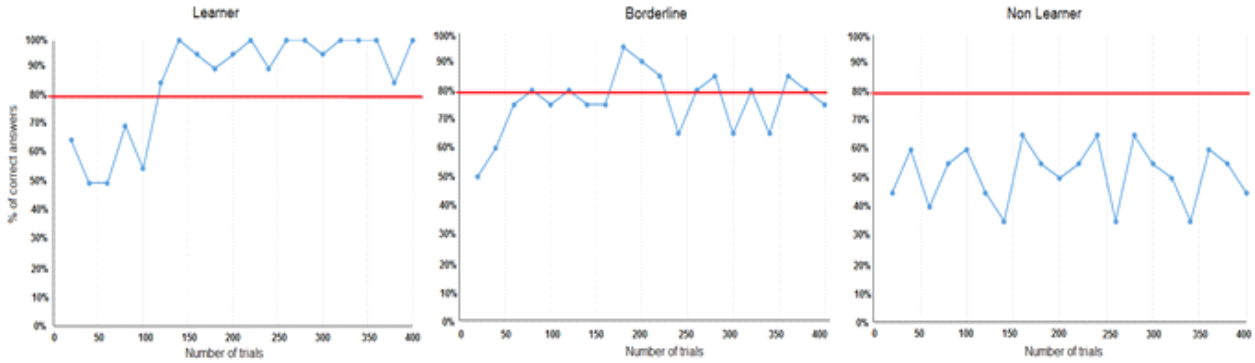

Figure 3. **Examples of the three patterns of learning observed for the 42 participants in our categorization task**. Left to right: Learners (29), borderlines (5) and non-learners (11). X axis is number of trials and Y axis is percentage of correct responses. Red line represents learning criterion of 80% sustained correct responses.

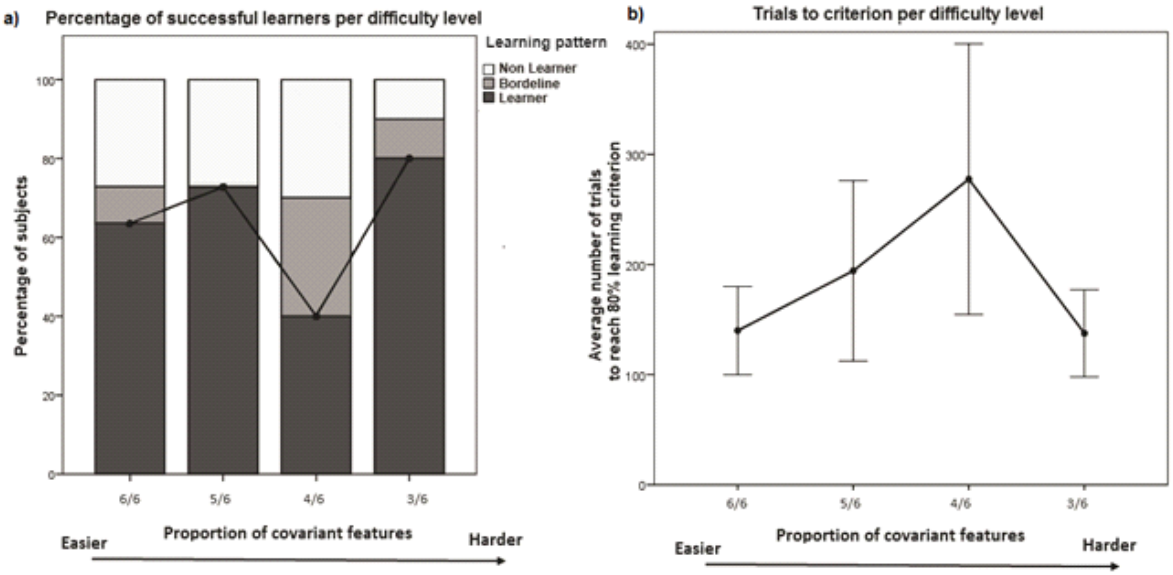

Figure 4. **Indicators of difficulty in Experiment 1**. On the left, Figure 4a) shows the percentage of participants that reached our learning criterion. On the right, figure 4b) shows the number of trials it took the learners to reach it. The four groups differed significantly but difficulty (percentage of unsuccessful learners and number of trials to reach criterion) did not increase monotonically as the proportion of covariant features decreased. Error bars represent ±2 SE.

We think this outcome pattern arises because our difficulty levels were not determined solely by the proportion of covariant features at each level, but also by which particular features were used at each level (see Appendix). The covariant features at each level were picked randomly but then that choice was the same for all participants at that level: only one random combination was used at each level.

Experiment 2

We chose the two apriori levels that had also proved significantly different aposteriori in Experiment 1, in terms of number of trials to criterion and proportion of learners/non-learners, and conducted a second experiment to test for changes in perceived similarity after the categorization task. The experiment was conducted under the same conditions as those described in the Procedure section of Experiment 1.

Method

Participants. Forty-one right-handed participants (23 Females, 18 Males) aged between 18 and 35 years were recruited as in Experiment 1. The difficulty condition for each participant was assigned randomly.

Stimuli. We used only two levels of difficulty, one that had proven easier in Experiment 1 (5/6 covarying features) and one that had proven harder (4/6 covarying features).

Procedure. Same as in Experiment 1.

Tasks. In Experiment 2, participants made pairwise similarity judgments on a subset of forty stimulus-pairs, once before the categorization training and once again after the training: A fixation cross was presented for 500 ms and then two stimuli appeared at the center of the screen one after the other for 1s each, with an inter-stimulus interval of 1s. After the second stimulus participants were to rate the similarity of the two stimuli on a scale of 1 to 9 on the keyboard with 1 corresponding to "very similar" and 9 to "very different." They were encouraged to make use of the full range of the scale. Of the total of forty pairs presented, 20 were within-category pairs (10 "Kalamites" and 10 "Lakamites") and 20 were between-category pairs (but of course before training, participants did not know the categories or their names). We recorded responses and reaction times during the task. The same set of 40 stimulus-pairs were presented in the same order for the similarity judgements before and after training. Following the first set of similarity judgements, participants began their visual category training with corrective feedback as in the first experiment (400 trials divided into four blocks with questionnaires in each pause).

Results

Learning. Forty-one participants completed the category-learning task, 21 assigned to the easier (5/6) and 20 to the harder (4/6) level. Twenty-four participants successfully reached our learning criterion (sustained 80% correct), 16 at the easier level, 8 at the harder level and were considered learners. Interestingly, at the easier level, 5 participants already had 80% accuracy from the outset, without training. This suggests that mere exposure to the 40 pairs of stimuli during the pairwise similarity judgement task, without feedback to indicate that there were two categories or which belonged in which, had been enough to induce passive learning (i.e., unsupervised learning) in these participants in the easier condition. We classified them as "immediate learners".

In the harder condition, two participants were classified as "Borderlines", showing the same pattern of partial learning described in Experiment 1 for this same difficulty. Four participants in the easy condition and ten in the hard condition were non-learners. See Table 2. It is important to point out that the average number of trials to learn does not include the immediate learners, who were performing above our criterion from the first 20 trials.

Similarity judgments. To compare similarity ratings before and after training, we calculated the "average pairwise distances", meaning the average score given by the participant for stimulus pairs between (B) or within (W) categories. We did so for the session before (pre) training and the session after (post), creating four variables "Bpre," "Bpost," "Wpre" and "Wpost." To estimate between-category separation we assessed

the changes in similarity ratings for between-category pairs. (diffB = Bpost – Bpre); we did the same for within-category compression (diffW = Wpost – Wpre). We then separately tested whether diffB and diffW differed from zero with separate correlated-sample t-tests for learners and non-learners. A repeated-measures ANOVA with learner/non-learner as the between-group factor tested the effect of successful learning on these perceived changes in pairwise similarity between and within categories. Combining both difficulty levels for the learners (including the immediate learners) the between-category separation (positive diffB) was significant, $t(24) = 6.785$, $p < 0.001$, but the within-category compression (negative diffW) was not, $t(24) = -1.551$, $p = 0.134$ (Figure 5). In repeated measures ANOVAS with Learning group as the between-subject factor, excluding the two borderlines, both Separation, $F(1, 37) = 5.320$, $p = 0.027$, and Compression, $F(1, 37) = 4.083$, $p = 0.050$, were significant.

Table 2

Number of Learners and Number of Trials before Reaching the Learning Criterion for Each Level in Experiment 2

Level	Immediate Learners	Learners	Borderlines	Non-Learners	Trials to learn: mean (SE)
Easier	6	10	0	5	106 (33)
Harder	0	8	2	10	262 (32)

The non-learners had no significant effects at all (only small, non-significant separation both within categories, $t(13) = 1.296$, $p = 0.217$, and between categories, $t(13) = 1.566$, $p = 0.141$. The changes in perceived similarity for both learners and non-learners are shown in Figure 5.

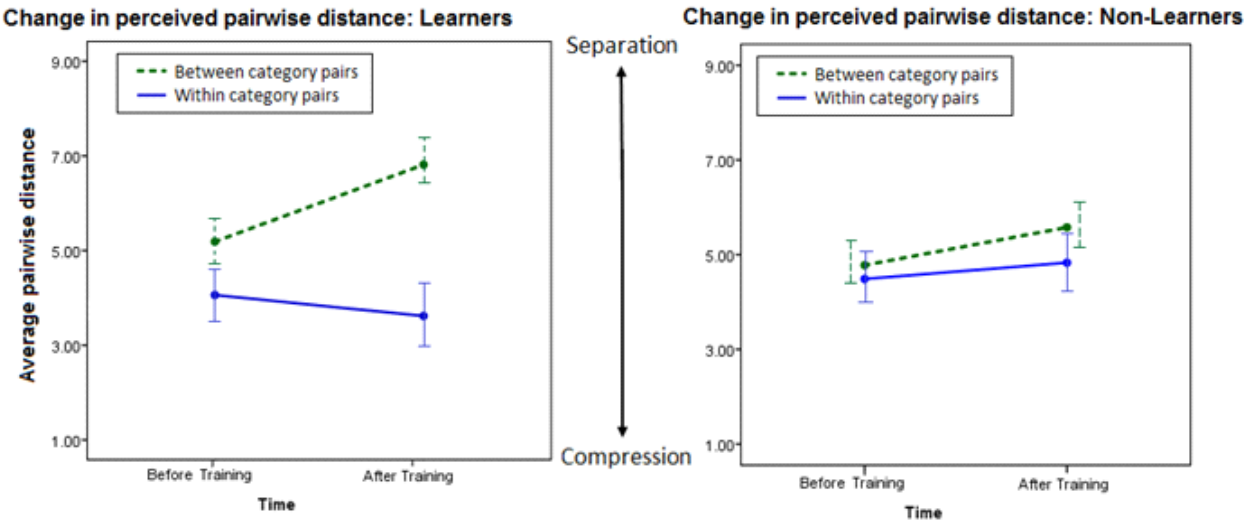

Figure 5. **Changes in average pairwise distance in learners and non-learners in Experiment 2**. Between categories (green-dashed) and within categories (blue-solid) similarity ratings before training compared to after training, for learners (left) and non-learners (right), averaged across the two difficulty levels. Learners showed significant between-category separation as well as within-category compression (the separation more pronounced than the compression). Non-learners showed no significant changes. Error bars represent ±2 SE., and those for between category pairs scores in the non learners have been purposefully displaced to displaced towards the outside of the graph to avoid overlapping with the error bars from within category pairs.

A difference between the easier and the harder condition was found for Successful Learners'. While there was significant between-category separation in both the easier (mean diffB = 1.82), $t(15) = 5.70$, $p < 0.001$, Cohen's $d = 1.32$, and the harder conditions (mean diffB = 1.35), $t(7) = 3.36$, $p = 0.012$, Cohen's $d = 1.194$, within-category compression was only significant in the easier condition (mean diffW = -0.89), $t(15) = -2.48$, $p = 0.025$, Cohen's $d = 0.87$ (Figure 6). The harder condition showed only a small, non-significant separation for within category pairs (mean diffW = 0.46), $t(7) = 1.17$, $p = 0.282$, Cohen's $d = 0.43$. These results corroborate the existence of the separation effect in both conditions.

Repeated-measures ANOVAS with difficulty as a between-subject factor found no significant hard/easy difference in between-category separation, $F(1,22) = 0.66$, $p = 0.426$, partial $\eta^2 = 0.028$, Observed Power = 0.12, but there was a significant hard/easy difference in within-category compression, $F(1,22) = 5.53$, $p = 0.028$, partial $\eta^2 = 0.194$, Observed Power = 0.62.

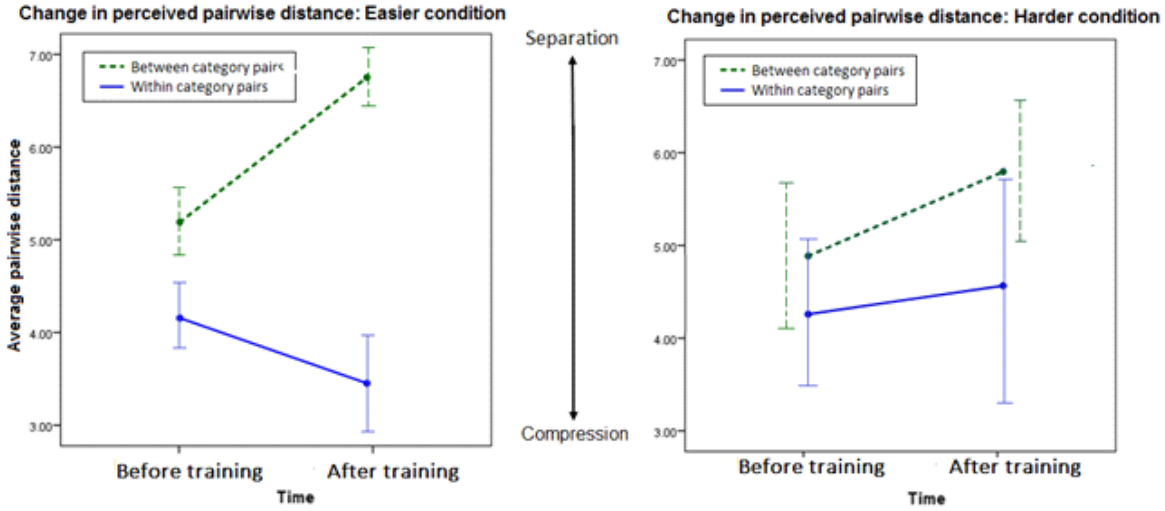

Figure 6. **Changes in average pairwise distance in the easier and harder task (in learners of Experiment 2).** Average between-category (green-dashed) and within-category (blue-solid) similarity ratings, before and after training. Left: easier condition (5/6 covariant features): significant separation between and compression within categories. Right: harder condition (4/6 covariant features) significant separation between categories but no effect within categories. Error bars represent ±2 SE. Same as Figure 5, error bars in the non-learners for between category pairs have been purposefully displaced to avoid overlap.

We think the explanation for the smaller between-category separation (and absence of within-category compression) in the harder condition may be that because learners in the hard condition learned significantly later than learners in the early condition, they had less practice categorizing correctly, reducing their separation/compression effects. To test this possibility, we combined the two difficulty conditions and recalculated compression and separation using the number of trials to reach the criterion of 80% correct (rather than k/N) as the index of difficulty. A one-way ANOVA with Linear Contrast tested the effect of the number of trials to criterion on the CP effects. The main effect of number of trials on separation was not significant, $F(8,9) = 2.124$, $p = 0.142$, but there was a significant negative linear trend: the more trials it took to learn, the smaller the separation effect, $F(8,9) = 8.345$, $p = 0.032$. This was also consistent with a significant Pearson product-moment correlation between number of trials to learn and the size of the between-category separation effect in the learners, $r = -0.535$, $n = 18$, $p < 0.05$ (Figure 7). There was neither a correlation nor a linear trend for the number of trials and the compression effect. A second possible explanation derives from feature

inequality: the black and white features were not homogeneous and some of them were more salient or noticeable for our participants (see Appendix for details).

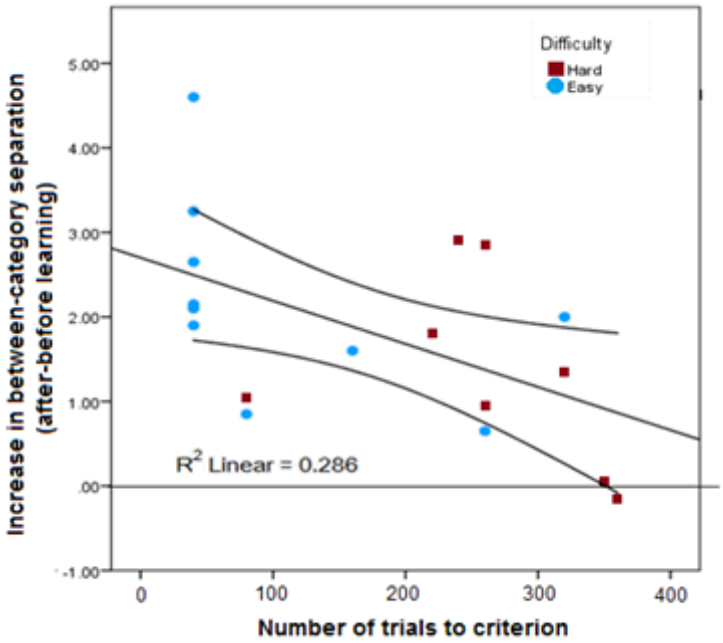

Figure 7. **Number of trials to reach criterion and degree of between-category separation after learning in Experiment 2**. Correlation between (1) increase in perceived between-category separation in learners (after learning minus before learning) and the number of trials it took them to reach criterion. The later the learning, the smaller the separation in both conditions (easy and hard).

Experiment 3

To test the stimulus characteristics that influenced our previous results, we created a new version of the textures and conducted an on-line version of the experiment to test the difficulty levels as well as the perceptual effects induced by learning to categorize them correctly.

Method

Online implementation of the experiment. Experiment 3 was implemented in Percept (Rivas & Prévost, 2016), our laboratory web platform, where behavioral experiments are built using the Django web framework, with a MySQL database contained in a VPN hosted by the laboratory. The experiment was constructed using the open source library jsPsych (De Leeuw, 2015) and written in JavaScript.

Participants. The experiment was conducted on 57 participants aged 18 - 35, recruited as in Experiments 1 and 2.

Stimuli. For this online version of the experiment, we modified our stimuli in the following way: (1) we made the components smaller and hence not individually detectable by participants, the density of our new stimuli was 300 x 300 pixels per texture, composed by 45 x 45 components of 6 x 6 pixels each; (2) we removed the specific disproportionately salient micro-component described by our previous participants (Figure 2-app); and (3) now only the number of covarying features was constant for a given

difficulty level, not the specific features themselves, which differed from participant to participant (e.g., the stimuli at difficulty level 2/5 always had two covarying features for each participantt, but which particular two features were used varied across participants; Figure 8).

Procedure. The experiment was conducted in each participants browser and would stop running if participants were idle for more than five minutes.

Tasks. We used the same three tasks as in the laboratory version of the experiment: similarity judgements before category training – category training – similarity judgements after category training. The categorization task was also divided into four blocks of one hundred trials each with pauses in which participants filled out an online questionnaire about their strategy. The similarity judgements consisted of 40 trials; participants had to respond by clicking with the cursor on a sliding bar that went from very similar to very different. The values in the sliding bar were transformed into a 0 to 100 scale.

Learning assessment. We classified our participants as "immediate learners", "learners", "borderlines" and "non-learners" following the same criteria we used in the two previous experiments.

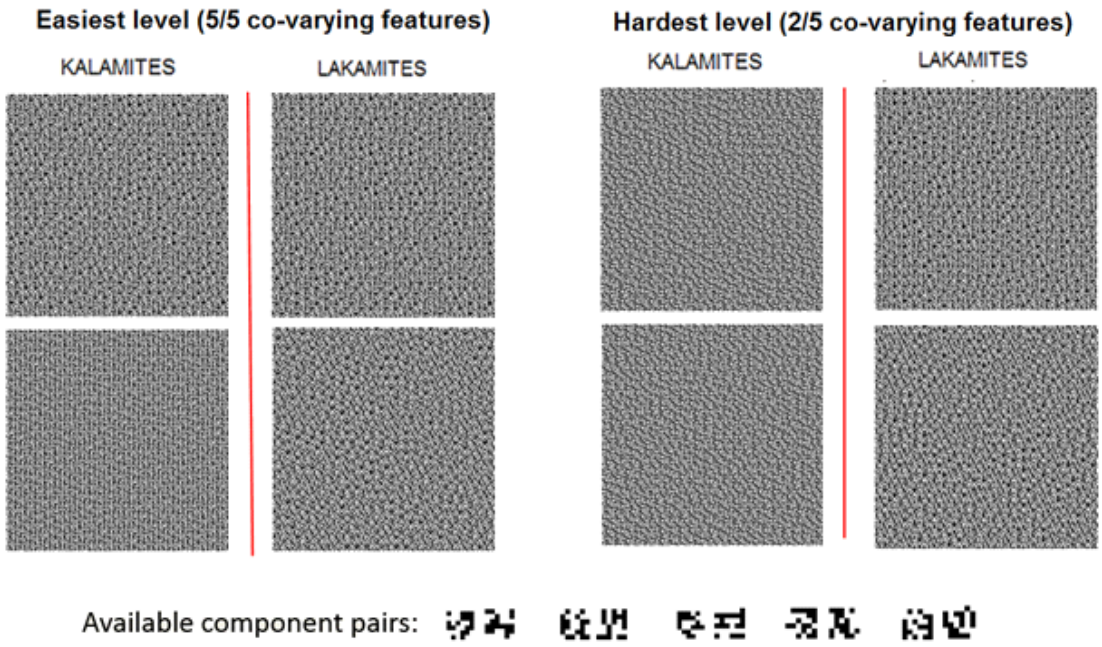

Figure 8. **Modified textures for Experiment 3 (online)**. Example of the new textures presented in the online experiments and (enlargements of) the micro-component pairs of which they are composed.

Results

Learning. Fifty-seven participants completed the online experiment, each randomly assigned to one of the four difficulty levels: 5/5(easiest), 4/5, 3/5, and 2/5(hardest). We classified our participants as "immediate learners," "learners," "borderlines," and "non-learners" following the same criteria we used in the two previous experiments. Forty-one participants were successful in learning the category, as determined by reaching our apriori criterion (four of them were immediate learners). Twelve participants did not succeed in learning and four were borderlines. The levels confirmed our assumption that fewer covarying features would make the category harder to learn, as indicated by the number of trials to criterion and the proportion of

successful learners as the difficulty increased (Figure 9). A one-way ANOVA confirmed a significant effect of difficulty on the number of trials before reaching the learning criterion, $F(3, 37) = 4.620$, $p = 0.008$.

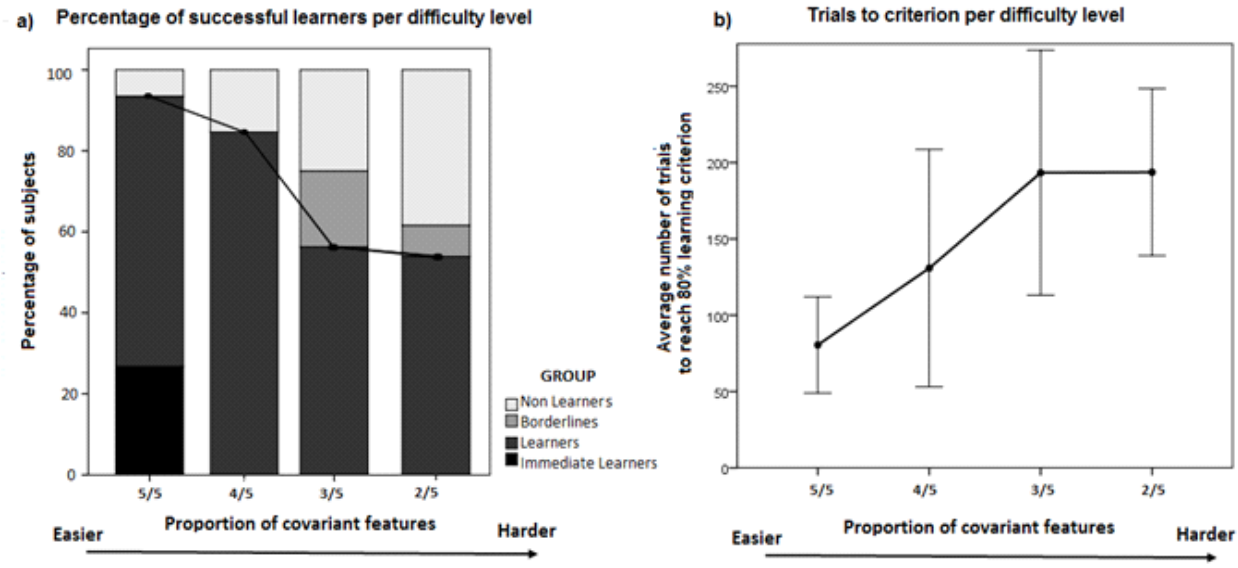

Figure 9. **Indicators of difficulty in Experiment 3**. On the left, Figure 9a) shows the proportion of learners decreases with difficulty. On the right, figure 9b) shows the number of trials it took the learners to reach it. The fewer the covariant features, the harder to learn the categorization. Error bars represent ±2 SE.

Similarity judgments. Combining the data for all four levels of difficulty reveals significant between-category separation, positive dB, $t(40) = 9.073$, $p < 0.01$, as well as within-category compression, negative dW, $t(40) = -5.512$, $p < 0.01$, for learners. Non-learners again showed only a small pairwise separation both between and within categories (Figure 10 as in Figure 5), significant between, $t(11) = 3.237$, $p = 0.006$, non-significant within, $t(11) = 1.705$, $p = 0.109$. We think this significant separation effect in the non-learners is likely to be a passive-exposure effect (unsupervised learning) in which, after being exposed to the stimuli, everything becomes slightly more distinct from everything else (Mackintosh & Bennett, 2014). We ran repeated measures ANOVAS testing Compression and Separation with learner/non-learner as the between-subject factor. The interaction with the learning group (learner or non-learner) for within-category compression was significant, $F(1, 55) = 18.655$, $p < 0.01$, whereas for between-category separation, $F(1, 55) = 0.592$, $p = 0.445$, it was not.

CP effect and task difficulty. A one-way ANOVA with Linear Contrasts tested the effect of task difficulty on separation and compression for learners. The effect of difficulty level on separation (diffB), $F(3, 37) = 1.95$, $p = 0.138$, was not significant, but it showed a siginificant weighted linear trend, $F(3, 37) = 5.13$, $p = 0.029$; for compression (diffW), difficulty had no significant effect, $F(3,37) = 1.30$, $p = 0.289$, and there was no linear trend, $F(3, 37) = 0.327$, $p = 0.629$ (Figure 11). The linear trend for separation, however, was downward as difficulty increased: this was consistent with what we had found in Experiment 2 but it was once again contrary to what we had predicted (i.e., increased separation with increased difficulty). Varying which of the features were the covariant ones within each level reduced the likelihood that this contrary outcome was an artefact of unintended inequalities in the salience of the features, individually or jointly, but it did not

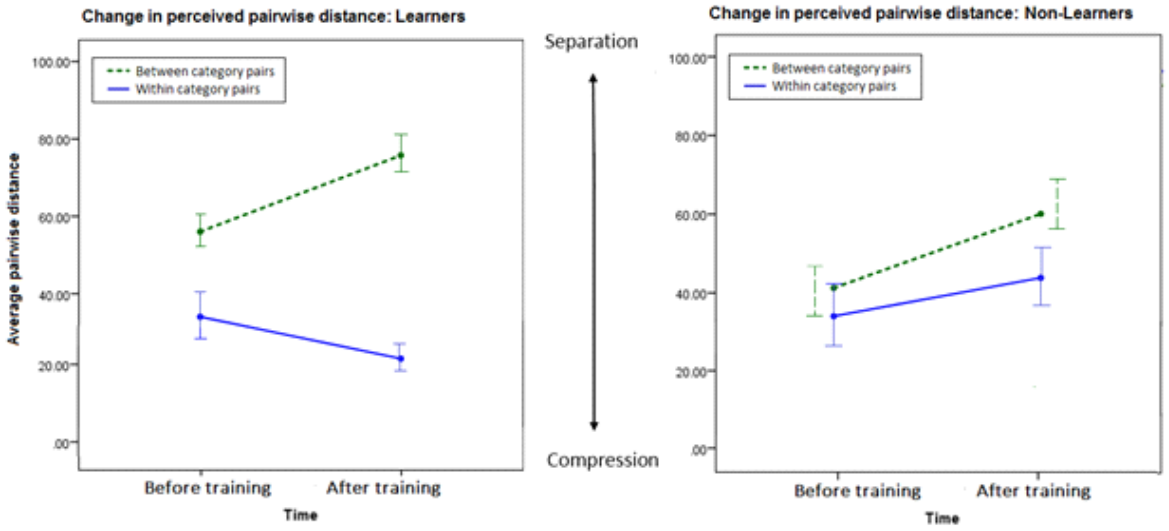

Figure 10. **Changes in perceived pairwise distances from before to after categorization training in Experiment 3**. All four difficulty levels combined. Left: Learners had both significant between-category separation and significant within-category compression. Right: Non-learners showed a small separation both between categories (significant) and within (non-significant), as in Figure 5. Error bars represent ±2 SE. Error bars for between-category pairs for non-learners have been displaced one more time to avoid overlap with within category pairs scores.

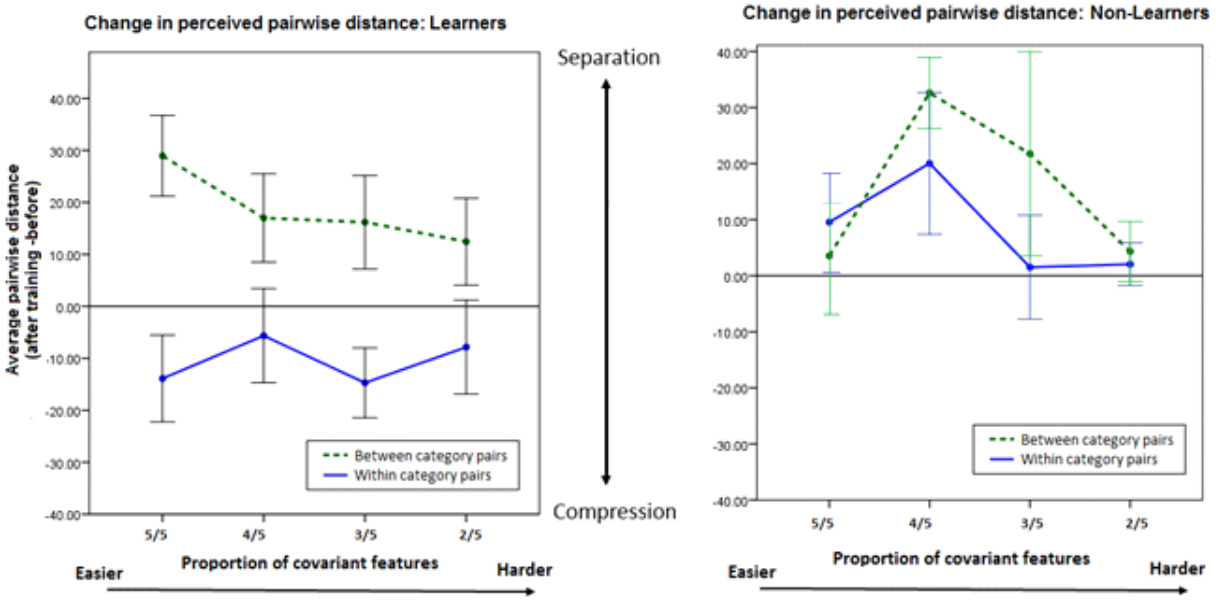

Figure 11. **Effect of learning difficulty on changes in perceived pairwise distance in Experiment 3**. Left: Learners have a significant downward linear trend for between-category separation: They show separation between categories at all difficulty levels, but it decreases when difficulty increases (i.e., separation is smaller when there are fewer covariant features). Learners show within-category compression at all levels, with no linear trend. Right: Non-learners show separation both within and between categories at all levels, with no linear trend related to difficulty level. Error bars represent ±2 SE.

eliminate it entirely. A more likely explanation is the second of the two proposed earlier: higher levels of difficulty mean later learning, fewer successful trials, and hence weaker learning in the late learners than the more practiced earlier learners. This is illustrated in Figure 12. In ongoing longer-term studies we are extending the training trials to reduce the difference between early and late learning. The issue of feature inequality is analyzed further in the Discussion.

As in Experiment 2, we combined the data for learners across the four conditions and tested the effect of the number of trials to criterion on the size of the separation effect using a one-way ANOVA with Linear Contrasts. There was no significant main effect, $F(17,19) = 1.112$, $p = 0.409$, but there was again a significant linear trend downward: the more trials it took to learn, the smaller the degree of separation $F(17,19) = 9.938$, $p = 0.005$. This was also apparent in the Pearson product-moment correlation between the number of trials to criterion and the size of the separation effect, $r = -0.512$, $p = 0.01$ (Figure 12).

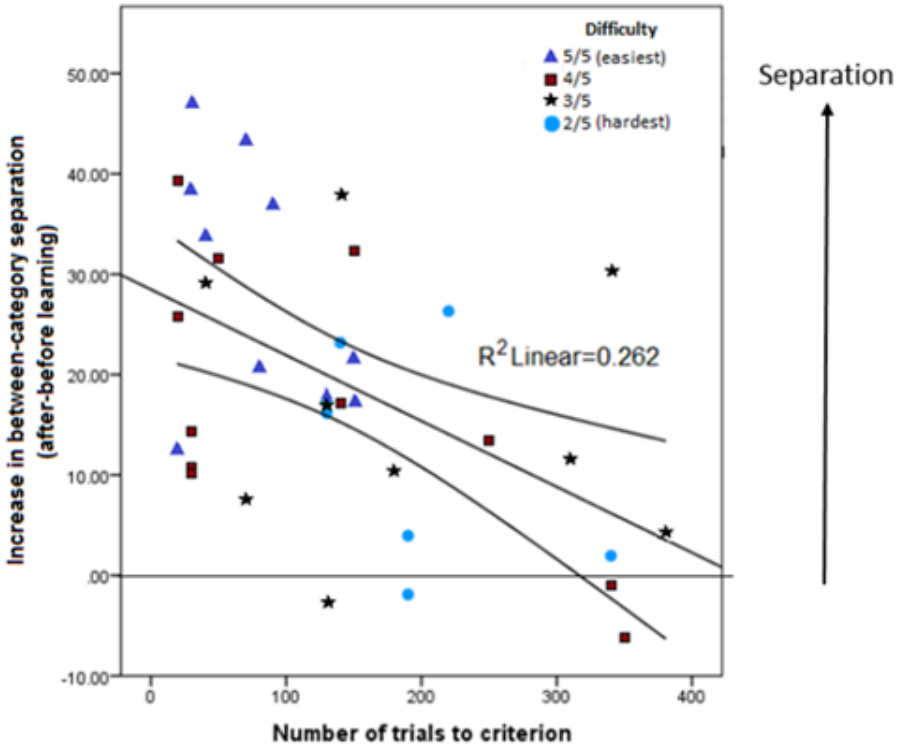

Figure 12. **Number of trials to reach criterion and degree of between-category separation after learning in Experiment 3.** Correlation between (1) increase in perceived between-category separation in learners (after learning minus before learning) and the number of trials to reach criterion. As in Experiment 2 (Figure 7), the later the learning, the smaller the separation in both conditions (easy and hard).

Experiment 4: Neural Net Model for Experimental Outcomes

Method

Building on prior simpler models (Harnad, Hanson, & Lubin 1995; Tijsseling & Harnad, 1997; Damper & Harnad, 2000), we modeled the category learning task with a general deep learning neural network (LeCun et al., 2015). The input consisted of binary vectors $\mathbf{x} \in \{0,1\}^N$ of dimension, $N = 20$ that had to be sorted into two categories according to the k binary features that covaried with category membership. As in the human experiments, we varied the proportion k/N of covarying features.

$$\mathbf{x} = [0\ 1\ 0\ 1\ 1\ 1 \ldots 0\ 1].$$

Each category is initialized by randomly creating two orthogonal binary vectors. At difficulty level k/N, stimuli are generated by randomly flipping the binary value of the N-k non-covarying components that are irrelevant to category membership. The other k components covary with category membership.

The neural network model was an unsupervised auto-encoder (Figure 13) feeding into a supervised classification layer (Figure 14) (Bengio, Courville, & Vincent, 2013). Through auto-association, this net first learns to generate as output the training stimuli it receives as input. Through exposure, the network learns a representation from which it is able to regenerate learned examples **x** when presented with partial, incomplete or noisy examples $\hat{\mathbf{x}}$

The forward and feedback activation of layer **h** and layer **x** are respectively given by

$$\mathbf{h} = f(\mathbf{W}\mathbf{x} + \mathbf{b}_\mathbf{h})$$

and

$$\mathbf{x} = f(\mathbf{W}^\mathbf{T}\mathbf{h} + \mathbf{b}_\mathbf{x})$$

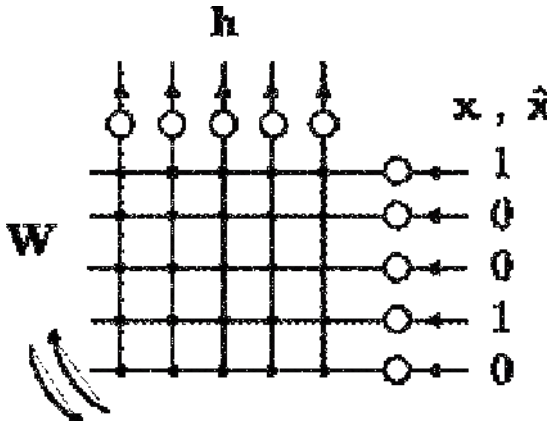

Figure 13. **General auto-encoder architecture.**

where f is a nonlinear activation function, **W** is the connection weights between the layers, and **b** is an activation bias. The connection weights are trained by repeated presentation of noisy examples $\hat{\mathbf{x}}$. During the learning process, **W** is progressively modified to minimize the error $E = \left\| \mathbf{x} - f(h(\hat{\mathbf{x}})) \right\|^2$ obtained when the network is attempting to regenerate examples **x** from noisy examples $\hat{\mathbf{x}}$.

Once the auto-encoder has learned a proper space of representation, these are passed forward to a second level of representation where the categories are learned using error-corrective feedback. On each trial the second layer of connection weights $\mathbf{W}_2$ is modified through gradient backpropagation based on the net's error -- the difference between its output and the correct category name -- strengthening correct connections and weakening incorrect ones.

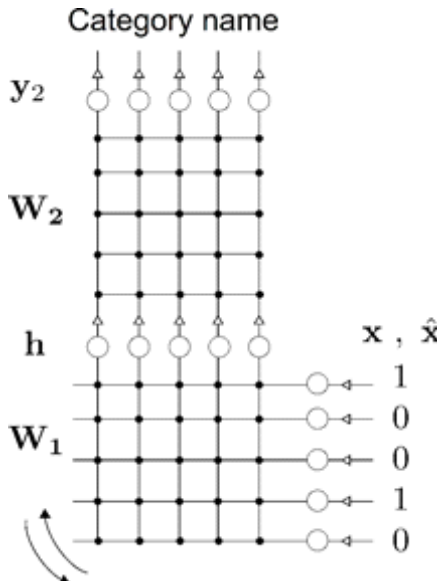

Figure 14. **General architecture of auto-encoder network feeding into a categorizer layer.**

Each experiment consists of training and testing the network with 600 examples of binary vectors $\mathbf{x} \in \{0,1\}^N$ with dimension N=20, as defined above, for two disjoint categories. A total of 50 experiments was run, each time with new randomized values defining the categories and the initial network connection weight matrices $\{ \mathbf{W_1}, \mathbf{W_2} \}$. For each experiment, the auto-encoder and the classification layers were trained until a 98% success rate was reached. The averaged categorization results over all experiments are summarized below. Figure 15 shows that with all other learning parameters fixed, the number of trials needed to learn to categorize successfully (i.e., the level of difficulty) increases as the proportion of co-varying dimensions (binary 0/1 features) decreases.

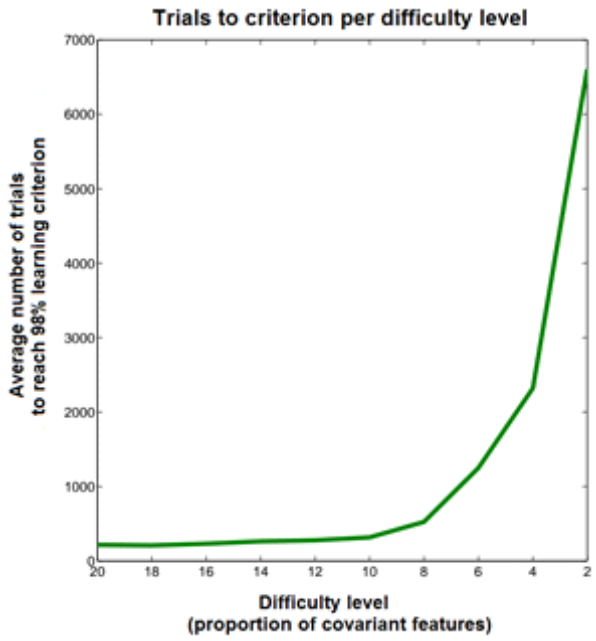

Figure 15. **Model predictions.** Number of learning trials increases as difficulty of the task increases (proportion of covariant features decreases) (Cf. Figure 9, Right.).

As in the laboratory experiments with human participants, the number of trials to reach criterion (which was 80% for the human participants and 98% for the nets) increased as the task was made more difficult by decreasing the proportion of the binary features that covaried with category membership (Figure 15). The important differences were that the features in the human experiments were two-dimensional visual textures composed out of a random distribution of 12 paired 6x6 (or 5 x 5) pixel patterns of black and white squares whereas for the nets they were merely digital 0/1 components of a one-dimensional input vector. In the human experiments, N was 6 and the proportion of covariants k/N tested ranged from 6/6 to 3/6 (in Experiments 1 and 2, and 5/5 to 2/5 in Experiment 3) whereas for the nets N was 20 and the proportion of covariants k/N ranged from 20/20 to 2/20. The most important difference was that being just binary 0 and 1 components of the input vector, the binary features of the stimuli for the nets were completely uniform and identical, whereas the paired micro-components of the visual textures that served as the input to the humans (Figure 2) were not uniform and identical, even though they all had the same number of black and white components.

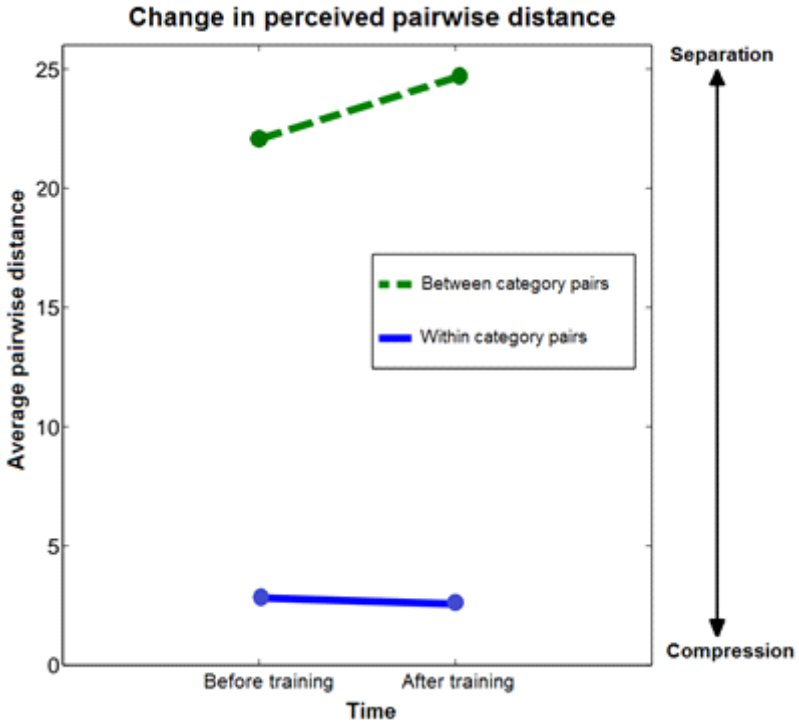

Figure 16. **Model predictions**. Pairwise distances within and between categories, averaged across all difficulty levels (20/20 – 2/20), before and after learning the categories (cf. Figure 7, Left, and Figure 10, Left).

As a measure of the pairwise distances between and within categories for the network, we calculated the pairwise vector distance between inputs, using as coordinates the activations on each of the three hidden units in the net's internal representation, averaged across all 600 inputs. We then compared these distances before learning (but after auto-encoding) with after learning the category, within and between each of the two categories.

As in the human experiments, category learning resulted in between-category separation and within-category compression -- the learned CP effect (Figure 16). The neural network simulations were repeated many times, varying many network parameters, but they almost always produce the same profile of between-category separation and within-category-compression. (Note that the separation/compression outcome was not imposed on the simulation: the net only had to learn to categorize. The CP effect was an observed correlate of having learned the category.)

Although the model produced the same pattern of results as the human experiments for (1) the increasing number of trials to reach criterion as the decreasing proportion of covariants increased task difficulty as well as for (2) the learning-induced separation/compression averaged across all the levels of difficulty (the learned CP effect), the model diverged from the human experiments in the correlation between task difficulty and the size of the CP effect. For the net, separation/compression increased with task difficulty (as had been predicted for the human participants; Figure 17) whereas for the human participants the relationship was

inconsistent, with a trend in the opposite direction (Figure 11, Left). We think the cause of the discrepancy is the feature inhomogeneity in the human visual stimuli, as analyzed further in the Discussion.

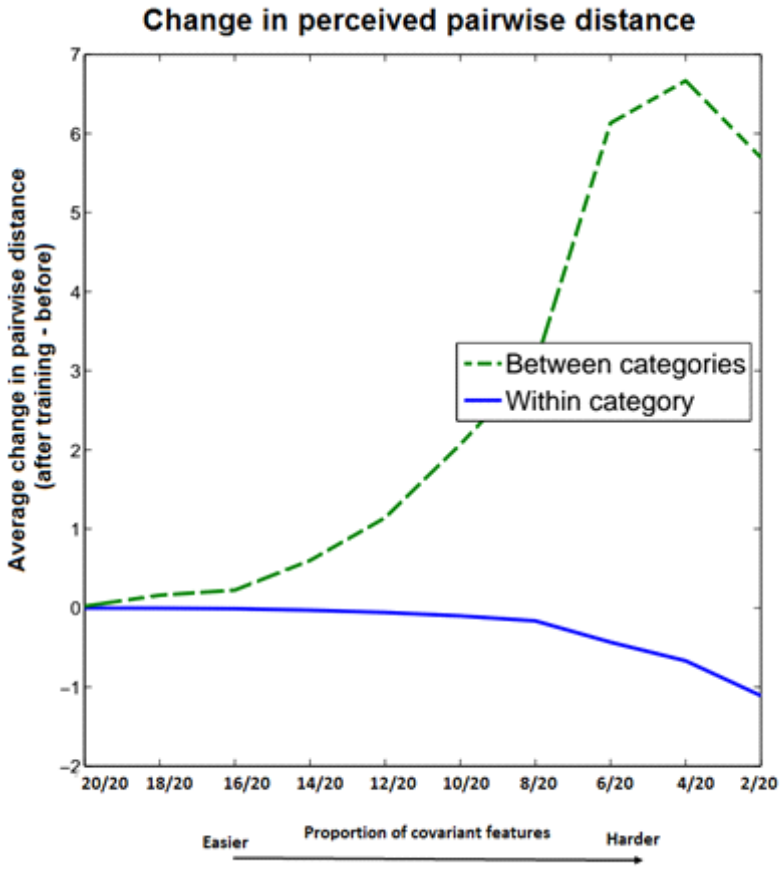

Figure 17. **Model predictions**. Pairwise distances increase between categories and decrease within categories, as proportion of covariant features decreases (difficulty increases) (Cf. Figure 11, Left.).

Discussion

Perceptual Effects of Category Learning

To categorize is to do the right thing with the right kind of thing. To learn to categorize requires detecting the features that distinguish the members of the category from the non-members: the features that covary with category membership. The lower the proportion of covariant features, the harder it is to learn the category, as indicated by the fact that it requires more trials with corrective feedback to learn successfully, and fewer participants succeed in reaching the learning criterion. We observe a perceptual change that occurs in successful learners only and is absent in those participants who failed to learn the category with the same number of training trials. The change is categorical perception (CP): differences between members of different

categories come to look bigger after successfully learning to categorize them compared to before (between-category separation), and differences between members of the same category (sometimes) look smaller (within-category compression). This same CP separation/compression effect as a result of category learning also occurs in our neural net model, but the model also shows a strong and consistent negative correlation between the size of the CP effect and the proportion of covariant features.

Dimensional Reduction

The neural net model suggests an explanation of why category learning causes categorical perception: dimensional-reduction (Edelman & Intrator, 1997; Folstein et al., 2012, 2014; Tenenbaum et al., 2000). The N binary features can be treated as dimensions of a (discrete) N-dimensional similarity space. Initially, the pairwise distance between inputs (which are points in this discrete space) is based on all N binary dimensions (i.e., the ordinary Euclidean distance between N-component vectors). Category learning occurs when the learner has successfully learned to detect the k covariant features and to ignore the N-k non-covariant features. This reduced k-dimensional subspace of the original N-dimensional similarity space acts as a kind of feature filter: Inputs (now k-component vectors rather than N-component vectors) from different categories are separated on the k covariant dimensions while inputs from the same category are compressed on them, compared to their prior distances in the full original N-dimensional similarity space.

Our neural nets, whose inputs are simply a vector of N binary features, show this pattern very clearly: The CP effect -- between-category separation in hidden-unit activation space (and, to a lesser degree, within-category compression) -- increases as the proportion of covariant features k/N decreases (i.e., CP grows as k/N shrinks). In our human experiments, too, significant CP (between-category separation and sometimes within-category compression) occurs after successful learning in almost all cases. The same linear trend as in the nets is also there in the human participants for the negative correlation between k/N (i.e., difficulty) and the number of trials required to learn the category. The correlation between difficulty and the size of the CP effect, however, is either absent or the linear trend is in the opposite direction (i.e., CP shrinks as k/N shrinks) in the human participants.

Feature Homogeneity and Practice Time

We think this partial discrepancy between the human performance and the neural net model is due in part to inequalities in the detectability and salience of the visual features for the human visual system. Unlike in an artificial neural network, it is very difficult to make multiple visual features equally distinct to the human brain. A further difference between the humans and the nets is that nets can be trained much longer, to much higher criteria (98% accuracy) whereas our experimental participants had only 400 trials to reach 80% accuracy. The human difficulty levels therefore had a confounding factor: Because reaching the 80% criterion at the higher levels of difficulty occurred later in the one-hour training session, it was also based on fewer successful learning trials. The weaker CP effect that we observed with higher difficulty may hence have been a consequence, not of the higher difficulty itself, but of the smaller number of remaining trials to practice and consolidate the learning once it had taken place, as seen in the significant negative correlation between the number of trials to learn and the size of the separation effect in both Experiment 2 and Experiment 3.

Conjunctive vs. Disjunctive Covariance

Another potential explanation of the partial difference between the human and net outcome for task difficulty could be related to conjunctive vs. disjunctive covariance: Did learning reduce dimensionality to k

or to 1? The covariance in this study was conjunctive: All k of the N covariant binary features had one value for the Kalamite category and the other value for the Lakamite category. Hence although it became harder to find the k covariant features as k/N became smaller, it was still true that, once found, any one of the k dimensions would have been enough all by itself to serve as a basis for categorizing correctly. In other words, both the human participants and the nets really only needed to find one out of the k covariants, regardless of the size of k. To analyze this further in future studies, we will test disjunctive covariants, in which the presence of any one (or more) of the k dimensions covaries with the category. Task difficulty should then correlate positively instead of negatively with k (because the bigger the k, the more independent features S must learn to detect), but the size of the CP effect (dimensionality reduction) should correlate negatively with k (because the smaller the k, the fewer the covariant dimensions).

Overlearning

The next step in this research will be a series of long-term studies with extended overlearning trials (Shibata et al., 2016; Smolen et al., 2016), spaced across weeks (in place of the trials massed in a single one-hour training session as in Experiments 1-3), with a much larger sample of inputs and a sustained learning criterion of 98% instead of 80%. Our expectation is that under such extended overlearning conditions and with a much more exacting learning criterion all participants will become successful learners eventually, with a robust category pop-out effect (Al-Rasheed, 2016) mediated by the feature filter underlying CP. We will also test whether the size of the (overlearned) CP effect itself correlates negatively with the k/N ratio (dimensional reduction), as in the neural net model (Figure 17).

Name Bias?

Is the separation/compression in similarity judgments really a perceptual effect or just a response bias from having learned the category name (a tendency to rate textures as more different when their names are different and more similar when their names are the same)? This frequently asked question cannot be answered rigorously on the basis of similarity judgments alone. It also needs to be tested by measuring psychophysical discriminability (rather than just similarity judgment), between and within categories, before and after category learning, using signal detection analysis to distinguish changes in discriminability (d') from response bias (B) (Goldstone & Hendrickson, 2009; Wood, 1976). In such an analysis the name can generate a response bias, B, toward one category or the other, but it cannot affect the discriminability parameter d', which is independent of B and measures a limit on the participants capacity to perceive differences. We are currently conducting such discriminability studies.

Electrophysiological Correlates

In ongoing work (Pérez-Gay et al., 2016) we have observed a late positive component of the visual event-related potential (ERP) that is present in participants who have successfully learned the texture category and absent in participants who have not. Late ERP components are thought to be cognitive and decisional (Barceló & Cooper, 2017). Early visual ERP components (such as N1) are correlated with perceptual discrimination (Balas & Conlin, 2015; Vogel & Luck, 2000). In participants who have successfully learned the texture category there is also an N1 effect that is absent in participants who have failed. The size of this early sensory component is also positively correlated with the size of the CP separation effect. This finding too makes it more likely that the separation/compression is a reflection of the perceived distinctness of the members of the two categories, rather than just a bias from the distinctness of their names.

Verbal and Non-Verbal Learning

In order to know whether its name is Kalamite or Lakamite, participants first have to learn to detect whether each stimulus is a member of category K or L. That is what changes between our successful learners' similarity judgments before and after learning the category (and hence its name). It should be noted that, as described in the methodology of the present study, although the names of the categories were Kalamites and Lakamites, when participants were doing the supervised trial-and-error learning, they were just pressing on the K and L keys of the keyboard in response to individual stimuli. Hence it is not clear how much of a role the overt naming of the category, rather than just a differential motor response, played in the learning itself (cf. Holmes & Wolff, 2012). After (successful) learning, the similarity judgments were made with a cursor on a sliding scale, in response to pairs of stimuli. It is not clear how much participants were likely to covertly name the category of each stimulus when comparing pairs for similarity.

Detecting Covariants or Memorizing Examples?

Did our participants really learn the covariance or did they just remember some special cases? With a learning criterion of 80% there is the possibility of partial learning of a remembered subset of the stimuli rather than the covariance underlying all cases. With training extended across days and weeks and an overlearning criterion of 98% this outcome is far less likely. The role of overt and covert naming vs. nonverbal motor responses can also be tested in long-term overlearning studies.

Categories and Language

Our findings provide support for the weak version of the Whorf Hypothesis (Wang, 2016), according to which language, in the form of learning to categorize and name things – as in categorizing novel textures as "Kalamites" and "Lakamites" – subtly changes how those things are perceived. The truly revolutionary power of language, however, is cognitive rather than just perceptual, and becomes possible after a sufficient number of category names has been perceptually "grounded" through learned feature-detectors and their resulting perceptual separation/compression (Blondin-Massé et al., 2013; Cangelosi et al., 2002; Vincent-Lamarre et al., 2016). For then the names of our perceptually grounded categories can go on to be combined and recombined into propositions that define still further categories verbally, allowing the members of our species (only) to transmit and receive categories across time and space through words alone, freed of the need to undergo the laborious, time-consuming (and potentially risky) direct perceptual learning through trial and error undergone by our laboratory participants, our neural nets and all species other than human.

References

Al-Rasheed, A. S. (2016). Categorical perception of color is lateralized to the left hemisphere: Evidence from present-absent and pop-out tasks. *International Journal of Social Science and Humanity*, *6*, 113–118.

Balas, B., & Conlin, C. (2015). The visual N1 is sensitive to deviations from natural texture appearance. *PloS One*, *10*(9), 1–115. doi: e0136471

Barceló, F., & Cooper, P. S. (2017). An information theory account of late frontoparietal ERP positivities in cognitive control. *Psychophysiology.* doi: 10.1111/psyp.12814

Baumann, N., & Kuhl, J. (2005). Positive affect and flexibility: Overcoming the precedence of global over local processing of visual information. *Motivation and Emotion Journal*, *29*, 123–134. doi:10.1007/s11031-005-7957-1

Bateson, P. (2000) What must be known in order to understand imprinting? In C. Heyes & L. Huber (Eds.), *The evolution of cognition*. Cambridge, MA: MIT press.
Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. PAMI, special issue Learning Deep Architectures*, *35*, 798–828. doi: 10.1109/TPAMI.2013.50.
Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, *111*, 4590–4595.
Blondin-Massé, A., Harnad, S., Picard, O., & St-Louis, B. (2013). Symbol grounding and the origin of language: From show to tell. In C. Lefebvre, B. Comrie, & H. Cohen (Eds.), *Current perspective on the origins of language.* Amsterdam, The Netherlands: John Benjamins Publishing Company.
Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT press.
Cangelosi, A., Greco, A., & Harnad, S. (2002). Symbol grounding and the symbolic theft hypothesis. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language*. London, UK: Springer.
Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Özgen, E., & Davies, I. R. (2012). Neural correlates of acquired color category effects. *Brain and Cognition*, *80*, 126–143.
Damper, R. I., & Harnad, S. (2000). Neural network modeling of categorical perception. *Perception and Psychophysics 62*, 843–867.
De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning and Memory 15*, 717–727.
De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavioral Research Methods, 47*(1), 1–12. doi: 10.3758/s13428-014-0458-y.
Edelman, S., & Intrator, N. (1997). Learning as formation of low-dimensional representation spaces. *Psychology of Learning and Motivation 36*, 353–380.
Fajen, B. R., & Turvey, M. T. (2003). Perception, categories, and possibilities for action. *Adaptive Behavior*, *11*, 276–278.
Fink, G. R., Marshall, J. C., Halligan, P. W., & Dolan, R. J. (1999) Hemispheric asymmetries in global/local processing are modulated by perceptual salience. *Neuropsychologia* , *37*, 31–40.
Fisher, D. H., Pazzani, M. J., & Langley, P. (Eds.). (2014). *Concept formation: Knowledge and experience in unsupervised learning*. San Francisco, CA: Morgan Kaufmann.
Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford, UK: Oxford University Press.
Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object discrimination: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 807–820.
Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*, 814–823.
Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: The case of shape and motion. *Frontiers in Psychology, 5*(1394), 1–7.
Folstein, J., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current Directions in Psychological Science*, *24*, 17–23.
Gao, H., Cai, X., Li, F., Zhang, S., & Li, H. (2016). How the brain detects invariance and inhibits variance during category induction. *Neuroscience Letters*, *626*, 174–181.
Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.
Gillebert, C. R., Op de Beeck, H. P., Panis, S., & Wagemans, J. (2008). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. *Journal of Cognitive Neuroscience*, *21*, 1054–1064.
Goldstone, R. L. (1994a). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157.
Goldstone, R. L. (1994b). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology General*, *123*, 178–200.
Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 69–78.
Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*, 27–43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*, 116–139.
Han, S. & Chen, L. (1996) Processing of global and local properties - an analysis with event-related brain potentials. *Science in China. Series C, Life Sciences*, *39*, 179–188.
Hanley, J. R. (2015). Color categorical perception. *Encyclopedia of color science and technology* (pp. 1–6). New York, NY: Springer-Verlag.
Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition*. New York, NY: Cambridge University Press.
Harnad, S. Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modelling: Steps toward principled integration* (pp. 191-206). New York, NY: Academic Press.
Harnad S. (2003). Categorical perception. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 169-177). New York, NY: Nature Publishing Group, Macmillan.
Harnad, S. (2005). To cognize is to categorize: Cognition is categorization In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization*. Amsterdam, The Netherlands: Elsevier.
Holmes, K. J., & Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? *Journal of Experimental Psychology: General*, *141*, 439–443.
Horst, J. S., & Simmering, V. R. (2015). Category learning in a dynamic world. *Frontiers in Psychology*, *6*(46), 1–4.
Hussein, B. A. S. (2012). The Sapir-Whorf hypothesis today. *Theory and Practice in Language Studies*, *2*, 642–646.
Jacobs, G. (2013). *Comparative color vision*. Santa Barbara, CA: Elsevier.
Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., & Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron*, *53,* 891–903.
Jitsumori, M., & Delius, J. D. (2008). Object recognition and object categorization in animals. In T. Matsuzawa (Ed.), *Primate origins of human cognition and behaviour*, (pp. 269–293). Tokyo, Japan: Springer.
Kang, X. (2014*). Categorization difficulty increases categorical perception*. (Unpublished masters disseration). University of Southampton, Southampton, UK.
Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*, 65–79.
Kimchi, R. (1992). Primacy of wholistic processing and global/local Paradigm: A critical review. *Psychological Bulletin*, *112*, 24-38.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.
Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*, 732–753.
Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*, USA, *2*, 1-8. doi: 10.1109/ICCV.1999.790410
Marino, L. (2017). Thinking chickens: A review of cognition, emotion, and behavior in the domestic chicken. *Animal Cognition*, *20*, 127–147.
Mackintosh, N. J., & Bennett, C. H. (2014). Perceptual learning in animals and humans. In J. Adair, K. Dion, L. Belanger, & M. Sabourin (Eds.), *Advances in Psychological Science* (Vol. 2, pp. 317–333). Montreal, Quebec, Canada: Psychology Press.
McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets?. *Memory & Cognition*, *6*, 462–472.
Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.
Navon, D. (1977) Forest before trees: The precedence of global features in visual perception, *Cognitive Psychology, 9*, 353–383.
Newell, F. N., & Bülthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition, 85*, 113–143.
Ng, T., & Beeman, M. (2017, March). Selective attention to global stimuli induces analytic problem solving. *Proceedings of the 24th Cognitive Neuroscience Society Annual Meeting*, San Francisco, CA.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
Notman L. A., Sowden P. T., & Ozgen E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, *95*, B1–B14.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual Perception, 155*, 23–36.
Ozgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: a perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General*, *131*, 477–493. http://doi.org/10.1037/0096-3445.131.4.477
Ozturk, O., Shayan, S., Liszkowski, U., & Majid, A. (2013). Language is not necessary for color categories. *Developmental Science*, *16*, 111-115.
Pagin, P., & Westerståhl, D. (2010). Compositionality I: Definitions and variants. *Philosophy Compass*, *5*, 250–264.
Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.
Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 2–10. doi:10.3389/neuro.11.010.2008
Pérez-Gay, F, Sabri, H, Rivas, D., Gregory, M., Morgan, R., Botero, N., & Harnad, S. (2016, April). Changes in event-related potentials induced by category learning. *23rd Annual meeting Cognitive Neuroscience Society*. New York, NY. https://doi.org/10.13140/RG.2.1.1780.5204
Pevtzow, R., & Harnad, S. (1997) Warping similarity space in category learning by human subjects: The role of task difficulty. *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization* (pp. 189–195). Edinburgh, Scotland: Department of Artificial Intelligence, Edinburgh University.
Pomerantz, J. R. (1983). Global and local precedence: Selective attention in form and motion perception. J*ournal of Experimental Psychology: General*, *112*, 16-40.
Porter, R., & Winberg, J. (1999). Unique salience of maternal breast odors for newborn infants. *Neuroscience & Biobehavioral Reviews*, *23*, 439–449.
Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology, 67*, 2425–2438.
Shibata, K., Machizawa, M., Walsh, E., Bang, J. W., Sasaki, Y., & Watanabe, T. (2016). Overlearning of a visual task makes the learning rapidly hyper-stabilized to protect it from being overwritten by training on a new task – A new role of overlearning since 1885. *Journal of Vision*, *16*, 1097–1097.
Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318–320.
Simanova, I., Francken, J. C., de Lange, F. P., & Bekkering, H. (2016). Linguistic priors shape categorical perception. *Language, Cognition and Neuroscience*, *31*, 159–165.
Smith, A. O., & Rangarajan, A. (2016). A category space approach to supervised dimensionality reduction. *arXiv preprint arXiv*:1610.08838.
Smith, C. L. (2017). Referential signaling in birds: The past, present and future. *Animal Behaviour*, *124*, 315–323.
Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
Smith, J. D., Zakrzewski, A. C., Johnson, J. M., Valleau, J. C., & Church, B. A. (2016). Categorization: The view from animal cognition. *Behavioral Sciences*, *6*, 12. doi:10.3390/bs6020012
Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: Mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, *17*(2), 77–88.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.
Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.
Thompson, R. K., & Oden, D. L. (2000). Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. *Cognitive Science*, *24*, 363–396.
Tijsseling, A., & Harnad, S. (1997). Warping similarity space in category learning by backprop nets. *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization* (pp. 263–269). Edinburgh, Scotland: Department of Artificial Intelligence, Edinburgh University.
Van Gulick, A. E., & Gauthier, I. (2014). The perceptual effects of learning object categories that predict perceptual goals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1307-1320.
Vincent-Lamarre, P., Blondin Massé, A., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science, 8*, 625–659.
Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology*, *37*, 190–203.

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaißert, N. (2013). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin & Review*, *21*, 976–985.

Wang, Y. (2016). Do we know more about Whorf?. *International Journal of Applied Linguistics and English Literature*, *5*, 215–223.

Ward, L. (1982) Determinants of attention to local and global features of visual forms. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 562–581.

Wegbreit, E., Suzuki, S., Grabowecky, M., Kounios, J., & Beeman, M. (2012). Visual attention modulates insight versus analytic solving of verbal problems, *The Journal of Problem Solving, 4*, 6. doi:10.7771/1932-6246.1127

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. *Advances in Psychology, 77*, 241–284.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, *60*, 1381–1389.

Zarate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition*, *8*, 161–185.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K., & Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition and Behavior Reviews*, *3*, 13–45.

Zhaoping, L. (2016). From the optic tectum to the primary visual cortex: migration through evolution of the saliency map for exogenous attentional guidance. *Current Opinion in Neurobiology*, *40,* 94–102.

Appendix

Categorization Strategies

Participants described their strategies in response to an open-ended question, hence responses varied widely. We identified some common patterns and divided their detailed strategies into three main categories. We classified participants as using a "global strategy" if they reported looking for patterns in the whole texture (Figure 1-app). (Examples: "Lakamites had more waves and were more structured, Kalamites had a more random distribution of dots," "The images with well-defined lines were Kalamites and the distorted images with unclear lines were Lakamites.") participants who reported looking at a particular location inside the texture (and categorizing on the basis of the presence or absence of a feature there) were classified as using "local strategies" (Examples: "Distinguished by the presence or absence of 'L' shaped boxes"; "I tended to go with Kalamites if I saw a small square"). For all participants, but especially learners, difficulty levels 5/6 and 3/6 had a higher proportion of local strategies. Among the local strategies, some participants even drew the shape they were using to determine the category (Figure 2-app). When participants could not verbalize what they were doing (Examples: "I knew which one was right but I couldn't figure out why", "the classification became intuitive") their strategy was classified as "undetermined"

Navon, 1977 (as subsequently expanded upon by Fink, Marshall, Halligan, & Dolan, 1997; Han & Chen, 1999; Kimchi, 1992; Ward, 1982) suggested that attention to global versus local features has an influence on perceptual processing. The focus on global vs. local features has turned out to be important for many processes related to perceptual learning, such as selective attention (Pomerantz, 1983), object and natural scene recognition (Lowe, 1999; Oliva & Torralba, 2006) and flexibility (Baumann, 2005). It has also been shown recently that selective attention to global versus local features of visual stimuli can induce different types of problem solving (Wegbreit et. al., 2012; Ng & Beeman, 2017).

About two thirds of the participants used a global strategy and one third used a local strategy (2% undetermined). There was no systematic relation between the level of difficulty (proportion of covariants) and the strategy used. The proportion of learners who used a local strategy (72%) or a global strategy (63%) did not differ significantly. Another potential measure of strategy advantage, the mean number of trials to reach criterion (for learners) also showed no significant advantage for the local over the global strategy (137 trials for local vs. 195 for global; $t(23) = -1.8$, $p = 0.069$).

Among the difficulty levels, there were also differences in the proportion of participants using local and global strategies. The two sets that turned out to be the easiest (3/6 and 5/6) had a larger proportion of local strategies among the learners. In the majority of cases, the local "feature" they were using emerged from some combination of our matrices (big black blotches, white squares, etc.). At what we had expected to be the hardest (but what turned out to be the easiest) difficulty level, half the participants were classifying the textures based on the presence or absence of one single feature that corresponded specifically to one of our matrices (Figures 4 & 5).

Two factors could explain the discrepancy between the number of covarying features and the resulting difficulty for each level: (1) For the visual system, our matrix features are not equal and homogeneous. Some (such as the feature depicted in Figure 2-app) are more easily detectable than others and draw the attention of the participant. Given that our features are binary, the presence or absence of the especially salient feature in Figure 2-app signals category membership and leads to easier learning of the category. (2) In addition, for the learners using global strategies, the distributed patterns arising from the interaction of the covarying features

also made the Kalamites more discriminable at level 3/6 than at the other levels, although this stimulus set had the lowest proportion of co-varying features. This indicates that it is not only the proportion of diagnostic features that determines difficulty but also their interaction in generating distributed global patterns.

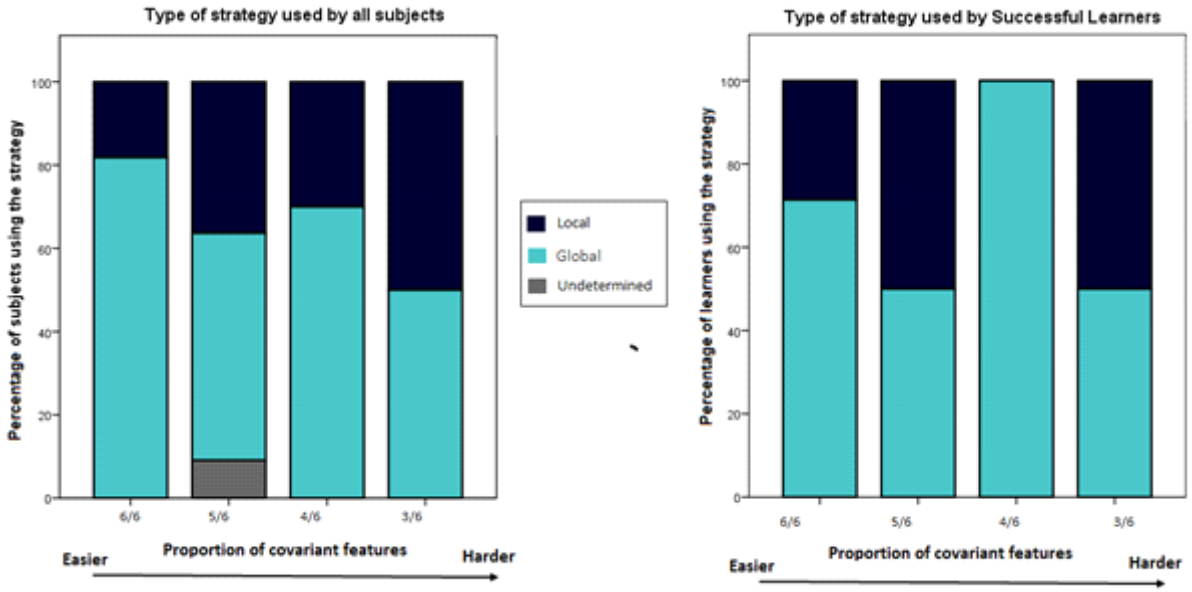

Figure 1-app. **Uneven distribution of strategies across difficulty levels**. Left: strategies across all participants (learners, non-learners, borderlines). Right: strategies of the learners only.

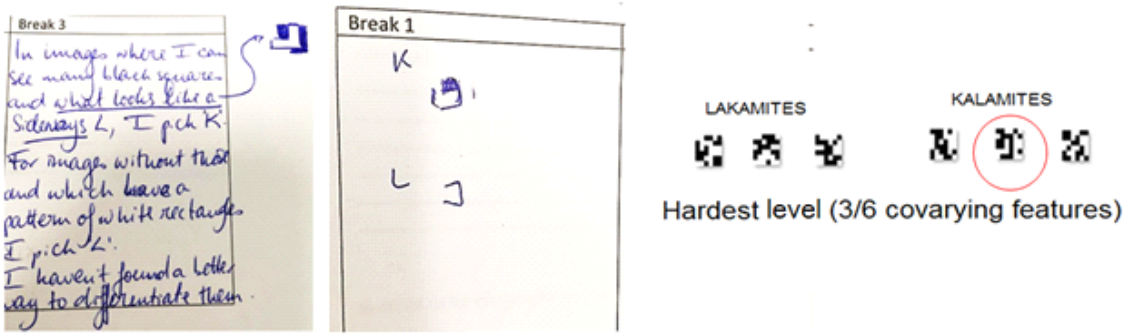

Figure 2-app. **Local strategies**. Two examples of local strategies described and drawn by two different participants assigned to difficulty 3/6 in Experiment 1. The local strategy they used corresponded exactly to detecting one of the invariant components of the KALAMITES.

Given this outcome for Experiment 1, we modified our stimuli for Experiment 3 as described under the methodology section, making our matrices smaller and randomizing which matrices were combined in each level's textures, so that only the number of co-varying matrices (features) would remain constant. After this change, all participants used global strategies, referring to distributed patterns in the textures rather than

focusing on smaller local elements (E.g.: “KALAMITES are fuzzy and scattered, LAKAMITES are fuzz with lines”, or “KALAMITES have a diagonal orientation”, or “LAKAMITES are rounder and more nebulous, KALAMITES are straighter and more structured”).

Financial conflict of interest: No stated conflicts.
Conflict of interest: No stated conflicts.

Submitted: *December 6th, 2016*
Resubmitted: *May 11th, 2017*
Accepted: *May 14th, 2017*