UNIVERSITY of CALIFORNIA PRESS | Collabra

de Leeuw, J R et al 2016 The Effects of Categorization on Perceptual Judgment are Robust across Different Assessment Tasks. *Collabra,* 2(1): 9, pp. 1–9, DOI: http://dx.doi.org/10.1525/collabra.32

ORIGINAL RESEARCH REPORT

The Effects of Categorization on Perceptual Judgment are Robust across Different Assessment Tasks

Joshua R. de Leeuw*, Janet K. Andrews†, Kenneth R. Livingston† and Benjamin M. Chin‡

Learned visual categorical perception (CP) effects were assessed using three different measures (similarity rating, same-different judgment, and an XAB task) and two sets of stimuli differing in discriminability and varying on one category-relevant and one category-irrelevant dimension. Participant scores were converted to a common scale to allow assessment method to serve as an independent variable. Two different analyses using the Bayes Factor approach produced patterns of results consistent with learned CP effects: compared to a control group, participants trained on the category distinction could better discriminate between-category pairs of stimuli and were more sensitive to the category-relevant dimension. In addition, performance was better in general for the more highly discriminable stimuli, but stimulus discriminability did not influence the pattern of observed CP effects. Furthermore, these results were consistent regardless of how performance was assessed. This suggests that, for these methods at least, learned CP effects are robust across substantially different performance measures. Four different kinds of learned CP effects are reported in the literature singly or in combination: greater sensitivity between categories, reduced sensitivity within categories, increased sensitivity to category-relevant dimensions, and decreased sensitivity to category-irrelevant dimensions. The results of the current study suggest that the cause of these different patterns of CP effects is not due to either stimulus discriminability or assessment task. Other possible causes of the differences in reported CP findings are discussed.

Keywords: categorization; categorical perception; compression; expansion; learning; similarity; online experiments; Bayes factor

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 review, see [1]). Learning that two stimuli belong to the same category can increase their similarity or perceptual confusability, an effect known as compression (e.g., [2]), while learning that two stimuli belong to different categories can have the opposite effect, often called expansion (e.g., [3–4]). Categorizing stimuli based on particular features may increase sensitivity to those features, regardless of whether the stimuli belong to the same or different categories, or it may reduce sensitivity to features that are not relevant to the categories (e.g., [3]).

Although there are these distinct 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, as the different CP effects implicate different kinds of mechanisms. For example, some models predict a change in overall sensitivity to a dimension after categorization training due to dimension-based attention weighting (e.g. [5–6]). However, sensitization to a particular range of values along a dimension requires different kinds of models that allow for changes in perceptual sensitivity at a subdimensional level (e.g. [7–8]).

Several studies have demonstrated that whether CP is observed at all depends on various factors including (1) the availability of verbal labels during perceptual testing [9], (2) the particular kind of perceptual assessment used to determine whether CP is present [10], and (3) how stimulus morphspaces are created [11]. However, few studies have reported differences in the kind of CP effect observed based on experimental manipulations, though there are exceptions. For example, Goldstone, Lippa, and Shiffrin [12] and Livingston and Andrews [13] both reported two qualitatively different CP effects exhibited by the same subjects when they were tested in two

* Indiana University, Bloomington, IN, USA

† Vassar College, Poughkeepsie, NY, USA

‡ University of Pennsylvania, Philadelphia, PA, USA

Corresponding author: Joshua R. de Leeuw (josh.deleeuw@gmail.com)

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 [14] used texture patterns and found larger expansion effects for stimulus sets that were harder to differentiate, which is consistent with this idea. This would also explain why studies using near-JND stimulus differences have tended to show expansion (e.g., [3–4]) while studies using stimulus differences well above JND have tended to report compression effects instead (e.g., [2]). However, in many cases, the difference in stimulus discriminability across studies coincided with a difference in dependent measure, complicating the interpretation.

In the experiment reported here, we directly tested the influence of both 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. Different tasks may not invite equivalent perceptual/cognitive strategies, so this is a reasonable potential source of different patterns of results between studies. For example, subjective rating tasks such as similarity judgments may invite strategic responses (altering the rating based on the category labels without warping of perceptual similarity) and thus produce CP effects even when objective measures such as same-different or XAB do not, especially in cases where perceptual learning is not necessary for categorization (i.e., when stimuli are highly discriminable). If this is the case, assessment task and stimulus discriminability should interact. We included both the XAB and same-different tasks since both are frequently used in the literature, though we expected them to produce similar results.

Method

Participants

The Vassar College Institutional Review Board approved the procedures used in this study and consent was obtained as part of the online data collection. We recruited 564 participants through Amazon Mechanical Turk (AMT).[1] Eight subjects were excluded from the analysis because of a bug that allowed them to complete more than one condition of the experiment, leaving 556 participants.

Materials

The experiment was developed using jsPsych, a software library for building online experiments [15]. Stimuli were cell-like shapes that varied on two dimensions, shape and tail length, created using Mathematica 8.0 (for details, consult the stimulus creation script available in the project repository). We generated two sets of stimuli: a high discriminability (HD) set and low discriminability (LD) set. The LD stimuli had half as much variation as corresponding HD stimuli (see **Figure 1**). The stimulus space was

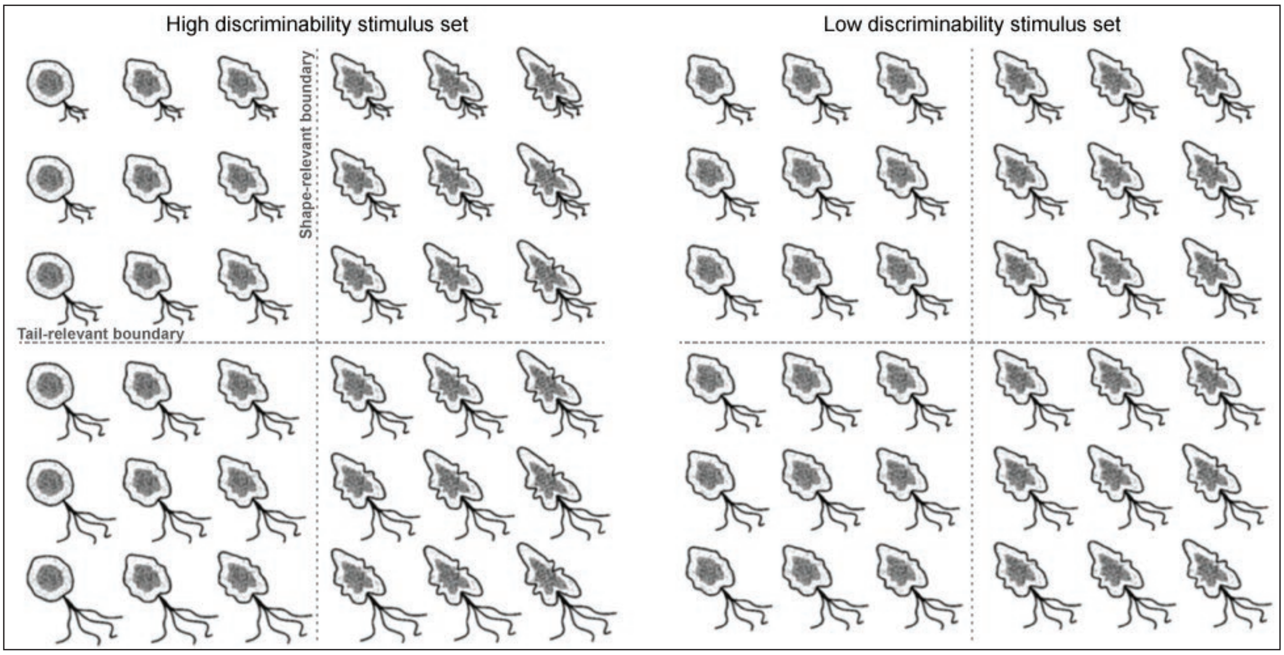
High discriminability stimulus set
Low discriminability stimulus set
Shape-relevant boundary
Tail-relevant boundary

Figure 1: The two sets of stimuli used in the experiment. Shape varies along the X axis and tail length varies across the Y axis.

6 (shape) by 6 (tail length) for both sets. We chose the values along each dimension by arbitrarily selecting the two endpoints and creating equal numeric intervals between neighboring items; however, we did not test whether the psychological distance between neighbors was equivalent throughout the space, nor do we expect that it is.

When shape was the category relevant dimension, the category boundary was between the 3rd and 4th stimuli on the shape dimension and the tail length dimension was irrelevant. When tail length was the category relevant dimension, the category boundary was between the 3rd and 4th stimuli on the tail length dimension and the shape dimension was irrelevant.

Procedure

Participants were randomly assigned to one of 24 conditions: 2 training type (control v. category training) X 2 stimulus sets (HD v. LD) X 3 type of assessment (similarity v. same-different v. XAB) X 2 category-relevant dimensions (shape v. tail length). There were 18–28 participants per condition after excluding subjects who failed the training phase (see Results). The category-relevant dimension variable was not of theoretical interest here but is included in the analyses for completeness, as suggested by an anonymous reviewer.

Training. Participants who were assigned to a training condition completed an adaptive training protocol by iterating through multiple blocks of training until category assignments were learned for all stimuli. 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. Participants were told that they would initially need to guess which category a cell belonged to, 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 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.[2]

Post-training assessment tasks. Participants completed one of three different assessment tasks. Those who received category training completed the task immediately after the category learning test, while control participants only performed the assessment task.

Similarity. In the similarity task, participants saw two stimuli sequentially. Each stimulus was visible for 750ms, with a blank screen displayed for 1000ms between stimuli. The participant dragged a movable slider to indicate how similar the two stimuli were using a continuous (no segmentation) scale 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 items in a pair). 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, participants saw two stimuli sequentially, each for 750ms, with a blank screen displayed for 1000ms between stimuli. They pressed one of two keys 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 non-identical pairs. The 27 non-identical pairs were selected according to the same policy as used in the similarity task, except that only 3 pairs per type were chosen instead of 4 in order to limit the overall length of the experiment.

XAB. In the XAB task, participants 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. Participants pressed a key to indicate whether A or B was identical to X. There were 4 blocks of 36 pairs of stimuli, selected in the same manner as in the similarity task.

Data and Analysis Files

Our data and analysis files are available in an Open Science Framework repository at https://osf.io/54rve/.

Results

Training

A total of 39 participants failed training and were excluded from the analysis. Of those 39 exclusions, 38 were for participants for whom the less salient dimension of tail length was the category-relevant dimension. Over all training conditions, the mean number of trials required to complete training was 220 with a standard deviation of 48. The minimum number of trials to complete training by any subject was 149, and the maximum was 386.

CP Effects

Between-category versus within-category pairs. The data set for analysis of between vs. within category pairs was restricted to pairs that varied only on the category-relevant dimension, with a maximum distance of 2 between the items. This emphasizes the effect of category learning on measures of CP. The pairs that were 3 units apart were not used because all such pairs were between category and have no within-category counterparts. The between-category and within-category scores for each participant consisted of mean rating for the similarity task and proportion of correct responses for the same-different and XAB tasks on the relevant subset of item pairs.

Dimension influence. This analysis was restricted to item pairs differing by 1 or 2 units on one dimension (and not differing on the other dimension). Each participant received a score (mean similarity rating or mean proportion correct) for the relevant dimension and the irrelevant dimension.

Conversion to a common scale. In order to make performance across the three different assessment tasks comparable and to directly test whether assessment task plays a role in producing learned CP effects, we took each participant's scores as explained above and transformed them as follows: First, the similarity scores were reverse coded so that higher scores represented greater differentiation of items, to be comparable to the same-different and XAB measures. For the between-category versus within-category pairs analysis, we determined the mean and standard deviation of all raw scores for a given assessment task for the items used in that analysis. Each participant's score was transformed by subtracting that mean and dividing by that standard deviation. The same thing was done for the dimension influence data. Transforming the data in this way provides a common scale for the three assessment tasks, making it easier to incorporate assessment task into the statistical analysis as an independent variable. The nature of the transformation is such that there is no possibility of a main effect of assessment task on performance; however, that would not be possible to determine using the raw scores either, and the transformation makes it possible to determine whether assessment task interacts with other independent variables to produce different patterns of CP effects, which is the primary goal.

Statistical analysis. We used the Bayes factor approach for all statistical analyses because of the importance of being able to quantify relative evidence for the null and alternative hypotheses concerning assessment method, stimulus discriminability, and their interactions with other independent variables. Statistical analyses were done using the BayesFactor package [16] in R [17].

The BayesFactor package places a sensible default, vaguely-informed prior on the standardized effect size of the model predictors [18]. The only parameter for the prior is a scaling factor, which controls how concentrated the prior probability density is around the standardized effect size of 0. We used a scaling factor of 0.5 on the prior, which puts half of the prior probability on standardized regression coefficients between -0.5 and 0.5. This reflects an expectation of finding small-to-moderate effects. Smaller scaling factors would generally increase the odds in favor of the alternative models, while larger scaling factors would increase the odds in favor of the null.

Between-category versus within-category pairs. The first analysis was designed to determine whether boundary CP effects occurred and, if so, whether they varied according to assessment task and/or stimulus discriminability. Using the transformed pair type dependent measure, we tested a set of linear models to quantify the support for each model. First we tested just for main effects of four of the five independent variables: stimulus set (HD v. LD), training type (control v. category training), pair type (between-category v. within-category), and category-relevant dimension (shape v. tail length). We did not test for a main effect of assessment type because the dependent variable was normalized to remove any main effects of assessment type (see previous section). For each effect, we found the Bayes factor for a model that included the effect and the random effect of subject against the model that just included the random effect of subject. The results are shown in the Appendix. There is very strong evidence for models including a main effect for stimulus set, pair type, and category-relevant dimension. The main effect of stimulus set can be seen in the smaller graphs on the right of **Figure 2a**, which show that HD performance consistently tends to be above 0 and LD performance tends to be below 0. The very strong main effect of category-relevant dimension is presumably due to the boundary analysis being restricted to item pairs differing only on the category-relevant dimension and, as shown by the training data, the categories were much harder to learn when they were defined by tail length rather than shape.

We then tested the model that includes all main effects, the random subject effect, and the training type by pair type interaction (which would constitute evidence for a learned CP effect) against the model that contains only the main effects and the random subject effect. The result was strong support for the model that includes the interaction and is shown in **Figure 2a**; $BF_{10} = 71.66$

The next analysis tested whether assessment type interacted with the CP effect. We tested all models that included the assessment type by training type by pair type interaction or other higher-order interactions of this 3-way interaction with the other independent variables, against the model that included only the CP interaction, the main effects, and the random subject effect. In other words, we tested all of the models in which the CP interaction (training type by pair type) interacted with assessment type. This set of model comparisons is a direct test of whether the assessment type modulates the CP interaction. None of the models that include interactions with the CP interaction received more support than the model with just the CP interaction. For all but one of the 15 tests, the Bayes factor was better than 3:1 in favor of the CP-only model (see Appendix for full results). A similar analysis was performed to test all models that include interactions between the CP interaction and the stimulus set and once again, none of the models that include those interactions received more support than the corresponding model without the interaction. However, the Bayes factors were generally equivocal between the CP-only model and the interaction of stimulus set with CP. The three-way interactions all had Bayes factors between 1 and 3 in favor of the CP-only model. Only the higher order interactions and models with multiple interactions had Bayes factors larger than 3 in favor of the CP-only model. The larger Bayes factors in favor of the CP-only model are to be expected as a natural result of increasing the complexity of the model when all the lower-order components of the model slightly favor the simpler CP-only model.

Dimension influence. The second analysis was designed to determine whether dimensional sensitivity CP effects occurred and, if so, whether they varied according to assessment task and/or stimulus discriminability. The

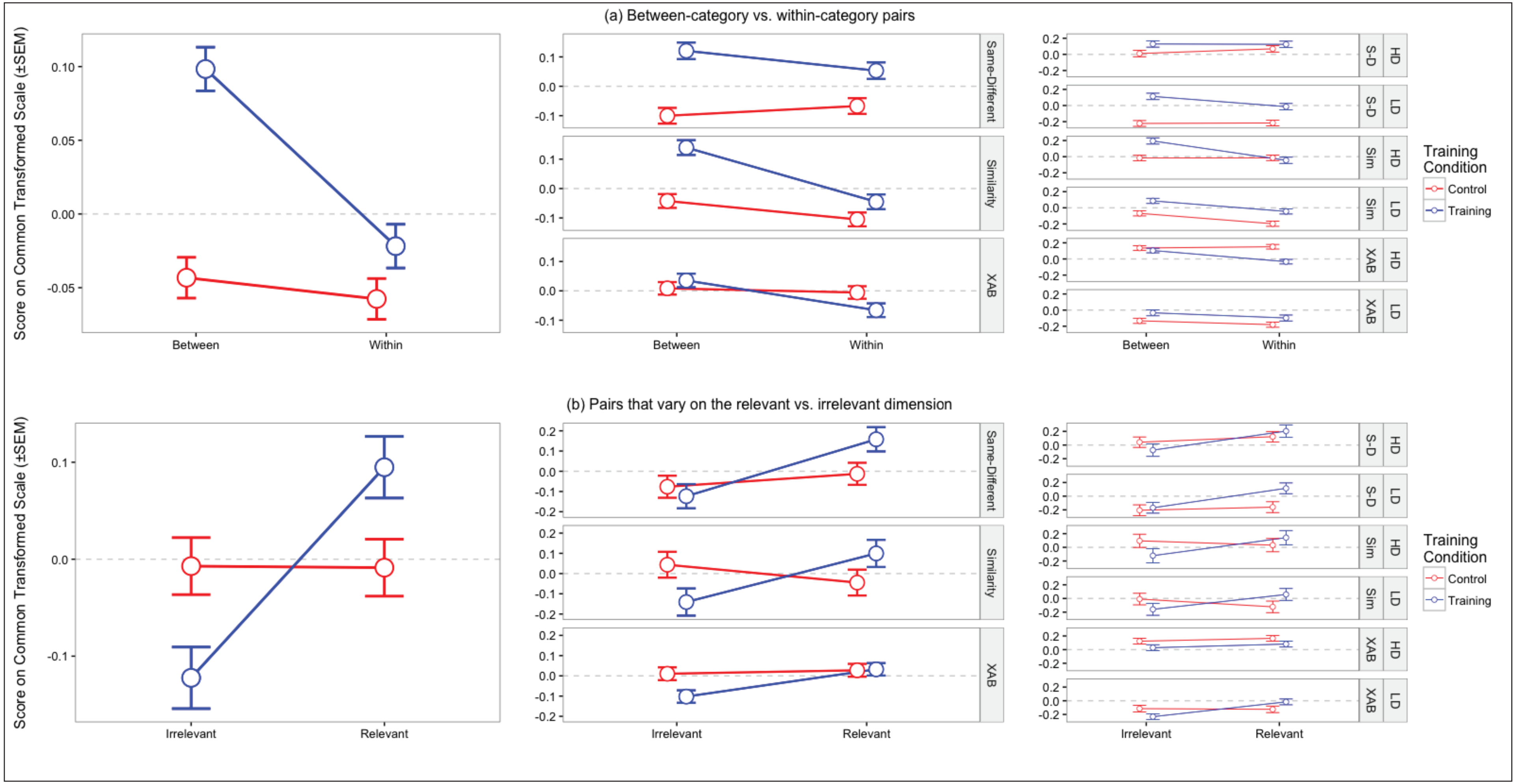

Figure 2: (a) Performance on between-category versus within-category pairs by the control and training groups. **(b)** Performance on pairs differing only on the irrelevant versus the relevant dimension by the control and training groups. In both panels, the large graph on the left shows the overall interaction across task type. The three graphs in the middle show the interaction for each task, and the six smaller graphs on the right show the interaction for each task and each stimulus type. Higher scores on the Y axis represent greater differentiation between items.

exact same analyses described above were carried out on the transformed dimensional influence dependent measure. There was very strong evidence for models including a main effect for stimulus set and varying dimension (see Appendix for results) and the varying dimension by training type interaction, which is shown in **Figure 2b** (BF_{10} = 166.43). For all models including interactions between the CP-interaction and assessment type and/or stimulus set, the Bayes factor was better than 4:1 for the CP-only model (see Appendix for full results).

Bayes factor t-tests were also performed to determine the specific nature of the CP effects shown in the interactions with training type. We used the default Bayes factor t-test method from the BayesFactor R package, with a scaling factor of 0.5 on the prior, indicating a prior belief that the standardized effect size will generally be small when the null is false. The four types of CP effects imply that the effect will be in a particular direction, e.g., compression requires that trained subjects show *less* differentiation on within-category judgments. We used alternative priors that built in this one-sided constraint [19]. For the boundary analysis, BF_{10} = 92.12 for the between-category pairs and BF_{01} = 13.12 for the within-category pairs, clearly demonstrating expansion but not compression. These data strongly favor the null in the case of compression. For the dimension analysis, BF_{10} = 11.16 for the relevant dimension and BF_{10} = 6.15 for the irrelevant dimension, suggesting stronger evidence of increased sensitivity to the relevant dimension, but also support for decreased sensitivity to the category-irrelevant dimension. We also tested variations in the scaling factor to examine the robustness of the BF t-tests to different prior beliefs. Even with a scaling factor of 1.0, double the scaling factor we used, the Bayes factors for expansion, acquired equivalence, and acquired distinctiveness were still all better than 3.5 to 1 in favor of the alternative. Scaling factors less than 0.5 further increased the relative support for the alternative models.

Discussion

This study was designed to use procedures typical of the research literature on learned visual CP effects. Accordingly, CP effects were examined in the context of successful category training with an artificial stimulus set. Of the four possible patterns of learned CP effects, we found positive evidence for expansion (enhanced ability to distinguish between-category item pairs as a result of learning the categories), increased sensitivity to the category-relevant dimension, and decreased sensitivity to the category-irrelevant dimension. There was no evidence for compression. In fact, the data indicated strong support for the null that there was no compression effect.

The main purpose of this study was to explore two candidate reasons, stimulus discriminability and assessment task, for why different studies on learned CP find different patterns of CP effects. The stimulus discriminability manipulation was effective in that performance was clearly reduced for the LD stimuli relative to the HD stimuli. However, there was no strong evidence that the CP effects were influenced by stimulus discriminability – the analysis was generally equivocal about whether stimulus discriminability interacted with boundary-based CP, and the Bayes factor favored the CP-only model for the dimension effect – despite some prior evidence that expansion is stronger when stimuli are more difficult to discriminate [14]. This result also suggests that the cause of expansion as opposed to compression effects in the literature is not due to stimulus discriminability.

Assessment task also did not influence the pattern of CP effects obtained; the evidence was overwhelmingly in support of the null hypothesis for interactions of assessment task with stimulus discriminability, category training, pair type, and dimensional relevance. This is an important result as no previously published study that we know of has directly compared different methods of assessing learned CP using the same stimuli and categories. This suggests that the variations in CP effects reported in the literature are not due to the use of different tasks, at least among similarity rating, same-different judgments, and XAB, which are the most common tasks used. It is reassuring that the CP effects obtained were so robust across these very different tasks, and this also suggests that current models of learned CP, which do not address the role of assessment task, may well not need to do so.

This result can be reconciled with the few previous studies that suggested that assessment task might play a role. Goldstone, Lippa, and Shiffrin [12] used morphed face stimuli and a similarity rating task. They found expansion using a traditional between-category versus within-category pair analysis, but found compression when stimuli were judged relative to a neutral, uncategorized face. Thus their differing patterns of results were not due to different assessment tasks but rather a different type of comparison stimulus. Livingston and Andrews [13] used artificial "alien" figures and obtained expansion using a similarity rating task and compression using a same-different task, but without rendering data from the two tasks comparable in the way that we did here and treating task as an independent variable, we cannot but sure that task had a significant effect on the pattern of learned CP results. Even if it did, it would be important to determine whether that result is replicable and genuinely conflicts with the results of the current study.

The only other apparent counterexample of which we are aware comes from Gerrits and Schouten [10], who obtained CP effects under some conditions and not others as a function of assessment task. These authors did directly compare performance using two different assessment tasks, but their study differed in many critical ways from a typical visual learned CP study such as the one reported here: the stimuli consisted of speech sounds and no category training was involved. The fact that there was no learning means that compression vs. expansion effects cannot be compared as was done in the present study. What Gerrits and Schouten [10] did observe was presence of CP for vowels using one measure (2I2AFC or two interval two alternative forced choice) but not when using a variant of that measure (4I2AFC or four interval two alternative forced choice). They conclude that the disappearance of the CP effect in the latter case is owing to features of the task that render listeners "incapable of using available phonetic information that would have improved their performance" ([10], p. 375). This seems a limiting case in which failure to process the information relevant for a category distinction results in failure to show CP, which is hardly surprising. The fact that there are measurement tasks that can have this effect is interesting but is not in itself inconsistent with the conclusion that assessment task does not affect whether CP effects are the result of compression vs. expansion when they do occur.

It is possible that different results would occur even for visual learned CP and these same three tasks if the stimuli and/or category structures were very different from those used in the current study. In fact, if individually unique and distinctive stimuli were used, a same-different or XAB task would surely fail to show learned CP effects that a similarity rating task might well show (as occurred, for example, in a study by Livingston, Andrews, and Dwyer [20] that produced compression using real tropical fish as stimuli and similarity ratings as the assessment method). But very few studies use such stimuli, and the current results suggest that factors other than assessment task and discriminability are likely the cause of the variations in learned CP effects reported in the literature.

One such factor may well be category structure, but very few studies have examined this directly. Reppa and Pothos [21] used simple geometric stimuli and reported differences in the pattern of learned CP results as a function of category structure. Categories with one relevant and one irrelevant dimension of variation and non-linearly separable categories both produced compression, while a diagonal category structure in which both dimensions were relevant produced expansion. Given that the categories in the study reported here were also based on one dimension of variation yet produced expansion with the same similarity rating task used by Reppa and Pothos, it is clear that more than just category structure is involved.

Another possibility is that the variation in reported findings is the result of statistical noise. Learned CP effects caused by a relatively short period of training are generally not very large effects, and sampling noise combined with low power could explain the variation in effects. Our data illustrate this point quite well. Note that in **Figure 2b** (central panel), the same-different and XAB tasks appear to show opposite results, with increased sensitivity to the relevant dimension in the same-different task and decreased sensitivity to the irrelevant dimension in the XAB task. If these tasks were analyzed separately, such as typically occurs when comparing results across experiments, we would wrongly conclude that task *does* matter because we would observe a statistically significant acquired equivalence effect in the XAB task and a statistically significant acquired distinctiveness effect in the same-different task. However, the full statistical model makes it quite clear that these variations are simply noise and that the general pattern is stable across task. An analogous situation may be affecting the literature as a whole when comparisons between the patterns of statistically significant results are made without considering whether or not the differences in the patterns themselves are statistically significant [22]. Comparing across different experiments in the literature is therefore often misleading.

A thorough examination of the learned CP literature shows that studies differ not only in the specific stimuli, features or dimensions, and category structures used, but also in the specific ways in which the effects of category learning are measured. Some of these differences are subtle but could be important for determining what conditions create each of the four types of learned CP effects identified at the outset. For example, for studies that only analyze between-category versus within-category pair data, it will be unknown whether dimensional sensitivity effects also occurred. In addition, both types of analyses may be influenced by the composition of the item pairs used for collecting and/or analyzing the data. Solid evidence that stimulus discriminability and assessment task do not determine the observed pattern of effects allows us to turn our attention to these other factors in order to better understand the complex phenomenon of learned CP.

Supplementary Files

The supplementary material for this article can be found here: http://dx.doi.org/10.1525/collabra.32

Acknowledgments

This research is supported by a National Science Foundation Graduate Research Fellowship to the first author under Grant No. DGE-1342962. We thank the Vassar College Phoebe H. Beadle Science Fund Endowment and Research Committee for financial support, and Samuel Schwamm and Calais Larson for assistance with the project. Earlier versions of this work were originally reported at the 35th and 36th Annual Meetings of the Cognitive Science Society.

Competing Interests

The authors declare that they have no competing interests.

Notes

[1] Samples from AMT tend to replicate laboratory findings, though category-learning tasks on AMT have produced mixed results [23]. Because 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.

[2] A data recording error resulted in missing data for more than 50% of the participants for this particular task, so we do not report any analysis of this data.

References

1. Goldstone, R. L., and Hendrickson, A. T. 2009. Categorical perception. Interdisciplinary Reviews: Cognitive Science 1: 69–78.
2. Livingston, K. R., Andrews, J. K., and Harnad, S. 1998. Categorical perception effects induced by category learning. Journal of Experimental

Psychology: Learning, Memory, and Cognition 24(3): 732–753. DOI: http://dx.doi.org/10.1037/0278-7393.24.3.732

3. Goldstone, R. L. 1994. Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General 123(2): 178–200. DOI: http://dx.doi.org/10.1037/0096-3445.123.2.178
4. Notman, L., Sowden, P. T., and Özgen, E. 2005. The nature of learned categorical perception effects: a psychophysical approach. Cognition 95(2): B1–14. DOI: http://dx.doi.org/10.1016/j.cognition.2004.07.002
5. Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115(1): 39–57. DOI:http://dx.doi.org/10.1037/0096-3445.115.1.39
6. Kruschke, J. K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review 99(1): 22–44. DOI: http://dx.doi.org/10.1037/0033-295X.99.1.22
7. Goldstone, R. L., Steyvers, M., and Larimer, K. 1996. Categorical perception of novel dimensions. In Proceedings of the 18th Annual Conference of the Cognitive Science Society. Hillsdale, New Jersey: Larwrence Erlbaum Associates. pp. 243–248.
8. Love, B. C., Medin, D. L., and Gureckis, T. M. 2004. SUSTAIN: a network model of category learning. Psychological Review 111(2): 309–332. DOI: http://dx.doi.org/10.1037/0033-295X.111.2.309
9. Kikutani, M., Roberson, D., and 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. DOI: http://dx.doi.org/10.3758/PBR.15.4.787
10. Gerrits, E., and Schouten, M. E. H. 2004. Categorical perception depends on the discrimination task. Perception & Psychophysics 66(3): 363–76. DOI: http://dx.doi.org/10.3758/BF03194885
11. Folstein, J. R., Gauthier, I., and 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. DOI: http://dx.doi.org/10.1037/a0025836
12. Goldstone, R. L., Lippa, Y., and Shiffrin, R. M. 2001. Altering object representations through category learning. Cognition 78(1): 27–43. DOI: http://dx.doi.org/10.1016/S0010-0277(00)00099-8
13. Livingston, K. R., and Andrews, J. K. 2005. Evidence for an age-independent process in category learning. Developmental Science 8(4): 319–25. DOI: http://dx.doi.org/10.1111/j.1467-7687.2005.00419.x
14. Pevtzow, R., and Harnad, S. 1997. Warping similarity space in category learning by human subjects: The role of task difficulty. In Ramscar, M., Hahn, U., Cambouropolos, E., and Pain, H. (Eds.), Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization. Department of Artificial Intelligence, Edinburgh University. pp. 189–195.
15. de Leeuw, J. R. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods 47(1): 1–12. DOI: http://dx.doi.org/10.3758/s13428-014-0458-y
16. Morey, R., Rouder, J., Love, J., and Marwick, B. 2015. BayesFactor: 0.9.12-2 CRAN. Zenodo. DOI: http://doi.org/10.5281/zenodo.31202
17. R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.
18. Rouder, J. N., and Morey, R. D. 2012. Default Bayes factors for model selection in regression. Multivariate Behavioral Research 47(6): 877–903. DOI: http://dx.doi.org/10.1080/00273171.2012.734737
19. Morey, R., and Rouder, J. 2011. Bayes factor approaches for testing interval null hypotheses. Psychological Methods 16(4): 406–419. DOI: http://dx.doi.org/10.1037/a0024377
20. Livingston, K. R., Andrews, J. K., and Dwyer, P. 2001. Ties that bind: Reconciling discrepancies between categorization and naming. In Moore, J., and Stenning, K. (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum. pp. 558–563.
21. Reppa, I., and Pothos, E. 2013. Predicting similarity change as a result of categorization. In Knauff, M., Pauen, M., Sebanz, N., and Wachsmuth, I. (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. pp. 1211–1216.
22. Gelman, A., and Stern, H. 2006. The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician 60(4): 328–331. DOI: http://dx.doi.org/10.1198/000313006X152649
23. Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS ONE 8(3): e57401. DOI: http://dx.doi.org/10.1371/journal.pone.0057410

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://dx.doi.org/10.1525/collabra.32.opr

How to cite this article: de Leeuw, J R, Andrews, J K, Livingston, K R and Chin, B M 2016 The Effects of Categorization on Perceptual Judgment are Robust across Different Assessment Tasks. *Collabra,* 2(1): 9, pp. 1–9, DOI: http://dx.doi.org/10.1525/collabra.32

Submitted: 18 August 2015 **Accepted:** 27 May 2016 **Published:** 28 July 2016

UNIVERSITY
of CALIFORNIA
PRESS
Collabra

Collabra is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS