The effect of the internal structure of categories on perception

Todd M. Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, 6 Washington Place
New York, NY 10003 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

A novel study is presented that explores the effect that learning internally organized categories has on the ability to subsequently discriminate category members. The results demonstrate the classic categorical perception effect whereby discrimination of stimuli that belong to different categories is improved following training, while the ability to discriminate stimuli belonging to the same category is reduced. We further report a new within-category perceptual effect whereby category members that share the same category label but fall into different sub-clusters *within* that category are better discriminated than items that share the same category *and* cluster. The results show that learners are sensitive to multiple sources structure beyond simply the labels provided during supervised training. A computational model is presented to account for the results whereby multiple levels of encoding (i.e., at the item-, cluster-, and category- level) may simultaneously contribute to perception. **Keywords:** category learning; categorical perception; perceptual learning

Introduction

Categorical perception (CP) refers to the tendency of observers who have learned a category to show a reduced ability to discriminate between items belonging to that category (i.e., acquired equivalence) while showing improved discrimination (i.e., acquired distinctiveness) for items that come from different categories. For example, a listener's ability perceive differences between speech sounds appears to be influenced by the structure of phoneme categories in their native language. Physical stimulus differences near the boundary of two phonemic categories (such as */be/* and */ge/*) appear exaggerated while similar differences within the phoneme category region appear reduced (Liberman, Harris, Hoffman, & Griffith, 1957). While there is some debate about the genesis of at least some of these effects (i.e., whether they reflect innate features of perceptual organization or are the expression of learned behavior), there is a growing body of evidence suggesting that learning arbitrary categorical distinctions for a variety of visual and auditory stimuli can effectively "warp" our perceptual abilities in the service of these categorizations (Harnad, 1987; Logan, Lively, & Pisoni, 1991; Goldstone, 1994).

One factor influencing why items that belong to the same category may be seen as more similar to one another while items from different categories are seen as more distinct may center on the the fact that, after learning, category members share a new commonality by virtue of the learned category label, name, or other association that category non-members do not. Consistent with this observation is the fact that almost all studies of learned CP use a supervised training procedure where stimuli are classified on the basis of trial and error (requiring an overt response) with corrective feedback. The advantage of this procedure is that it is possible to objectively measure category learning performance. However, it remains a somewhat open question if learned CP effects are restricted to cases where subjects make a differential response to each category or if other aspects of category organization, such as the similarity structure or distribution of items within a category, may also exert an influence on perception. For example, computational models of category acquisition such as the rational model (Anderson, 1991) or SUSTAIN (Love, Medin, & Gureckis, 2004) suggest that learners are sensitive to sources of sub-category structure such as "clusters" of highly similar items even in the absence of explicit reinforcement.

The focus of the present article is to provide direct evidence for the combined contribution of both explicitly reinforced (i.e., supervised) and incidental (i.e., unsupervised) learning on categorical perception. In particular, we examine how sources of within-category structure (that are irrelevant for making a successful categorization response) influence the ability to later discriminate category members. Consistent with previous reports of CP, we find clear evidence that learners in our task become better at discriminating between items that belong to different overt categories during and following training. However, the ability to discriminate within-category differences was influenced by the internal structure of the category: items which fell in the same "cluster" of within-category items became harder to discriminate while items that shared the same overt category label, but belonged to separate "clusters" became better discriminated. These results are not anticipated by standard theoretical accounts of CP as the warping of an internal representation space under pressure to reduce categorization error (e.g., Harnad, Hanson, & Lubin, 1995) or through selective attention given to categorization-relevant information (Kruschke, 1992; Nosofsky, 1986). We conclude by presenting simulations with an extension of the SUSTAIN model that is simultaneously sensitive to multiple levels of category structure (e.g., at the item-, cluster-, and category- level) which successfully explains the pattern of results.

An Experiment

In our study, subjects were asked to discriminate items that varied along two poorly defined and arbitrary dimensions.

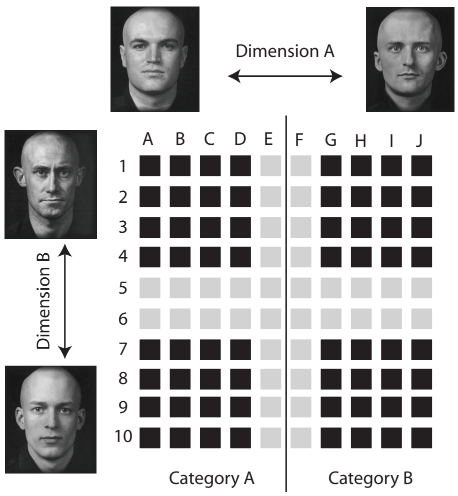

Figure 1: Stimuli varied along two arbitrary dimensions (A and B) defined by morphs sequences that interpolated between the faces shown. A 10x10 grid of faces that blended these dimensions was created, but the light gray stimuli were withheld from the category learning phase of the experiment inducing a source of within-category structure (i.e., each category was naturally defined by two distinct prototypes). The vertical line shows an example category boundary subjects were taught during the learning phase (for half the subjects this line was horizontal through the space).

Later, they learned to categorize these stimuli into two groups via trial and error with corrective feedback. Successful categorization required attention to only one of the stimulus dimensions. However, the structure internal to each of the two categories was not uniform: within each category there were two sub-clusters or sub-prototypes of items. Our hypothesis was that if the principal cause of CP effects is the addition of a shared category label or response, then we would find increased discrimination accuracy for items that varied along the category-relevant boundary, with reduced accuracy for items that belonged to the same overt category. In effect, the pressure to categorize items into groups would cause learners to generalize over the structure internal to the categories that was irrelevant for classification. Alternatively, if learners are sensitive to both the demands of the categorization task *and* the distribution of exemplars within each category (as are models such as SUSTAIN and the rational model), we might find changes to the discriminability of items within each category consistent with the induced clustering of items (despite the fact that this clustering was irrelevant for success in the categorization task).

Methods

Participants and Apparatus 120 students at Indiana University participated in partial fulfillment of a class requirement. Subjects were randomly assigned to one of two conditions based on which dimension (A or B in Figure 1) was the categorization boundary. The experiment was administered on standard Macintosh computers over a single one-hour session.

Stimuli Stimuli were constructed that varied along two arbitrary and equally salient dimensions. Each dimension was created by taking two bald male faces as endpoints and creating a morph sequence which interpolated between. Figure 1 shows the four faces used in our study. One dimension (A) was created by morphing between the faces along the top of the figure, while a second dimension (B) was created by morphing between the two faces along the left edge. The actual faces used to construct the dimensions were selected according to preliminary work which allowed us to select dimensions that were roughly equally salient and roughly orthogonal when subjected to a MDS analysis. Using a blending technique described in Steyvers (1999), a 10x10 matrix of faces was created such that each face was defined by half its value on dimension A and half on Dimension B. Previous studies have show that subjects originally may have little sense of the dimensional structure of these faces, but with practice can isolate the relevant aspects needed to categorize along either dimension (Goldstone & Steyvers, 2001).

For some subjects dimension A was the category boundary, while for other's dimensions B was used. A "clustering" was induced within each category by removing a subset of the full 100 items during training. Figure 1 shows the basic structure of the categories. The light grey items in this figure were never presented to subjects during the category learning. As a result, the distribution of examples from each category were distorted such that there were two distinct sub-prototypes or clusters of items. However, this structure was incidental to performance in the main categorization task which required attention along the orthogonal dimension.

Procedure

The experiment was divided into two phases: a baseline discrimination phase and a mixed category learning and discrimination phase.

Phase 1: Baseline Discrimination The pre-category learning discrimination phase assessed each subject's baseline ability to discriminate between pairs of faces that varied along one or both stimulus dimensions. This initial baseline discrimination phase served two functions. First, it allowed us to assess for each subject, the ability to discriminate stimuli prior to category training. Second, it allowed us to evaluate the a-priori discriminability of each dimension and to confirm that each dimension was roughly equally salient.

The baseline discrimination phase was divided into 3 blocks consisting of 56 trials each. On each discrimination trial, subjects were asked to study a single face which was presented in the center of the display for 500 ms. Following this short study period, the target face disappeared for 500 ms (during which time the screen was blank) before two faces appeared side by side on the display. One of these faces exactly matched the studied face, while the other was a foil item. Subjects were asked to judge which of the two items they had just studied and indicated their response using the computer keyboard. The trial terminated after subjects entered their response and the next trial began after an inter-trial interval of 1500 ms. No feedback was provided about accuracy on discrimination trials.

Sixteen target items were selected from the four corners of each of the four "clusters" of items shown in Figure 1 (e.g., A1, D1, A4 and D4 for the cluster in the top left). Foils were stimuli that were two increments away along either one or both dimensions. For example, foil items for stimulus A1 were A4, D1, or D4. Likewise, foils for stimulus D10 were selected from the set D7, G7, G10, A10, or A7. Note that which item was considered the (studied) target and which was the foil was randomly determined any given trial. Each of the 3 baseline blocks tested all target-foil combinations once.

For the purposes of analysis, discrimination trials were classified according to their relation to the category structure used in the following learning phase. Discriminations were considered *within-cluster* when both the target and the foil item belonged to the same category *and* the same cluster (such as A1 vs. D1 or G7 vs. J10 in Figure 1). Trials where the target and foil shared the same category, but were from different clusters (such as A4 vs. A7 or G4 vs. J7) were classified as *within-category, between-cluster* discriminations. Finally, if the target and foil belonged to different categories (such as D1 vs. G1 or D7 vs. G10) they were classified as *between-category-and-cluster*.

Phase 2: Mixed Category Learning and Discrimination In the second phase, subjects completed three mixed blocks of category learning and discrimination trials. On each trial, a single face appeared for 500 ms in the center of the display. Subjects were asked to study the item carefully. However, they did not know for certain which type of judgement they would be asked to make about the item (categorization or discrimination). On category learning trials, immediately after the study phase, the face disappeared for 300 ms and subjects were asked to indicate if the studied item belonged to category 'A' or 'B' (note that the stimulus was not visible while they made their judgement). After the response was registered, subjects were again shown the target stimulus along with a prompt indicating if their previous judgment was correct or incorrect along with the correct label for the current item. On discrimination trials, the trial proceeded in an identical fashion, although instead of being asked to judge the category of the previously studied item, subject made the same target/foil discrimination used in the phase 1. No feedback was provided during discrimination trials. There were a total of 64 categorization judgements and the same 56 discrimination trials in each block (a total of 120). Trials were mixed so that in each set of 15 consecutive trials, 8 were categorization trials, and 7 were discrimination trials.

The benefit of this mixed procedure is two-fold. First, by mixing category learning and discrimination trials throughout the learning phase, we gain additional insight into the evolution of perceptual learning abilities in relation to category knowledge. Second, trials in this phase of the experiment were constructed so that at the time of encoding the target, subjects did not know if they would next be asked to categorize or identify the face. As a result, we are able to detect a stronger influence of category knowledge on perception in a relatively short training session. While previous work has established that the impact of category knowledge on perception abilities can be lasting, our technique can detect strong effects in a shorter training period and may be of general use for perceptual learning researchers.

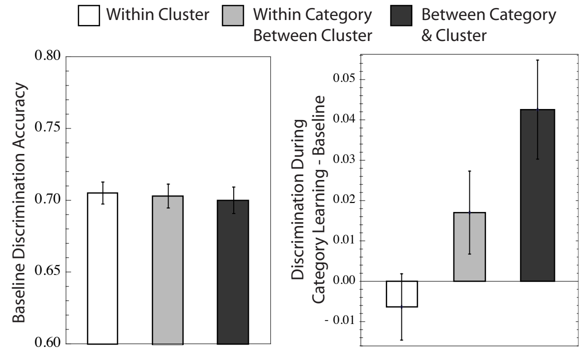

Figure 2: *A*: Pre-category learning discrimination performance for the three key types of discriminations. Prior to learning, subjects have roughly equivalent discrimination ability for each type of comparison (but are well below ceiling). *B*: The overall change in discrimination performance relative to baseline while learning the category. Subjects show increased discrimination between categories and between the clusters inside categories, but reduced discrimination of within cluster discriminations. All error bars are standard errors.

Results

Baseline Discrimination Ability As shown in Figure 2A, subjects demonstrated relatively robust (but below ceiling) discrimination ability for within-cluster discriminations (M=.71, SD=.09), between-cluster, within-category discriminations (M=.70, SD=.10), and between-cluster, between-category discriminations (M=.69, SD=.11). However, prior to category training, discrimination accuracy for all three of the types of comparisons were not significantly different ($F(2,238)$=.308, Mse=0.002, p=.74). We also considered if there were a-priori differences in discrimination along dimension A or B (irrespective of category structure). Overall, subjects ability to discriminate differences between stimuli that varied along dimension A only slightly exceeded that of dimensions B (M=.67, SD=.11 vs. M=.63, SD=.11, $t(119)$=2.9, $p<.005$), suggesting a reasonable balance in the baseline discriminability of the two dimensions. Note that in all subsequent analyses we considered the effect of category learning on discriminations *with respect* to this baseline ability which

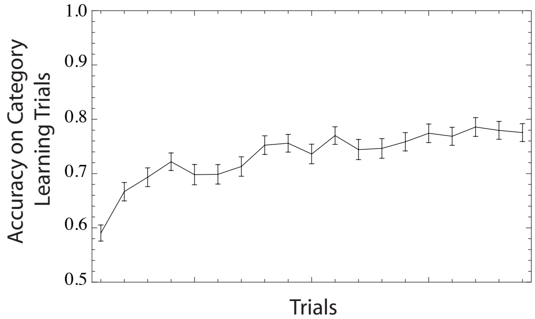

Figure 3: Category learning performance as a function of trials in the task considered in block of 10 trials as a time. Subjects asymptote around 80% correct.

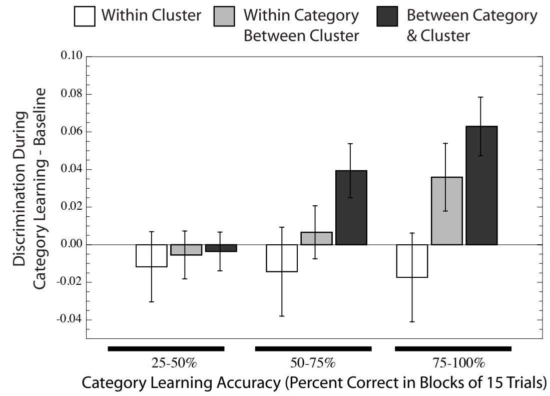

Figure 4: The evolution of the CP effects as a function of categorization accuracy. All error bars are standard errors.

takes into account this difference on a subject-by-subject basis.

Category Learning Performance Figure 3 shows participant's accuracy on category learning trials over the three training blocks of Phase 2 considered in sets of 10 trials as a time. Overall the categorization task was somewhat difficult. For example, during the last block of trials, subjects were still only about 80% accurate at category judgements. However, this is likely due to the fact that at the time of study for each trial, subjects were uncertain if they would be tested in categorization or discrimination and the stimulus was not present on the screen when they made their judgement (requiring memory for the presented item). Nevertheless, subject shows substantial category learning, particularly early in the task.

Discrimination Performance during Category Learning Of particular interest is the performance on discrimination trials during the three category learning blocks (phase 2). For each subject, baseline discrimination ability for *within-cluster*, *between-cluster-within-category*, and *between category and cluster* from phase 1 was computed. Then, discrimination ability during the category learning phase was computed for the same classes of stimuli comparisons. Figure 2B shows the change in discrimination ability during the category learning blocks compared to baseline. There was a strong effect of discrimination type ($F(2,238) = 8.24$, Mse=.07, $p <.0004$). Post-hoc tests revealed a near significant (Bonferroni corrected $\alpha = .016$) difference between *within-cluster* vs. *between-cluster, within category* judgements ($t(119) = 2.18$, $p <.031$), and between *between-cluster, within category* vs. *between category and cluster* judgements ($t(119) = 1.99$, $p < .05$) with a reliable difference between *within-cluster* and *between category and cluster* judgements ($t(119) = 3.9$, $p < .001$).

Considering only those subjects who reached 80% or better at classification in the final block, N=81) we found a more robust effect. In this group, both *between-cluster, within category* and *between category and cluster* judgements significantly improved ($t(80)$=2.66, $p <.01$ and ($t(80)$=4.55, $p <.001$, respectively) while *within-cluster* judgement did not ($t(80) < 1$). In addition, there was a significant difference (at the adjusted $\alpha = .016$ level) between *within-cluster* vs. *between-cluster, within category* judgements ($t(61) = 2.5$, p =.014), no difference *between-cluster, within category* as compared to *between category and cluster* judgements ($t(61) < 1$) and a strong difference between *within-cluster* and *between category and cluster* judgements ($t(61) = 3.05$, $p = .003$).

Our ability to examine the effect category learning on discrimination was limited by the fact that learning appeared early in the task. For example, a 2x2 repeated measures design with comparison type (*within-cluster*, *between-cluster, within category*, and *between category and cluster*) and learning block (1-3) revealed a main effect of comparison type ($F(2,714) = 13.806$, $Mse = .21$, $p < .001$) but no effect of block ($F(2,952) = 1.51$, $p = .21$) or interaction ($F(4,952) = .29$, $p = .88$). As a result, we sought to assess more directly the impact that learned category knowledge had on participant's discrimination abilities. First, we divided the category learning blocks into non-overlapping segments of 15 trials each. Within each segment, we then analyzed both the discrimination performance (as a function of contrast type) and overall categorization performance. Figure 5 shows the results of this analysis. For each subject we considered sub-blocks where category learning performance ranged from 25-50%, from 50-75%, and from 75-100% (there were very few observations where categorization accuracy fell below 25%). Within these three levels of categorization accuracy, we found that discrimination trials show a progressive improvement from near baseline levels to gradually increasing performance for discriminations that crossed either the category or cluster boundary. A repeated measures ANOVA with both accuracy range (3 levels) and discrimination type (3 levels) as within subject variables found a significant effect of category learning accuracy ($F(2,874) = 4.75$, Mse=.12, $p <.009$), a significant effect of discrimination type ($F(2,874) = 4.38$, Mse=.11, $p <$.013), but the interaction failed to reach significance ($F(4,874) = 1.46$, Mse=.04, p =.21). Note however that when categorization performance was high (i.e., >75% correct), subjects show better discrimination accuracy (at the adjusted $\alpha = .016$) for *between-cluster, within-category* judgements compared to *within cluster* judgements ($t(119) = 2.89$, $p < .005$), and between *between category and cluster* compared to *within category* ($t(119) = 3.83$, $p < .001$), but not between *between-cluster, within-category* judgements and *between category and cluster* judgements ($t(119) = 1.16$,$p = .25$).

Modeling Analyses

In order to account for this pattern of results, we evaluated a number of computational models of category learning.

Backpropogation Networks

One theoretical account of learned categorial perception effects was provided by Harnad, Hanson, and Lubin (1991) whereby simple three layer backpropogation networks were first trained to auto-associate input patterns with an identical output. Following this initial phase, the networks were then taught to simultaneously auto-associate input patterns with identical output patterns along with the additional target of predicting a category label for each item. Measuring the similarity in the pattern of hidden unit activations before and after category training show that these networks exhibit a between-category expansion (increased differences) and within-category compression (decreased differences) (Harnad, Hanson, & Lubin, 1995; Goldstone, Steyvers, & Larimer, 1996). In effect, the encoding that develops in the internal layer of the network reorganizes in response to the demands of categorization, changing the similarity relationships between hidden unit states in the model.

In our first set of simulations, we sought to test if this simple account of CP could explain our results. Following Harnad, et al. we trained simple three layer BP[1] networks to auto-associate stimuli that varied along two continuous valued dimensions (with input values ranging from 0.1 to 1.0) with identical output patterns. Once the networks reached 0.01 of their target values, we then trained the networks to simultaneously auto-associate and categorize a subset of the input patterns (withholding those shown in grey in Figure 1) and compared changes in hidden unit space activation patterns before and after category training. Figure 5 shows the results of this analysis in terms of the same contrasts use in the analysis of our experiment. After training, the BP network predicts strong between category expansion (improved category discrimination along the boundary) but no differential effect on the between-cluster comparison.

[1]We tested a variety of network sizes and found similar results with each. The results reported here are for networks with 8 hidden units, learning rate = 0.1, momentum = 0.1.

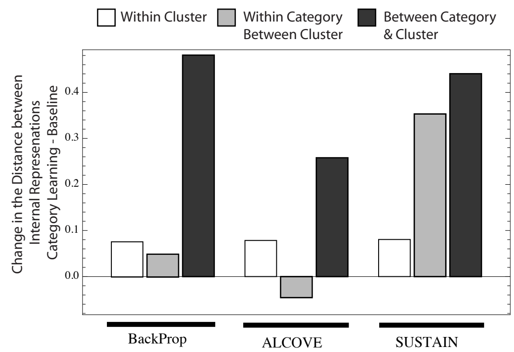

Figure 5: Change in predicted baseline discrimination compared to that following category learning for the three models tested.

ALCOVE

A similar problem faces models which assume a warping of the stimulus space via selective attention. For example, models such as ALCOVE (Kruschke, 1992) or the GCM (Nosofsky, 1986) assume that as attention is increased along a particular stimulus dimensions, differences along that dimension are accentuated. Critically, attention in these models is differentially allocated to dimensions which are predictive for category learning. Figure 5 shows the results of a simulation applying ALCOVE to our task. The exemplar memory in the model was initialized to hold all 100 stimulus items, and the attention weights on both dimensions were initialized to 0.5. Prior to category learning, we measured the pattern of activation in the hidden layer of the network in response to each item as in the Harnad, et al. simulations (see also Goldstone, Steyvers, & Larimer, 1996 for a similar approach). Then, we trained the ALCOVE network to predict the category membership of the training items shown in Figure 1 (for the same number of trials as human subjects). During the category learning phase the network adjusted its attention weights, particularly along the category-relevant dimension. Following this training, we recomputed the pattern of exemplar node activations in response to each item. As predicted, the model shows heightened discrimination for items that cross the category boundary. Interestingly, the model predict decreased discrimination along the within-category-between-cluster boundaries due to the fact that attention shifts away from this dimension (as it is irrelevant for categorization). Finally, we found slightly increased discrimination for within-cluster discriminations due to the fact that at least some of these discriminations also spanned the category-relevant boundary. Intuitively, increased attention to category relevant information is insufficient to account for the changes in discrimination performance observed in human subjects.

SUSTAIN

In our final simulation we applied a variant of the SUSTAIN model of category learning (Love et al., 2004). Unlike strictly exemplar or prototype models, SUSTAIN assumes that categories are represented in terms of a set of clusters which capture regularities both within and between categories. For example, when learning rule-plus-exception categories, SUSTAIN creates one cluster in memory to capture rule-following items and a separate cluster to capture the exception. The cluster recruitment process in the model is driven by both the goals of the learner as well as the similarity structure of experienced items. As a result, the model is able to predict how the internal structure of a category can influence the types of memory representations that are acquired.

As in the simulations described above, input to the model was encoded on two continuous valued dimensions. Since SUSTAIN was not directly designed to model perceptual discrimination tasks we simply assumed that the discriminability of items depended on two factors: the psychological distance or similarity between the items and the overlap in acquired category representations. Thus, to model discrimination of stimulus A and B, we computed the the degree to which a cluster centered at stimulus A would be activated by stimulus B (since these values are symmetric it makes no difference which is the target or which is the foil). This value indexed the similarity of the items independent of the acquired category knowledge (but takes into account changes in similarity due to attention). We then assumed that two sources of category information contribute to discrimination performance. First, we calculated the pattern of activations across the clusters that SUSTAIN recruited during category learning. In addition we computed the pattern of activation on the category output nodes (i.e. the node used to generate a category response). The sum of the perceptual similarity measure and the euclidean distance between the combined vector representing the cluster and category unit activation indexed the final predicted perceptual discrimination. Consistent with the structure of the task SUSTAIN created four clusters on average. Figure 5 shows the changes in the model's discrimination following category learning. Unlike the previous simulations, SUSTAIN clearly captures the basic pattern with lower sensitivity for items that share the same cluster following learning relative to items that cross either the category or cluster boundary. However, in these simulation SUSTAIN shows enhanced discrimination relative to baseline for items that fall in the same cluster (due to in part the definition of baseline as being based on perceptual information alone).

Discussion

Learners in our task appear to have picked up on sources of structure within each category that were irrelevant to the primary categorization task. These results show that the effect of learned CP can extend beyond simply the labels provided during category learning. The flexible ability to learn about category structures other than those that are overtly reinforced allows for adaptive behavior in a dynamic environment where what is good or bad at one moment can quickly change (e.g., substance A and B are poison, C and D are good, later A and C are good, B and D are poison). In that sense, our results share some similarity to findings such as pre-differentiation whereby simply giving animals exposure to stimuli prior to learning an associative distinction can facilitates the later task (Hall, 1991). However, to our knowledge, the empirical results presented here are the first demonstration of direct unsupervised influences on perceptual discrimination. In addition, our results present an interesting challenge for models of category learning. In general, the results appear consistent models such as SUSTAIN which are sensitive to patterns of within-category structure. While previous simulations with SUSTAIN have validated the cluster-based approach to category representation based on overall fits to empirical data, here we find direct evidence for subject learning clusters of items within overt categories.

Acknowledgments

This work was funded by NIH-NIMH training grant T32 MH019879-12 to T.M. Gureckis.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Goldstone, R. (1994). Influence of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178-200.

Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585-612.

Goldstone, R., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *1*, 116-139.

Goldstone, R., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the 18th annual conference of the cognitive science society.* Hillsdale, N.J.: Erlbaum.

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

Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Harnad, S., Hanson, S., & 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 modeling: Steps toward principaled integration* (p. 191-206). Boston: Academic Press.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358-368.

Logan, J., Lively, S., & Pisoni, D. (1991). Training japanese listeners to identify english /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*, 874-886.

Love, B., Medin, D., & Gureckis, T. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), 309-332.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Steyvers, M. (1999). Morphing techniques for generating and manipulating face images. *Behavior Research Methods, Instruments, & Computers*, *31*, 359-369.