Distinguishing learned categorical perception from selective attention to a dimension: Preliminary evidence from a new method

Janet Andrews (andrewsj@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Joshua de Leeuw (jdeleeuw@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Rebecca Andrews (reandrews@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Cole Landolt (clandolt@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Chrissy Griesmer (chgriesmer@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Abstract

A novel experimental method is motivated and applied in an effort to test for effects of category learning on perceptual discrimination so as to clearly distinguish category boundary effects of expansion and compression from changes in sensitivity to stimulus dimensions. The method includes a control group performing a task that, like category learning, requires attention to one systematically varying stimulus dimension rather than another. Discrimination accuracy is tracked over time and measured using a psychophysical staircase procedure tailored to individual participants that doesn't rely on memory. Initial results suggest improvement in discrimination accuracy over time, particularly on the dimension relevant to the categorization or control task, but no evidence of category boundary effects or effects of category learning on dimension perception stronger than those of the control task. Possible reasons for this and directions for further research are briefly discussed.

Keywords: categorical perception; categorization; learning; expansion; compression; dimensional modulation; selective attention

Introduction

It is well known that various kinds of experience can produce perceptual learning, i.e., improved ability to distinguish objects, features, or values on a dimension (Goldstone, 1998). One of the processes that is claimed to have special effects on the perceptual judgment of stimuli is learning to categorize the items, the phenomenon known as learned categorical perception (CP) (Goldstone & Hendrickson, 2009). Learned CP effects reported in the literature include boundary effects whereby items placed in different categories become more distinguishable, sometimes called expansion, and/or items placed in the same category become less distinguishable, sometimes called compression. However, these are not always clearly distinguished from dimension-wide effects where there is sensitization to the category-relevant dimension(s) and/or desensitization to the category-irrelevant dimension(s).

There are potentially many tasks besides category learning that require or benefit from greater attention to one dimension rather than another whereas only category learning would be expected to produce the boundary effects of expansion and/or compression. It is therefore very important that measures of learned CP carefully distinguish dimensional effects from boundary effects, something that previous research has not necessarily done. An important goal of the work reported here is to develop a method that distinguishes boundary effects of category learning from dimension-wide effects and, if category learning does cause dimension-wide effects, to determine if it does so to a greater extent than a task that doesn't involve category learning.

One reason that learned CP effects are of theoretical interest is that they may provide key evidence of genuine top-down effects on perception, an issue of considerable current controversy (Firestone & Scholl, 2016). But since the vast majority of learned CP evidence is based on measures that rely on memory (e.g., successive judgments of pairs of stimuli for same-different or similarity judgments), it is hard to argue that they are genuinely perceptual effects rather than reflecting higher level cognitive processes. Another purpose of the method adopted here is to eliminate the role of memory and determine if learned CP effects still occur. (Of course, even if they do, other challenges raised by Firestone and Scholl might still need to be addressed.)

An examination of the existing body of learned CP research also reveals a bewildering pattern of effects and non-effects (compression vs. expansion vs. both, boundary effects with or without accompanying dimensional effects and vice versa, etc.). Researchers rarely have specific predictions regarding which effects will or won't occur and often don't distinguish clearly between them or test for all of them. As noted above, our study will clearly distinguish boundary effects from dimension-wide effects of category learning.

A recent p-curve meta-analysis of this body of research (Andrews, de Leeuw, Larson, & Xu, 2017) found a low level of statistical power, suggesting that it may be unproductive to try to interpret the patterns of effects and non-effects in the existing literature, since low statistical power is likely to produce both false positive and false negative results. Without a firm grasp on which learned CP effects do and don't occur under what conditions, it will be very difficult to make progress understanding the theoretical basis of learned CP or modeling the relevant mechanism(s). In addition to simply running better powered studies, another strategy to increase the informativeness of the data that are collected is to use analysis techniques such as Bayesian statistics that indicate the relative support for different hypotheses regarding learned CP effects, including the null hypothesis of no effects.

Another important methodological feature that renders previous results difficult to interpret is the fact that learned CP experiments almost always use a before-after comparison, a control group that only performs the final task performed by the learning group after category training, or at most, a control group that receives passive exposure to the category training stimuli. The goal of the research reported here is to address this and the other features of learned CP research that render its results ambiguous. Our approach relies on the use of a new method for tracking the effects of learning to categorize a set of patterns *over time* and in *comparison to the effects of performing an appropriate non-category-based control task*. Tracking over time is important for addressing another ambiguity when effects are only measured after training: expansion effects cannot be distinguished from a combination of compression and sensitization to the category-relevant dimension. These could potentially be distinguished if they emerge at different rates or times over the course of training. In order to track effects of learning over time, we test for changes in discrimination ability using a psychophysical staircase procedure throughout the entire experiment, alternating with classification or control task trials.

Because we use *simultaneous* stimulus presentation to avoid memory effects, a standard same-different or XAB task would allow successful performance based on the comparison of meaningless pixel-level features. We therefore developed a stimulus set where the potentially category-relevant dimensions vary both systematically in one respect (e.g., number/density of dots inside a circle) and also randomly (e.g., the exact location of the dots). This means that two stimuli with the same values on the two systematically varying dimensions will not be identical, much in the same way that individual instances of real world categories are usually unique. This allows us to use a variation on same-different judgments that highlights the role of the dimensions and works with simultaneous presentation, as explained in the method section.

The above features of our method make it different from the usual learned CP experiment in a number of ways, but we think it is essential to determine whether learned CP will occur under these more controlled conditions. If it does not, we can systematically re-introduce more traditional methodological features, such as successive presentation on the discrimination test, to determine which are necessary to produce the effects in order to better understand them. While we only report one experiment and acknowledge that our method likely needs adjustment to be fully successful in achieving its goals, our hope is that by sharing our work at this stage we can obtain useful feedback to inform and guide our next steps.

Method

All study materials, data, and analysis scripts are available at this OSF site: https://osf.io/msq57/.

Participants

A total of 101 participants (52 women; mean age 34.8; age range 18-72) were recruited using the online crowdsourcing platform Prolific and paid $4 for participating. Data from 8 participants were missing or incomplete leaving a final total sample size of 93.

Stimuli

Stimuli for this experiment were sunbursts. The number/density of dots and lines was systematically varied across stimuli but the exact placement of the dots and lines and the length of the lines were random (see Figure 1). For each participant in the experimental group (see below), category membership was randomly assigned to be based on either line or dot density. The density of dots or lines in a particular stimulus ranges from 300-2000 dots and 30-550 lines. (Each range is treated as 0.0-1.0 here.)

Procedure

The software jsPsych was used to create the experiment (de Leeuw, 2015). **Phase 1** used a same-different task variant we call the odd-one-out task. Four sunbursts appeared simultaneously: three had the same dot and line densities and one differed on one of those dimensions. Participants had 4 seconds to press a number key (1-4) to indicate the odd one out and receive feedback (see Figure 2).

At the beginning of Phase 1, the dimension that differed in the odd one differed by a large amount from the others. This distance was subsequently adjusted through a staircase

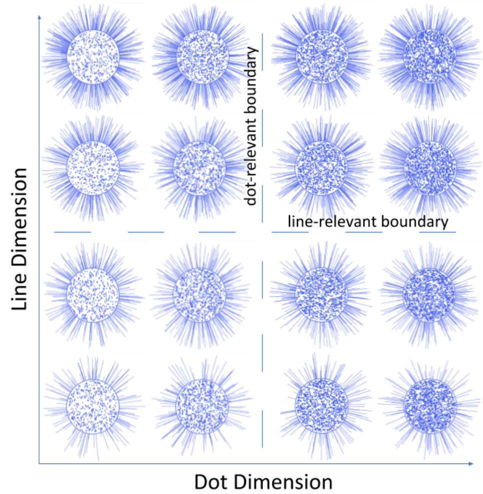

Figure 1. The stimulus space illustrating the two dimensions and the two possible sets of categories.

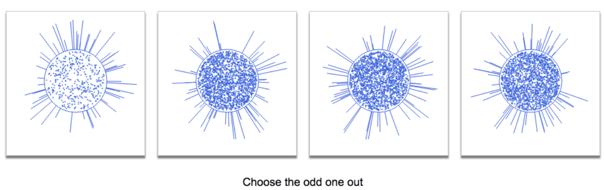

Figure 2. An odd-one-out trial display in Phase 1.

procedure, decreasing or increasing by 15% depending on whether the response was correct or incorrect. Trials continued until at least eight reversals occurred on each dimension. The goal of Phase 1 was to identify an approximation of each individual participant's just noticeable difference (JND) on each dimension, defined as the average of the distances of the last four reversals.

In **Phase 2**, odd-one-out trials alternated with one of two other tasks, classification or number judgment, to which participants were randomly assigned. For both tasks, a single sunburst appears with a question and participants press a key to answer and receive feedback (see Figure 3).

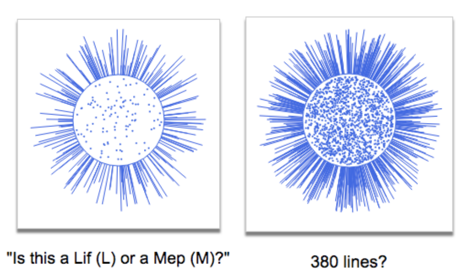

Figure 3. A classification task trial display (left) and a number judgment task trial display (right) in Phase 2.

The number judgment (control) task is to say "more" (M) or "less" (L) in response to a question about the number of dots or lines, where the number varied from trial to trial. For a given participant, the number judgment questions are always about just one of the two dimensions, randomly assigned, so that the control task matches the category learning task in relying on attention to one "relevant" dimension to answer correctly. For the classification ("experimental") group, the randomly assigned relevant dimension defined the category boundary as shown in Figure 1.

The specific stimuli used in Phase 2 odd-one-out trials were initially based on each JND value from Phase 1 for each participant and dimension. The sets of four stimuli (see Figure 2) were of three types as shown below in Figure 4. For both BE (between category) and WI (within category) comparisons, the odd one out differed from the other three only on the relevant dimension while for IRR comparisons, it differed only on the irrelevant dimension. All 48 possible adjacent stimulus pairs were used as the basis for the odd-one-out trials and drawn from the participant's JND-based dimensional space at a given moment.

Phase 2 trials proceeded in 40 blocks each containing six odd-one-out trials (one BE, two WI, and three IRR trials to sample the stimulus space evenly) and six classification or number judgment trials in a random order. The staircase procedure on the odd-one-out task was continued individually for each participant throughout Phase 2 just as in Phase 1, but separately for these six comparison subtypes. This controls for discriminability differences due to stimulus magnitude (e.g., Weber's law). Since adjacent dimensional values were already near JND level, the proportion change from one trial to the next of that subtype was reduced from 15% to 5% and the maximum distance allowed between dimension values was .33.

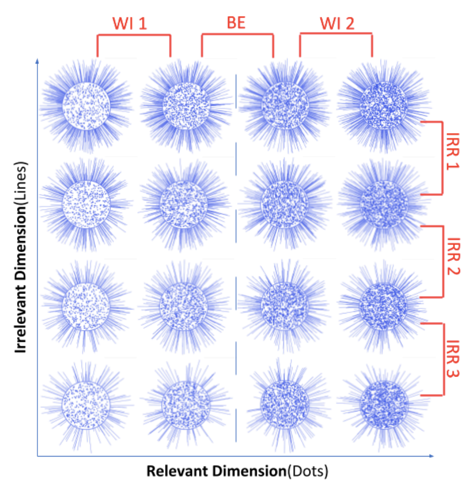

Figure 4. Illustration of the six comparison subtypes for the odd-one-out trials with dots as the relevant dimension.

Analysis Plan

A traditional learned CP analysis takes a behavioral measure such as similarity rating or same-different accuracy and compares the experimental (category learning) and control groups on that measure for between-category vs. within-category pairs. Our experiment tracked changes in the size of the distance between the two dimensional values used in odd-one-out task trials. Therefore, our learned CP measure was the change in this value for a given dimension from the beginning to the end of Phase 2. If participants improved on the odd-one-out task, their scores will be negative since they will become able to accurately judge smaller differences, and a larger negative score represents more improvement. Because differences in speed of discriminating between-category vs. within-category pairs are sometimes taken as evidence for CP, we also used mean correct reaction time over the last four blocks on odd-one-out trials as an alternate measure. We standardized RTs within subject by converting them to z scores. Note that for both of these measures, a smaller score reflects better performance.

It is traditional for the above types of analysis to adopt some criterion of successful category learning and exclude participants who don't meet it. However, the choice of the criterion is arbitrary, may well influence the results, and is not explicitly motivated in learned CP research. In addition, because our continuous staircasing procedure kept dimensional differences between adjacent stimuli near JND, we expected category learning to be relatively difficult and produce a wide range of performance levels. Since it seems reasonable to predict that learned CP measures should positively correlate with category learning success (see Gureckis & Goldstone, 2008 for a similar approach and positive evidence), we only reported that type of analysis.

Results

Figure 5 shows an example of a result of the Phase 1 staircase procedure for illustrative purposes. Participants whose Phase 1 JND on either dimension exceeded the maximum of .33 allowed in Phase 2 by more than .05 were excluded from subsequent analysis since the Phase 2 staircasing procedure would not apply correctly to them. This produced a final n of 72 (35 control, 37 experimental).

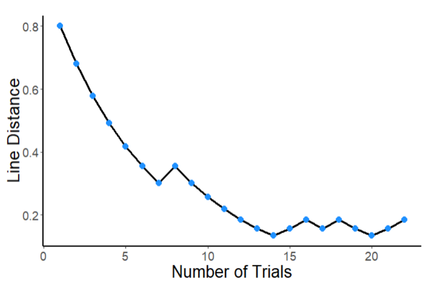

Figure 5. Example outcome of Phase 1 staircase procedure.

The mean proportion correct over Phase 2 on the classification task was .678 (SD = .145) and on the number judgment task it was .807 (SD = .13).

Phase 2 began with dimensional differences based on each individual participant's Phase 1 JND. Did the staircase procedure continuing throughout Phase 2 (in alternation with the classification or number judgment task) produce further perceptual learning? Figure 6 shows that in general, averaging across all participants, it did, particularly on the relevant dimension comparisons, as one might expect. Using the mean distance change for each participant averaging over the three odd-one-out trials differing on the relevant dimension (BE, WI1, and WI2) in the final block, the mean of the entire sample (M = -2.57) was significantly less than zero ($t(71)$ = -3.547, p < .0001). This was not the case for the irrelevant dimension (averaging over IRR1, IRR2, and IRR3 trials) (M = -0.72, $t(71)$ = -0.996, p = .16). A one-tailed paired samples t-test yielded a significant difference between relevant and irrelevant mean distance change ($t(71)$ = -2.014, p = .024).

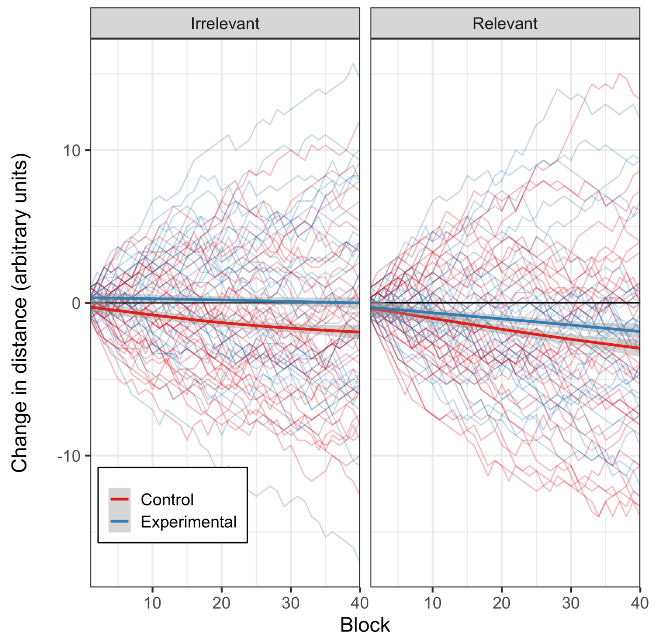

Figure 6. Overall perceptual learning in the experiment; the y axis represents number of staircase steps, e.g., a change in distance of -10 means the staircase has gotten 10 steps more difficult, indicating improved discrimination accuracy.

The left panel in Figure 7 illustrates the pattern that would be expected to hold for the control group, with better performance on the number judgment task coinciding with better performance on the odd-one-out task only (or to a greater degree) for the dimension relevant to the number judgment task, and no difference in the patterns for between and within category comparisons. The right panel shows what the pattern would be if the experimental group showed learned CP boundary effects, with better classification performance associated with better odd-one-out performance on between-category comparisons (expansion) and/or worse odd-one-out performance on within-category comparisons (compression) relative to the control group. If the experimental group were to show stronger sensitization to the relevant dimension or desensitization to the irrelevant

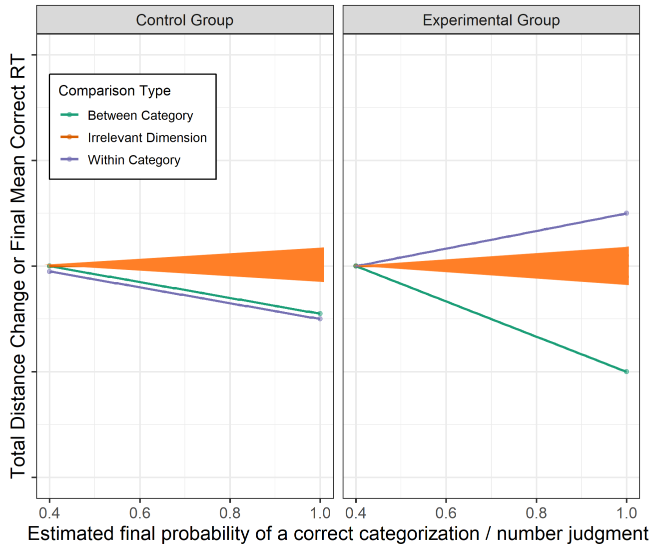

Figure 7. Relationship between classification or number judgment task performance (x axis) and either dependent variable (y axis) predicted by learned CP boundary effects.

dimension relative to the control group, the patterns would be slightly different, but we will focus on the boundary effects that are typically what is meant by learned CP.

Figure 8 shows the actual relationship in our data between total distance change over Phase 2 (y axis) and an estimate of the probability of a correct response at the end of Phase 2 on the number judgment (left) or classification (right) task (x axis) obtained by fitting a logistic regression model for each individual subject.

Overall there is a weak negative relationship such that discrimination performance tended to be better when classification or number judgment was more accurate, perhaps reflecting a general effect of effort. These data were analyzed using a Bayesian linear model to predict total distance change from three variables: comparison type (between, within, or irrelevant), group (control or experimental), and estimated final performance on the number judgment or classification task. The model also included the three-way interaction between these three variables since, as shown in Figure 7, this would have to be present if learned CP effects occurred. The analysis produced a BF_{10} of 119 for the estimated final performance variable, supporting the effort effect mentioned previously. To assess evidence for the critical three-way interaction, we determined the ratio of the BF_{10} for the full model containing the three predictor variables and the three-way interaction (.94) to the BF_{10} for the model containing just the three predictor variables (7.18). This yielded a BF_{10} of .131 indicating moderate support for H0 and therefore no evidence for learned CP.

The same analysis was performed for the RT measure (see Figure 9) and showed only one result favoring the alternative hypothesis and that was for comparison type

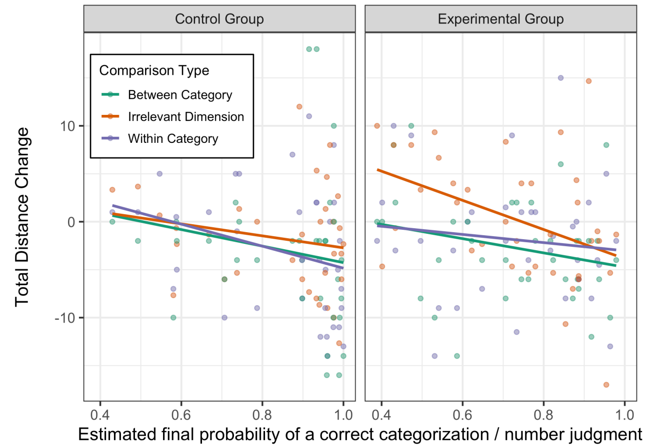

Figure 8. Relationship between estimated final performance on the classification or number judgment task and actual discrimination accuracy improvement over Phase 2 for the three comparison types.

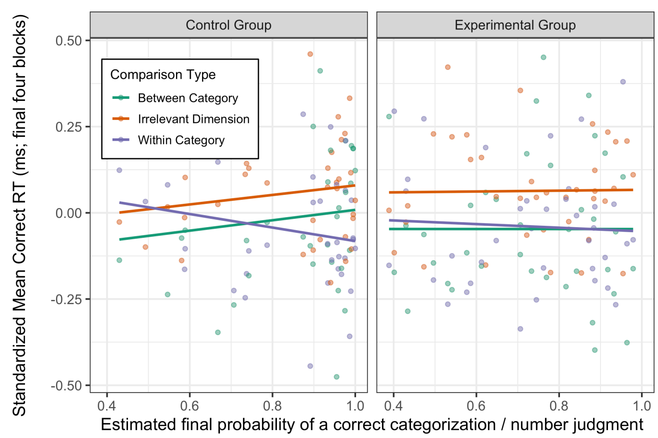

Figure 9. Relationship between estimated final performance on the classification or number judgment task and standardized mean correct RT on the last four blocks of Phase 2 for the three comparison types.

(BF_{10} = 41). The graph shows this to be due to irrelevant dimension odd-one-out trial responses being slower in general than relevant dimension trials of either type. The ratio of the BF_{10} for the full model containing the three predictor variables and the three-way interaction (.095) to the BF_{10} for the model containing just the three predictor variables (.942) yielded a BF_{10} of .101. This constitutes fairly strong support for H0 and thus no learned CP effects.

Discussion

This experiment employed a novel methodology designed to rigorously test for learned CP effects. Stimuli varied systematically on two dimensions, only one of which was relevant for either category learning or a control task. The stimuli also varied in random low-level features to allow for simultaneous presentation in the discrimination ("odd-one-out") task to eliminate reliance on memory. A staircasing procedure was used to initially determine the JND for each participant on each dimension and this staircasing continued with discrimination trials alternating with classification or control task trials to allow for continuous measurement of discrimination ability on each dimension.

The results provide evidence of sensitization to the relevant dimension for both the classification task and a control task that was comparable in requiring attention to one of the two dimensions. This was seen in significant discrimination performance improvement from the beginning to the end of Phase 2 for the sample as a whole on the relevant but not the irrelevant dimension. The interesting question then is whether there were differences in discrimination performance between the two groups that fit any of the patterns consistent with learned CP.

We did not report traditional analyses of learned CP effects, comparing successful category learners to the control group on our odd-one-out performance measures as a function of comparison type, due to the arbitrariness of setting a criterion for successful learning and the fact that our continuous staircasing procedure kept discrimination across the category boundary difficult. Instead, we examined whether learned CP effects appeared in the form of different *relationships* between category learning performance and discrimination performance as a function of comparison type and in relation to the control group.

The only effects we found were a positive correlation between success on the classification or number judgment task on the one hand and the odd-one-out task on the other, and slower response times by the end of the experiment for odd-one-out trials that required distinguishing stimuli differing on the irrelevant dimension. The critical three-way interaction between group, comparison type, and level of classification or number judgment performance that would be required in order to demonstrate any variety of learned CP effects was lacking for both dependent measures, and the analyses showed more than anecdotal support for its absence.

Note that if learned CP effects had occurred in this experiment, our continuous measurement of discrimination ability on the three types of comparisons would have been valuable for tracking the emergence of different types of effects (e.g., expansion vs. compression) and would have potentially allowed us to distinguish otherwise similar end results (i.e., expansion vs. a combination of compression and relevant dimension sensitization). However, since we did not obtain any learned CP effects overall, we were not able to take advantage of this capability.

There are many possible reasons for these negative results, due to the ways in which our methodology deviated from typical learned CP experiments. Perhaps the constantly changing stimulus set and its randomly varying sub-features below the dimensional level prevented learned CP from occurring. Or it may be that constantly alternating between a classification task and the odd-one-out task interfered with learned CP. If learned CP effects depend on memory and thus require tasks with a delay between stimuli in order to occur, our simultaneous stimulus presentation would be the cause. Or it could be that, previous evidence of boundary effects notwithstanding, so-called learned CP effects are really due to paying attention selectively to one dimension rather than another, and thus also occur as a result of other tasks besides category learning such as the number judgment task used by our control group.

We believe it is very important to determine the conditions under which learned CP effects do and do not occur, which has not been addressed sufficiently in the literature. Our negative results can provide a useful initial reference point. One strategy for building on this would be to next conduct a traditional version of the experiment utilizing a fixed set of the same stimuli and and a successive presentation version of our discrimination task to establish whether learned CP effects do occur under those conditions. If they do, methodological changes can then be incorporated one at a time, such as simultaneous rather than successive discrimination testing and comparison to a control group that performs a task requiring attention to one dimension, to determine which manipulations change and/or eliminate learned CP effects. This would allow us to make real progress in understanding the phenomenon of learned CP and its scope and limits.

Acknowledgments

This project was supported by the Vassar College Undergraduate Research Summer Institute.

References

Andrews, J. K., de Leeuw, J. R., Larson, C., & Xu, Xiaoqing. (2017). A preliminary p-curve analysis of learned categorical perception research. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1550-1555). Austin, TX: Cognitive Science Society.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*, 1-72.

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

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

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30thh Annual Conference of the Cognitive Science Society* (pp. 1876-1881). Austin, TX.: Cognitive Science Society.