Learned Visual Categorical Perception Effects Depend on Method of Assessment and Stimulus Discriminability

Joshua R. de Leeuw (jodeleeu@indiana.edu)
Department of Psychological and Brain Sciences, Program in Cognitive Science
Indiana University, Bloomington, IN

Jan Andrews (andrewsj@vassar.edu)
Ken Livingston (livingst@vassar.edu)
Department of Cognitive Science
Vassar College, Poughkeepsie, NY

Abstract

Learned categorical perception (CP) effects were assessed using three different measures and two sets of stimuli differing in discriminability, both of which varied on one category-relevant and one category-irrelevant dimension. Two different kinds of analysis produced patterns of results that depended on both of these variables and show that categorical perception effects are sensitive to variations in assessment task and stimulus discriminability. Only the similarity-rating task produced evidence of between-category expansion effects, suggesting that participants used different strategies for subjective and objective tasks. Generally, there was evidence that category training caused a decrease in the salience of category-irrelevant variation, but when the assessment task cued participants to category-irrelevant differences they were equally apt at identifying category-irrelevant variation as a control group.

Keywords: Categorization; categorical perception; compression; expansion; learning; similarity; online experiments.

Introduction

Learning to categorize stimuli in a new way can change how those stimuli are perceived or judged. This phenomenon is called learned categorical perception (CP). There are numerous kinds of reported CP effects (for a recent review, see Goldstone & Hendrickson, 2009). Learning that two stimuli belong to the same category can increase their similarity or perceptual confusability, an effect known as compression (e.g., Livingston, Andrews, & Harnad, 1998), while learning that two stimuli belong to different categories can have the opposite effect, often called expansion (e.g., Goldstone 1994; Notman, Sowden, & Özgen, 2005). Categorizing stimuli based on particular sets of features may increase the relative influence of those features at the expense of other stimulus features, regardless of whether or not the stimuli belong to the same or different categories (e.g., Goldstone, 1994).

Although there are several possible consequences of learning to categorize stimuli, most studies of learned CP only report finding one of the possible effects, even though different kinds of CP effects are not logically mutually exclusive. Determining why categorization training leads to different CP effects in different experimental contexts is an important step towards a thorough understanding of the mechanisms that cause CP.

Several studies have demonstrated that whether CP is observed at all in certain scenarios depends on various experimental factors, such as the availability of verbal labels during perceptual testing (Kikutani, Roberson, & Hanley, 2008), the particular kind of perceptual assessment used to determine whether CP is present (Gerrits & Schouten, 2004), and how stimulus morphspaces are created (Folstein, Gauthier, & Palmieri, 2012). However, few studies have reported differences in the kind of CP effect observed based on experimental manipulations, although Goldstone, Lippa, & Shiffrin (2001) and Livingston & Andrews (2005), both reported two qualitatively different CP effects exhibited by the same subjects when they were tested in two different ways. Thus, one possible reason for the diversity of CP effects is that different tasks used to assess CP are sensitive to different aspects of CP or invoke different processes, only some of which exhibit particular CP effects.

In addition to the task used to measure CP, it is possible that incidental differences in stimuli between experiments are responsible for the different kinds of CP effects observed. If subjects learn that stimuli with small differences belong in different categories, successful categorization necessitates the ability to differentiate the stimuli, which might naturally lead to expansion effects. However, if the differences between stimuli are obvious prior to training, then expansion might be less likely to occur. Pevtzow and Harnad (1997) found larger expansion effects for stimulus sets that were harder to differentiate, which is consistent with this hypothesis.

In this experiment, we directly tested the influence of stimulus discriminability and assessment task on CP. We created an artificial stimulus set that varied on two dimensions, and selected two subsets of stimuli from this set, one with only half as much variation between neighboring pairs as the other. We expected that the stimulus set with less variation would be more likely to produce expansion effects, while the stimulus set with more variation would be more likely to produce compression effects. Three commonly used tasks were implemented to assess CP: a similarity rating task, a same/different task, and an XAB forced-choice task. We expected task to interact

with discriminability, due to differences in the demands of each particular task. Subjective rating tasks (e.g. similarity judgments) may invite strategic responses (altering the rating based on the category labels, and not warping of perceptual similarity), and thus produce CP effects even when objective measures (e.g. same/different, XAB) do not, especially in cases where perceptual learning is not necessary for categorization. We included both the XAB and same/different tasks since both are frequently used in the literature, although we expected them to produce similar results.

Method

Participants

We recruited 290 participants through Amazon Mechanical Turk (AMT)[1], paying between $0.50 and $1.25 for participation, depending on the projected length of the experiment. Eight subjects were excluded from the analysis because of a bug that allowed them to complete more than one condition of the experiment, leaving 282 participants.

Materials

The experiment was developed using jsPsych, a software library for building online experiments (de Leeuw, 2014). Stimuli were cell-like shapes that varied on two dimensions, shape and tail length. We generated two sets of stimuli: a high discriminability (HD) set and a low discriminability (LD) set. The LD stimuli had half as much variation as corresponding HD stimuli. See Figure 1. The stimulus space was 6x6 for both sets. The category boundary was between the 3rd and 4th stimuli on the shape dimension. Thus the shape dimension (chosen arbitrarily) was always the relevant dimension for categorization, and the tail length dimension was always irrelevant.

Procedure

Subjects were randomly assigned to one of twelve conditions: 2 training type (control v. category training) X 2 stimulus sets (HD v. LD) X 3 type of assessment (XAB v. same/different v. similarity). The N per condition ranged from 20 to 28.

Training Subjects who were assigned to a training condition completed an adaptive training protocol by iterating through multiple blocks of training until all stimuli were learned. In each block of the procedure, all 36 stimuli were shown one at a time and categorized as either a 'Tig' or a 'Bep' by the participant. Subjects were told that they would initially need to guess a cell's category but would receive feedback indicating the correct category for each cell. When a stimulus had been correctly categorized in four consecutive blocks, the stimulus was removed from the training set. If at any point fewer than five stimuli were left in the training set, stimuli that had already been learned were randomly included in the block (but these stimuli were considered learned, even if an error was made). Once all

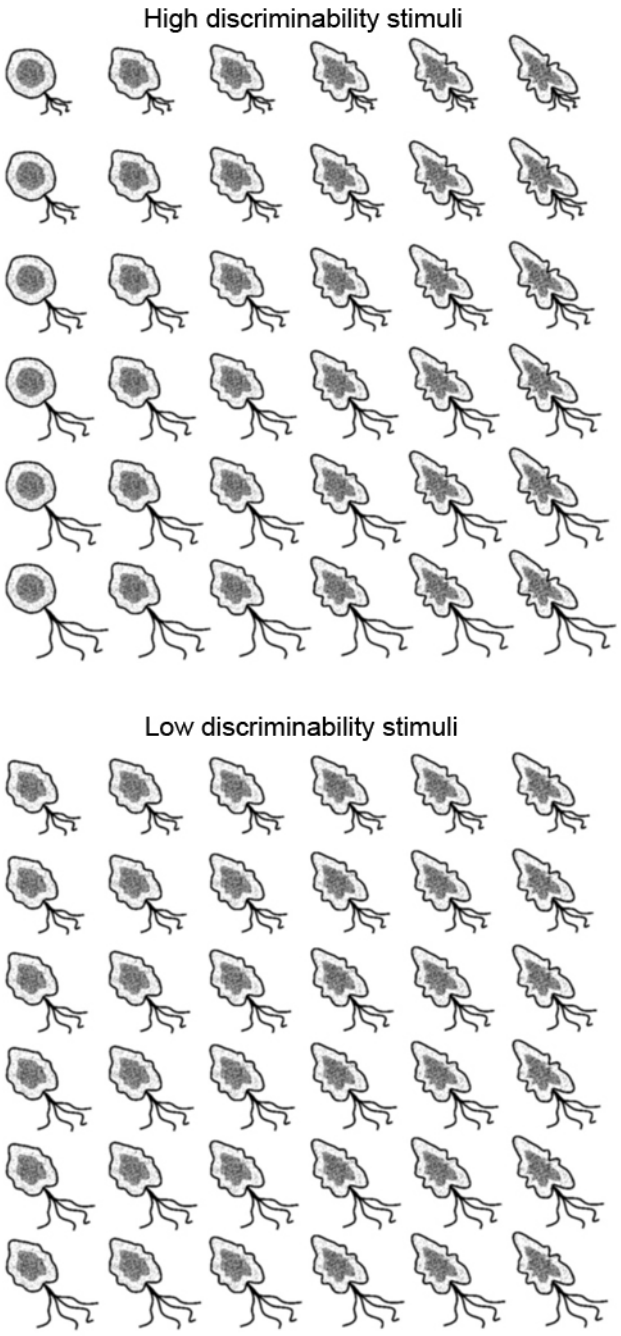

Figure 1: Stimuli used in the experiment. The left three columns of each set belonged to one category, and the right three columns belonged to the other category.

[1] Samples from AMT tend to replicate laboratory findings, though category-learning tasks on AMT have produced mixed results (Crump, McDonnell, & Gureckis, 2013). Since the key methodological consideration for our study is that all participants in learning conditions have acquired the category structure before doing the discrimination trials, we used an adaptive training protocol that ensures participants learned the category structure before progressing to the testing phase.

stimuli had been learned, the training ended. If a participant got fewer than 60% of the items correct in a round, then the round did not count towards the four consecutive blocks (to prevent guessing strategies). If a participant got fewer than 60% correct for 5 consecutive rounds, then training ended and the participant was considered to have failed training.

Post-Training Categorization Test Once participants completed the adaptive training, they categorized each of the 36 stimuli without feedback in a single block. Stimuli were presented in a random order and remained on screen until the participant gave a response. A blank screen was displayed for 1500ms between stimuli.

Post-Training Assessment Tasks Subjects completed one of three different assessment tasks. Subjects who received category training completed the task immediately after the category learning test, while control subjects only performed the assessment task.

Similarity In the similarity task, subjects saw two stimuli sequentially (each stimulus was visible for 750ms, with a blank screen displayed for 1000ms between stimuli). They dragged a movable slider to indicate how similar the two stimuli were. The scale was anchored by the labels "most similar" and "least similar". There were 9 different pair types that could be presented: (tig-tig pairs, bep-bep pairs, or tig-bep pairs) X (1, 2, or 3 city block units of distance between pairs). Each of the 9 different types was selected exactly 4 times, but the particular exemplars that made up each pair were selected at random from all possible pairs that satisfied the constraints.

Same/Different In the same/different task, subjects saw two stimuli sequentially. Each was visible for 750ms, with a blank screen displayed for 1000ms between stimuli. They pressed a key to indicate whether the stimuli were the same or different. There were 4 blocks of 54 pairs of stimuli. Each block consisted of 27 identical pairs and 27 pairs with variation. The 27 pairs with variation were selected according to the same policy as used in the similarity rating task, except only 3 pairs per type were chosen instead of 4, to limit the overall length of the experiment.

XAB In the XAB task, subjects saw a target stimulus (X) for 750ms, followed by a blank screen for 1000ms, and then the simultaneous presentation of two stimuli (A and B) for 750ms. Subjects pressed a key to indicate whether A or B was identical to X. There were 4 blocks of 36 pairs of stimuli, selected as in the similarity rating task.

Results

One subject failed training and was excluded from the analysis. Subjects in the high discriminability conditions completed training in fewer trials (M = 188, SD = 30) than subjects in the low discriminability conditions (M = 211, SD = 36), $t(141) = 4.103$, $p < 0.0001$.

We looked for CP effects in two ways. First, we compared performance on within-category pairs (the stimuli in each pair belonged to the same category) to performance on between-category pairs (the stimuli in each pair belonged to different categories). Our second analysis examined the effect of one dimension of variation (either the irrelevant or relevant dimension) on performance when the other dimension was held constant.

Between-Category Versus Within-Category Pairs

We restricted the data set to pairs that only varied on the category-relevant dimension, with a maximum distance of 2 between the items. We did this to emphasize the category boundary aspect of this analysis. Since the second analysis focuses on the relative change in importance of the dimensions, we used the first analysis to look only at changes related to the category boundary. The pairs that were 3 units apart were not used in this analysis because all such pairs were between category. We conducted three separate 2 (training type: control v. category training) X 2 (pair type: between v. within category) X 2 (stimulus set: high discriminability v. low discriminability) ANOVAs, one for each type of assessment.

For same/different judgments there were no significant effects that involved training type as a factor. For similarity judgments there was a significant interaction of training type and category type, $F(1, 90) = 5.02$, $p = 0.027$. Subjects who received category training rated between category pairs as less similar than subjects who did not receive training, but there was no difference in their ratings of within category pairs. For XAB judgments there was a significant three-way interaction between training type, category type, and stimulus set, $F(1, 97) = 4.83$, $p = 0.03$. Control subjects were more accurate than training subjects on within category judgments for the high discriminability stimuli but not for low discriminability stimuli. See Figure 2.

Dimension Influence

To measure the influence, ϕ, of changes along a dimension, d_1, while the distance along the other dimension, d_2, remains at a fixed value, c, we used a measure defined as follows:

$$\phi(d_1|d_2 = c) = \frac{1}{n} \sum_{i=min(x|d_2=c)}^{max(x|d_2=c)} \sum_{j=min(x|d_2=c)}^{i-1} \delta_{i,c} - \delta_{j,c}$$

$$n = y(y+1)/2$$

$$y = \max(x|d_2 = c) - \min(x|d_2 = c)$$

Where x represents a distance along the dimension d_1, and $\delta_{a,b}$ is the average value of the dependent measure for all pairs with distance along d_1 = a and distance along d_2 = b. Constraining the summation with the max and min functions is necessary because the range of presented distances along d_1 is determined by the distance along d_2 (since we constrained the total city-block distance between pairs to be

no more than 3). The $1/n$ component serves to average the differences that were summed over. Intuitively, this measure is computing something similar to a slope of how much the dependent variable changes as the distance on d_1 changes. The value of this measure is large when changes along the dimension produce correlated changes in the dependent variable, i.e. increasing the distance along the dimension (while holding the distance along the other dimension constant) causes an improvement in discriminability. We calculated the influence of the relevant dimension at each value of the irrelevant dimension, and vice versa. This calculation was performed for each subject individually. We conducted six separate 2 (training type: control v. category training) X 3 (distance on the constant dimension: 0, 1, or 2) X 2 (stimulus set: high discriminability v. low discriminability) ANOVAs, one for each combination of the 3 types of assessment and 2 dimensions of influence (relevant or irrelevant).

We found no significant changes in the influence of the category relevant dimension as a result of training. We did find some changes in the influence of the irrelevant dimension as a function of training. For same/different judgments we found a main effect of training, $F(1, 82) = 12.47$, $p = 0.0007$. Subjects who received category training were less likely to be influenced by the category irrelevant dimension. For similarity judgments, we found both a main effect of training, $F(1, 90) = 9.21$, $p = 0.003$, and a training X stimulus set interaction, $F(1, 90) = 5.25$, $p = 0.02$. Subjects who received category training were less influenced by the irrelevant dimension, and this effect was stronger with the high discriminability stimuli. For the XAB task, there were no significant changes in influence of the irrelevant dimension as a result of training. See Figure 3.

Discussion

We found evidence of CP effects with all three tasks and with both stimulus sets, but each combination of task and stimulus set produced a somewhat different pattern of results. Our specific findings were: (1) The influence of the irrelevant dimension decreased for the same/different and similarity tasks but not for the XAB task. (2) The high discriminability stimulus set produced a compression effect in the XAB task for stimuli varying on the relevant dimension. This was the only task/stimulus set where such an effect was observed. (3) The between vs. within category pairs analysis also revealed an expansion effect, but only for the similarity task.

At a basic level, this supports the idea that learned CP effects depend on the task by which they are assessed and to a lesser extent on characteristics of the stimuli that are used. The specific task-stimulus set combination not only affected what type of CP effect occurred, but whether any CP effect occurred at all. While the current results do not suggest any definitive general conclusions about how task or stimulus structure affects CP, there were some common threads that point to broader principles.

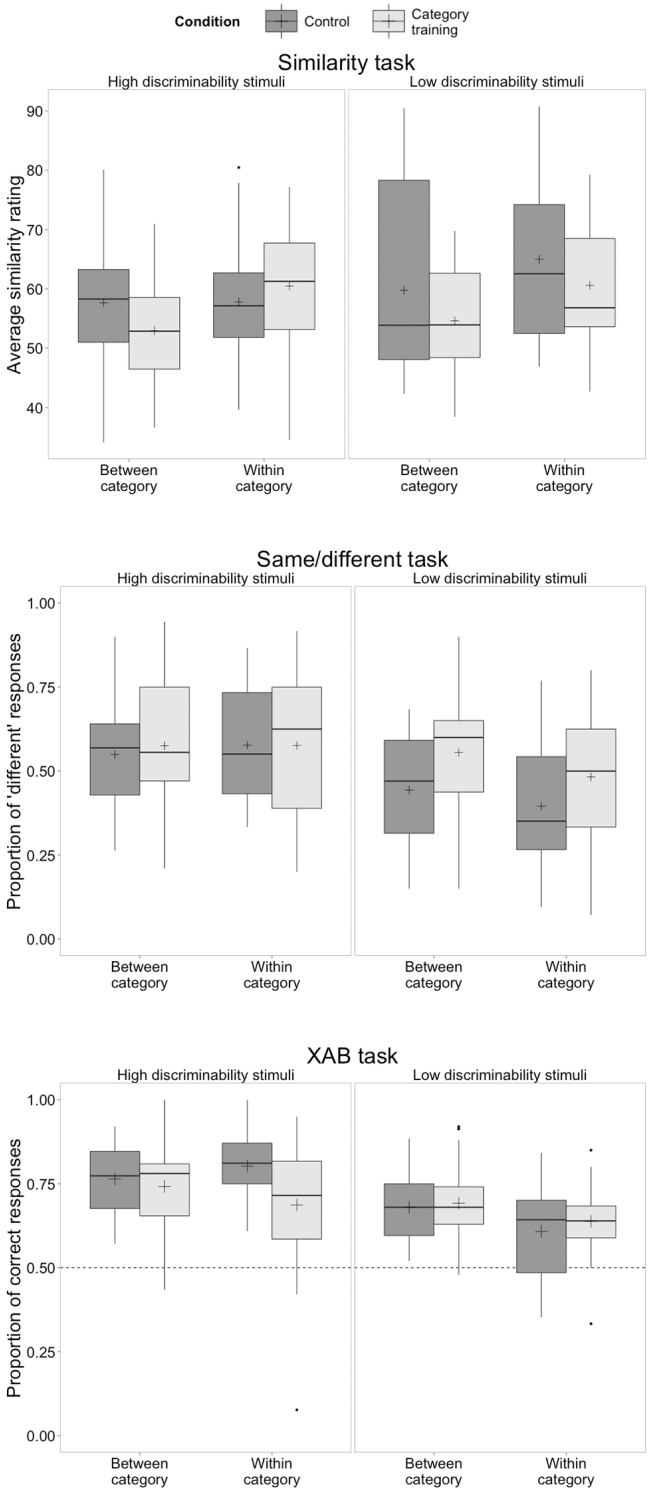

Figure 2: Performance on each assessment task for between-category versus within-category pairs. The + symbols denote the mean value, and the whiskers show the range of values that are within 1.5 * IQR (Inter-quartile range). Outliers are shown individually as dots. The dotted line in the XAB graph represents chance performance.

First, we tended to find evidence of compression-like effects, with the only exception being a decrease in similarity ratings for between category pairs. In two of the three tasks, the influence of the irrelevant dimension was reduced for subjects receiving category training, constituting a compression effect, and in no case did category training increase the influence of the relevant dimension. Here, it seems likely that our choice of stimuli was a major factor. We created stimuli that made it easy to attend to one dimension and not the other, and the two dimensions were not equally salient. Control subjects were influenced by the relevant dimension far more than the irrelevant dimension even without category training. Changes on the irrelevant dimension may have been both harder to detect and easier to ignore. This inadvertently created a scenario that was prone to compression-like effects, but not expansion. However, if true, this emphasizes the general conclusion that CP effects are sensitive to stimulus design. Indeed, research by Notman, Sowden, and Özgen (2005) shows expansion with extremely low-discriminability stimuli using a same/different task as a measure. We hypothesize that decreasing the discriminability of our stimuli, or even just making the far less discriminable tail dimension the relevant dimension, should produce more, and more robust, expansion effects, strengthening the same/different task results shown in Figure 2 where the low discriminability stimuli appear to show an expansion-like pattern.

While the stimuli might have been biased to produce a certain pattern of results, we still found different effects using the different measures, indicating that tasks are an important consideration in assessing CP.

The similarity task was the only one to produce an expansion like effect. Goldstone, Lippa, and Shiffrin (2001) note that learning that two stimuli are in different categories may reduce similarity ratings simply because the stimuli belong to different groups. This might seem a fitting explanation of our data given the lack of evidence that the relevant dimension became more influential or discriminable in any of our tasks, but it is clear from previous research that using a similarity task is more associated with compression than expansion effects (e.g., Livingston, Andrews, & Harnad, 1998; Livingston, Andrews, & Dwyer, 2001). Other factors such as category structure may mediate what kind of effect is observed with a similarity rating task (Reppa & Pothos, 2013), potentially explaining the apparent discrepancy between these results.

The XAB and same/different tasks, though both speeded perceptual judgment tasks, yielded very different patterns of results. The same/different task showed a clear decrease in the influence of the irrelevant dimension for trained subjects, but there was no such change observed with the XAB task. This may be because the XAB task forces subjects to compare two different stimuli, which in some cases vary only on the irrelevant dimension. Thus, in a case where subjects have learned to ignore variation on the irrelevant dimension, the XAB task would alert them to the fact that they are missing information when they are confronted with two stimuli that appear to be the same.

In addition, the XAB task revealed a somewhat puzzling interaction where within category pairs in the high discriminability condition became more confusable as a

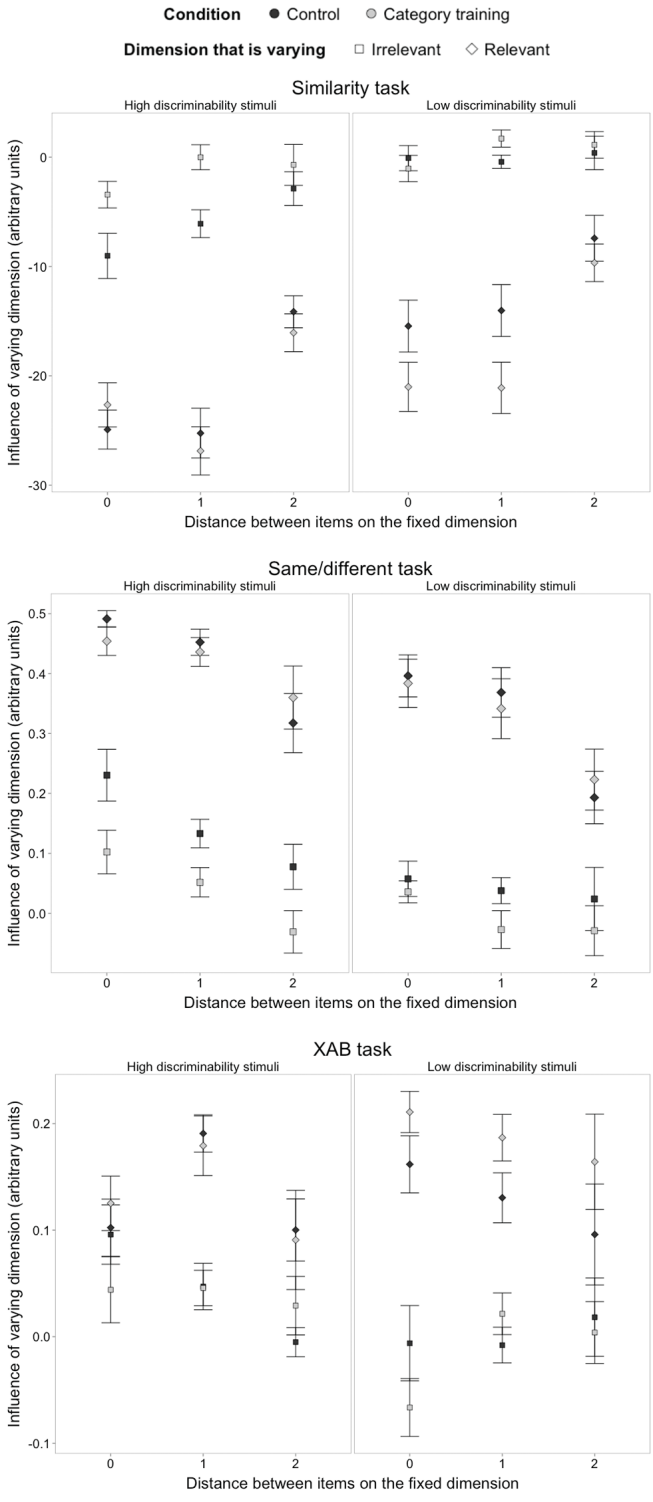

Figure 3: Influence of relevant and irrelevant variation for each assessment task (see text for explanation of measure).

within pairs related to training condition with the same/different task. The XAB task may bias responses according to whether the target stimulus is closer to the center of the stimulus space than the foil stimulus (Hanley & Roberson, 2011; Hendrickson, Carvalho, & Goldstone, 2012). We looked for evidence of this bias in the XAB, HD data, and found a nearly significant pattern in which the trained group with the HD stimuli was more likely to pick the correct answer when the target was more extreme (farther from the center of the stimulus space) than when the foil was more extreme, $t(23) = 2.05$, $p = 0.051$, while there was no such difference for the control group. This result is consistent with the idea that the bias results from prolonged exposure to a stimulus space (Hendrickson et al., 2012) and reflects a labeling process (Hanley & Roberson, 2011), since only the trained group acquired labels, and they had longer exposure than the control group.

The current results highlight the variability and complexity of learned CP effects and the difficulty of comparing effects across studies that differ in such factors as task used to assess CP, stimulus discriminability, salience of dimensions, and type of category structure. This difficulty is exacerbated by the variety of analyses that can be used to explore CP effects, such as the two used here that produced different patterns of effects across the three tasks. A great deal of research suggests that learned CP effects are a critical manifestation of the category learning processes, but our results strongly suggest that to understand category learning will require that we model not just the learning process that gives rise to CP, but also the various tasks and stimuli used to assess it.

Acknowledgments

We thank the Vassar College Phoebe H. Beadle Science Fund Endowment for financial support. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1342962. We thank Rob Goldstone for helpful data analysis suggestions.

References

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57401. doi:10.1371/journal.pone.0057410.

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods.* Advance online publication. doi:10.3758/s13428-014-0458-y

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 807–20.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–76.

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

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Interdisciplinary Reviews: Cognitive Science, 1*, 69-78.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Hanley, J.R. & Roberson, D. (2011) Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review*, *18*(2), 355-363.

Hendrickson, A. T., Carvalho, P. F., & Goldstone, R. L. (2012). Going to Extremes: The influence of unsupervised categories on the mental caricaturization of faces and asymmetries in perceptual discrimination. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Soc.* (pp. 1662-1667). Austin, TX: Cognitive Science Society.

Kikutani, M., Roberson, D., & Hanley, J. R. (2008). What's in the name? Categorical perception for unfamiliar faces can occur through labeling. *Psychonomic Bulletin & Review*, *15*(4), 787–794.

Livingston, K. R., & Andrews, J. K. (2005). Evidence for an age-independent process in category learning. *Developmental Science*, *8*(4), 319–25.

Livingston, K. R., Andrews, J. K., & Dwyer, P. (2001). Ties that bind: Reconciling discrepancies between categorization and naming. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 558–563). Mahwah, NJ: Erlbaum.

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*(3), 732–753.

Notman, L., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition*, *95*(2), B1–14.

Pevtzow, R., & Harnad, S. (1997). Warping similarity space in category learning by human subjects: The role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization* (pp. 189–195). Department of Artificial Intelligence, Edinburgh University.

Reppa, I., & Pothos, E. (2013). Predicting similarity change as a result of categorization. In M. Knauff, M. Pauen, N., Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1211-1216). Austin, TX: Cognitive Science Society.