AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2018 American Psychological Association
0278-7393/19/$12.00

2019, Vol. 45, No. 7, 1166–1176
http://dx.doi.org/10.1037/xlm0000609

Bias to (and Away from) the Extreme: Comparing Two Models of Categorical Perception Effects

Ryan M. Best and Robert L. Goldstone
Indiana University

Categorical perception (CP) effects manifest as faster or more accurate discrimination between objects that come from different categories compared with objects that come from the same category, controlling for the physical differences between the objects. The most popular explanations of CP effects have relied on perceptual warping causing stimuli near a category boundary to appear more similar to stimuli within their own category and/or less similar to stimuli from other categories. Hanley and Roberson (2011), on the basis of a pattern not previously noticed in CP experiments, proposed an explanation of CP effects that relies not on perceptual warping, but instead on inconsistent usage of category labels. Experiments 1 and 2 in this article show a pattern opposite the one Hanley and Roberson pointed out. Experiment 3, using the same stimuli but with different choice statistics (i.e., different probabilities of each face being the target), obtains the same pattern as the one Hanley and Roberson showed. Simulations show that both category label and perceptual models are able to reproduce the patterns of results from both experiments, provided they include information about the choice statistics. This suggests 2 conclusions. First, the results described by Hanley and Roberson should not be taken as evidence in favor of a category label model. Second, given that participants did not receive feedback on their choices, there must be some mechanism by which participants monitor their own choices and adapt to the choice statistics present in the experiment.

Keywords: categorical perception, categorization, face perception

There has been some debate in the last several years concerning the cause of categorical perception effects (CP), which manifest as better discriminability between objects belonging to different categories than objects from within the same category, when controlling for the physical differences between the objects. The colors of a rainbow, for instance, vary continuously along a spectrum, but we perceive them as distinct bands according to the color categories that we know, with clear distinctions between bands of color but not within them. Such effects have been found for both auditory and visual stimuli (e.g., Pisoni & Tash, 1974 in the auditory domain; Goldstone, Steyvers, & Rogosky, 2003 in the visual domain; for a review see Goldstone & Hendrickson, 2010). Two kinds of explanation for this phenomenon have been proposed, which we call *perceptual effect* (PE) explanations and *category label* (CL) explanations. The PE explanations include those by Goldstone and colleagues (Goldstone, Steyvers, Spencer-Smith, & Kersten, 2000), and generally posit that people's perception of objects is affected by their categorization of those objects. The CL explanations include those by Hanley and Roberson (2011), who argued that perception of an object activates a category label, and CP effects can be explained by inconsistent use of these labels near category boundaries.

Here we present results from three experiments that show CP effects. We show that the pattern of results that led Hanley and Roberson (2011) to propose that their CL account can be reversed by changing the distribution of the stimuli in a CP experiment. Simulation results from two models show that both PE and CL explanations, when combined with information about the choice statistics in each experiment, can account for the results.

Perceptual Effects Versus Category Labels

The PE explanations of CP effects share the central idea that these effects occur because people's representations of objects, either in memory or as they relate to other parts of the environment, are affected by their categorization of the objects. As an example, Goldstone et al. (2000) described a neural network model of CP in which weights between input nodes and hidden nodes are updated in an unsupervised manner, and weights between these hidden nodes and category output nodes are updated with supervised learning. Given different category labels for the same stimuli for the supervised learning, the network will learn different weights between the input and hidden layers. Thus, the network's representations of the stimuli in these weights depend on the labels given to the stimuli. Though other PE accounts differ in their details, they share this feature of representational warping resulting from categorization experience (e.g., Harnad, Hanson, & Lubin, 1995; Livingston, Andrews, & Harnad, 1998).

Hanley and Roberson (2011) challenged the PE accounts and claim that an effect of categories on perception is not necessary to

This article was published Online First July 19, 2018.

Ryan M. Best and Robert L. Goldstone, Department of Psychological and Brain Sciences, Indiana University.

Ryan M. Best is now at Capital One.

Correspondence concerning this article should be addressed to Ryan M. Best, 8100 Gaylord Parkway, Apartment 1234, Frisco, TX 75034. E-mail: rmbestcogs@gmail.com

explain CP effects. Instead, they argued that such effects can be explained by the perceiver's inconsistent use of category labels near category boundaries. The authors supported their argument with an analysis of several existing experiments that tested CP using a two-alternative forced choice XAB task. In this task, participants are briefly presented with a stimulus (the X stimulus) that comes from a continuum that spans the boundary between two categories (e.g., a shade of blue on a continuum from blue to green). Then, participants are shown the target stimulus alongside a foil from the same continuum (A and B, respectively) and are asked to identify which one is identical to the original (X) stimulus. Here, CP manifests as better accuracy on trials where the target and foil come from different categories (i.e., between-category trials) and worse accuracy on trials where the target and foil come from the same category (i.e., within-category trials).

Hanley and Roberson (2011) found that accuracy on within-category trials in their experiments was not uniformly worse than on between-category trials. Instead, the accuracy depended on whether the target or foil was more extreme (i.e., farther from the category boundary; see Figure 1, Panel a). In the data sets they analyzed, accuracy on within-category trials was only worse than on between-category trials when the foil was more extreme than the target. The authors argued that these results cannot be explained by perceptual warping. Their explanation instead relies on category labels to account for the CP effects. Their model assumes that when participants make a choice in the XAB task, they do not have access to an accurate representation in memory of the original stimulus. Instead, participants use the category labels of the target and the foil to make a decision. Objects that are near the category boundary are categorized more slowly and less consistently than those that are farther from the boundary, leading to the asymmetry on within-category trials.

In the next section, we present the results of three XAB experiments using faces morphed on a continuum from Black to White. All three experiments show the typical CP result, with participants more accurate on between-category trials than within-category trials. The first two also show a pattern of results that is the opposite of what Hanley and Roberson's (2011) account predicts with regard to the bias toward selecting extreme faces; participants are more accurate when the foil is more extreme than the target. The third experiment, using the same stimuli as the first two, shows the pattern of results that Hanley and Roberson's model predicts. The major change between the first two experiments and the third experiment that can account for the difference, as well as Hanley and Roberson's original results, is the distribution of stimuli along the continuum (details in the next section). This suggests that the pattern of results that Hanley and Roberson's account is supposed to explain is a consequence of the stimulus distribution in their experiments.

Experiment 1

Method

Participants. One hundred nine workers on Amazon's Mechanical Turk completed the experiment for $2.00. The racial makeup of the participants was 61% White and 12% Black. We excluded 1 participant from analyses because the participant's accuracy was below chance.

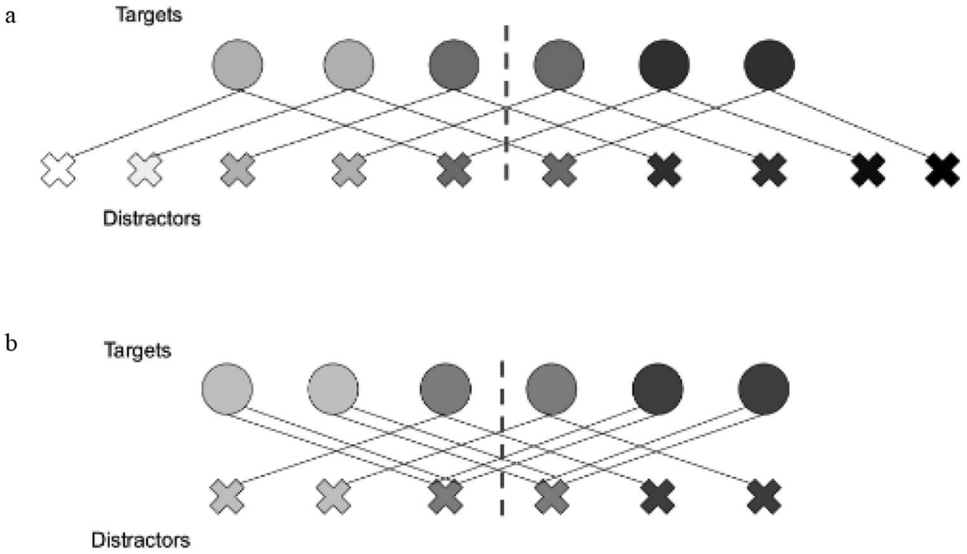

Figure 1. Panel a: Stimulus distribution used in Experiments 1 and 2. Circles represent targets, and X's represent distractors. Circles and X's that are aligned vertically represent the same face. Each line connecting a circle to an X represents one pair presented in the XAB task. For example, the leftmost circle has lines connecting it to the leftmost X and to the fifth X from the left, indicating that it was presented equally often with each of those faces as distractors. The dotted line marks the category boundary. Faces that are farther from the category boundary are considered more extreme. The leftmost X represents a more extreme face than the leftmost circle, for instance. Panel b: Stimulus distribution used in Experiment 3. Target–foil pairs connected by two lines were presented twice as often as those connected by one line.

Materials. The stimuli used in all three experiments were 10 faces created by morphing between two endpoint faces. The endpoints were pictures of a Black male and a White male, selected from a collection of photographs of bald heads (Kayser, 1985). The morphed faces were created using a program written by Steyvers (1999). For the two endpoint faces, 60 control points were placed at landmarks on the face, such as the corners of the eyes, tip of the nose, and so forth The morphs were created by moving each control point to a location some distance along the line connecting the control points on the endpoint faces, and making the grayscale value for each pixel a weighted average of the values at the corresponding points on the endpoint faces. Figure 2 below shows the morphed faces resulting from this procedure. The third face from the left, for example, was created by placing each of its control points [1/4] of the way from the control points on the White endpoint face to the Black endpoint face, and assigning the grayscale values of the pixels accordingly. Each face was displayed at a size of 396 pixels wide × 460 pixels tall.

Procedure. All three experiments were run using jsPsych (de Leeuw, 2015). After receiving instructions, participants completed 360 XAB trials, split into six blocks. The middle six faces in Figure 2 were used as targets. Each of these six faces was a target on 10 trials per block, randomly distributed over the block. On each trial, participants saw the original (X) face in the center of the screen for 500 ms, followed by a blank screen for 500 ms. The target was then displayed on the screen along with the foil, playing the roles of the A and B stimuli. The target and foil faces were displayed at the same height on the screen as the original face, with one face to the left of the center and the other to the right. On each trial, the target face location was randomized to either the left or the right position. Text displayed below the faces reminded participants to use the *Q* key to select the left face, or the *P* key to select the right face as being identical to the original face. The target and foil remained onscreen until the participant indicated that either the left or right face was the target. Once participants made their choice there was then another blank screen for 500 ms, followed by the start of the next trial. In between blocks, participants were instructed to take a break for as long as they needed. At the start of each block after the first one, participants saw a reminder to complete each trial as quickly and accurately as possible. After the main XAB task was complete, participants completed 360 categorization trials, split into five blocks. On each trial, participants saw one of the six faces closest to the middle of the morph continuum in the center of the screen and were asked to indicate whether the face was White or Black. Participants responded using the *Q* and *P* keys, with the keys randomly assigned to Black or White between participants. Each face was shown an equal number of times in each block, and presentation order was randomized within blocks. At the end of the experiment, participants completed demographics questions and were thanked and debriefed.

Figure 1a shows the stimulus distribution used in the first two experiments. Each target face had two possible foils it could be paired with—one that was two faces closer to the Black end of the continuum, and one that was two faces closer to the White end of the continuum. Pilot testing using directly adjacent faces as target-foil pairs resulted in undesirably low accuracy. Each pairing appeared five times per block in Experiment 1. Contrast this distribution with Figure 1b, which shows the stimulus distribution used in the third experiment. The distribution in Figure 1b is representative of previous CP experiments that used one-dimensional continua, including each of those reanalyzed in Hanley and Roberson (2011). (See the original studies reported in that article for specific stimuli, e.g., Roberson & Davidoff, 2000, which used a blue–green color continuum, and Kikutani, Roberson, & Hanley, 2010, which used face continua.) The two distributions in Figures 1a and 1b each control for something that the other does not. In the distribution used in Experiments 1 and 2, each target face is equally likely to appear with a foil that is blacker or a foil that is whiter. In Experiment 3, this is not the case; the four most extreme faces in Experiment 3 only appear with foils that are closer to the opposite side of the continuum. In the distribution used in Experiment 3, every face that is a choice could potentially be a target. In Experiments 1 and 2, this is not the case; the four most extreme choice faces are only foils and never targets. All experiments using one-dimensional continua will either have stimuli that are never targets, as in Figure 1a, or targets that are never paired with more extreme foils, as in Figure 1b.

The two distributions also differ in the probability that the extreme face is the foil or the target. It can be seen from the Figures that in Experiments 1 and 2, the middle faces are equally likely to be targets or foils, whereas in Experiment 3, the middle faces are more likely to be foils than targets. Specifically, for Experiments 1 and 2, the probability of the extreme face from a pair being the foil is 0.66, whereas in Experiment 3 this probability is 0.33. This difference could be important if participants use a face's history of being a target or a foil as a cue in making their decision. The experiments presented here were approved by Indiana University's Institutional Review Board.

Results

For each face, Figure 3 plots the proportion of categorization trials that the face was categorized as Black. Although there is more uncertainty around the midpoint of the continuum, these results justify our intuition that the midpoint is also the category boundary.

In all three experiments, we analyzed the XAB data using Bayesian parameter estimation in JAGS (Plummer, 2003), with the

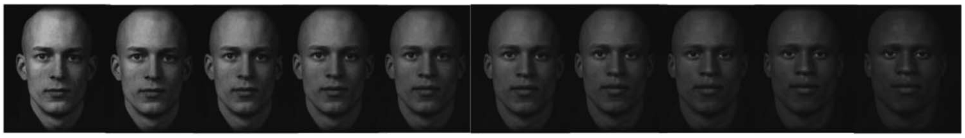

Figure 2. Face morph continuum used in all 3 experiments. Adapted from *Heads by Alex Kayser* (pp. 23, 71), by Alex Kayser, 1985, New York, NY: Abbeville Press. Copyright 1985 by Alex Kayser.

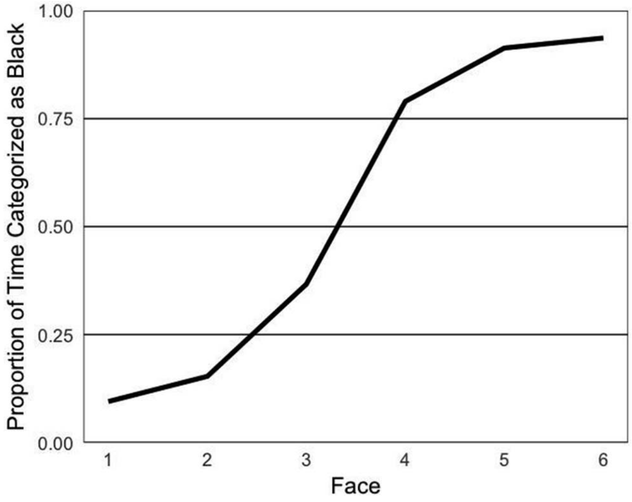

Figure 3. Proportion of categorization trials in which each face in Experiment 1 was categorized as Black.

help of code provided in Kruschke (2014). The model we estimated was a logistic regression, with the result of each trial (correct vs. incorrect) as the dependent variable. The model had a random intercept for each participant and deflections for extremeness (target extreme vs. foil extreme) and trial type (between-category vs. within-category). All three chains in JAGS were well-behaved, with an effective sample size greater than 10,000 for all statistics reported here, indicating that the search through parameter space was thorough and did not get stuck at local maxima. All estimates are reported in terms of sum-to-zero deflections from the mean, on the scale of log-odds.

There was a main effect of trial type, with higher accuracy on between-category pairs than within-category pairs (95% highest density interval [HDI] of difference: [0.06, 0.20]). This difference is consistent with CP. There also was a main effect of extremeness, with participants having higher accuracy when the foil, not the target, was more extreme (95% HDI of difference: [0.28, 0.42]). This is the opposite of the effect of extremeness found by Hanley and Roberson (2011), where accuracy was higher when the target was more extreme. There was also an interaction, such that the difference in accuracy when the target or foil was more extreme was larger for within-category pairs than between-category pairs (95% HDI of the double-difference, or within-category difference minus between-category difference: [0.06, 0.13]). See the leftmost graph of Figure 8 for a graph of these results.

For each face, Figure 4 shows participants' accuracy for trials when that face was the target and trials when that face was the foil. Face 1 is the whitest face and Face 10 is the blackest face. It can be seen from the graph that participants were more accurate on Black faces than White faces. Given that most of the participants were White, this is conceptually consistent with previous findings from Levin (2000), in which discrimination accuracy was higher for Black than White faces on a Black–White continuum. This supports the feature selection hypothesis, which predicts higher discrimination accuracy when looking at outgroup than ingroup members, which is the reverse of the pattern that is widely assumed to occur.

Figure 4 also suggests an alternative way of looking at the extremeness result described above. For the two central faces, 5 and 6, accuracy is higher when those faces are targets than when they are foils. For Faces 3 and 8 (the most extreme faces that were ever targets), accuracy was higher when those faces were foils than when they were targets. This can be viewed as a bias toward selecting the central faces as targets and against selecting the extreme faces. Figure 4b visualizes this as the difference in accuracy for each face when it is in the target and foil role, excluding faces that were never targets. The central faces have positive values, indicating that participants were biased toward saying those faces were targets. The extreme faces have negative values, indicating that participants were biased against saying those faces were targets.

Discussion

The results of Experiment 1 contradict the results shown by Hanley and Roberson (2011). There was a CP effect, with partic-

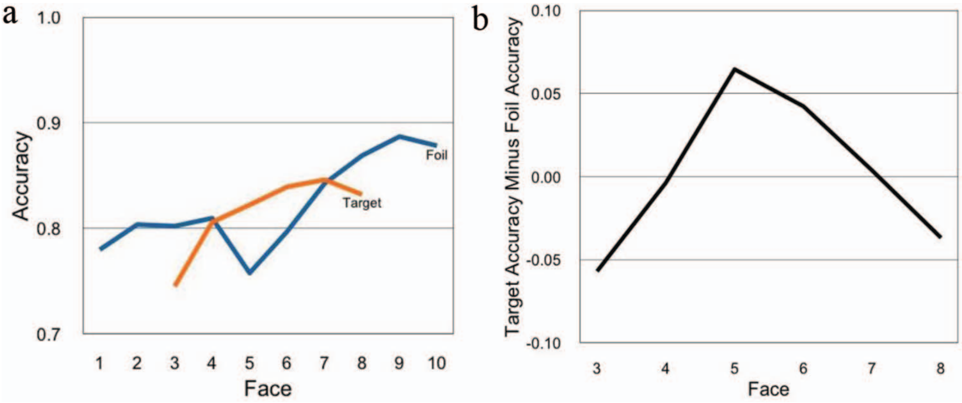

Figure 4. Panel a: Accuracy for each face on trials where that face was the target (orange [light gray] line) and when it was the foil (blue [dark gray] line) in Experiment 1. Panel b: For each face in Experiment 1 that was ever a target, difference in accuracy between trials where that face was the target and trials where that face was the foil. See the online article for the color version of this figure.

ipants more accurate on between-category trials than within-category trials. However, participants were more accurate when the foil was more extreme than the target, which is contrary to previous findings. One explanation for this reversal is that the stimulus distribution in this experiment is different from those used in previous experiments, and participants here chose the less extreme faces more often because such faces were, in fact, more likely to be correct. However, it is possible that the pattern of results was due to the particular stimuli used. Previous experiments that used faces as stimuli and showed an advantage for pairs with the target more extreme than the distractor (Hanley & Roberson, 2011; Hendrickson, Carvalho, & Goldstone, 2012) used morphs between individuals of the same race. If participants in previous face morph experiments conceptualized the faces they saw as being a continuum between individuals, although participants in this experiment saw the faces as being a continuum between races, there could be an explanation for the results that does not rely on the stimulus distribution but instead points to a difference in the way participants thought about the stimulus continuum. Experiment 2 was designed to test this by "introducing" participants to the two endpoint faces before the XAB task, giving them practice seeing and identifying the endpoints to lead them to see the morph faces as existing on a continuum between two individuals, rather than between two races.

The sizes of the effects in Experiment 1 were greater than we had anticipated. Because of this, we recruited fewer participants in Experiments 2 and 3. We also gave participants fewer trials to complete, due to the sizes of the effects and feedback from participants in Experiment 1 who said that the experiment was too long and that they found it difficult to focus near the end of it.

Experiment 2

Method

Participants. Thirty-eight workers on Amazon's Mechanical Turk completed the experiment for $1.00. The racial makeup of the sample was 74% White and 11% Black.

Materials. Experiment 2 used the same faces from Experiment 1. There was an introduction phase, where participants were introduced to two of the faces. These were the most extreme faces from the morph continuum, that is, the leftmost and rightmost faces in Figure 2.

Procedure. Before reading instructions for the XAB task, participants were "introduced" to the two most extreme faces as John and Eric. They saw the two faces on the screen one at a time, along with a name and a short story about each person. Identity (whether the White face or Black face was John and the other face was Eric) was randomized, as was order of presentation (White face first or Black face first). Participants were instructed to memorize each face and to look at them for 30 s each. Participants then completed an identification task in which they identified each face as John or Eric twice, with feedback. We hoped this active task, as opposed to passively viewing the faces, would encourage participants to pay attention to the faces. The next part of the experiment was the same as Experiment 1, except that there were only 300 XAB trials in five blocks. Participants completed 68 categorization trials in one block. To capture more information about participants' categorization of faces near the category boundary, the frequency of presentation was organized in a stair step: The middle two faces were shown 16 times each, the next two faces out from the middle were shown 12 times each, and the next two faces were shown six times each, with the presentation order randomized. Because of a coding error, 7 participants did not do the categorization task.

Results

The results of Experiment 2 were similar to those of Experiment 1. As in Experiment 1, Figure 5 shows that the midpoint of the stimulus continuum is also the category boundary. There was a main effect of trial type, with higher accuracy on between-category than within-category trials (95% HDI of difference: [0.09, 0.31]). Just as in Experiment 1, there was also a main effect of extremeness, with higher accuracy for trials in which the foil was more extreme than the target (95% HDI of difference: [0.19, 0.42]). Again, there was an interaction such that extremeness had a greater effect on within-category trials than between-category trials (95% HDI of double-difference: [0.07, 0.18]). See the middle graph of Figure 8.

Figures 6a and 6b show that the pattern for individual faces is also similar to the pattern in Experiment 1. Again, participants were biased toward selecting central faces as targets, and against selecting faces near the ends as targets. They were more accurate when central faces were targets than when they were foils, and more accurate when faces near the ends were foils than when they were targets. There was also higher accuracy overall for Black faces than White faces.

Discussion

Despite the manipulation in which the continuum endpoints were introduced as specific individuals at the start of the experiment, Experiment 2 produced the same pattern of results as Experiment 1. This suggests that construal of the continuum as being between racial categories rather than between identities does not

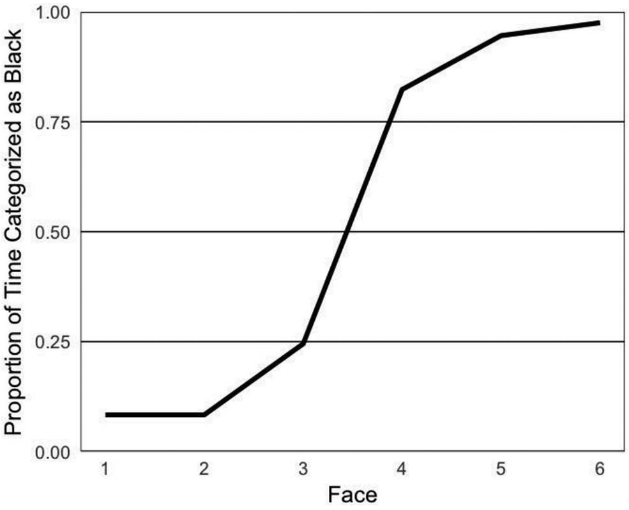

Figure 5. Proportion of categorization trials in which each face in Experiment 2 was categorized as Black.

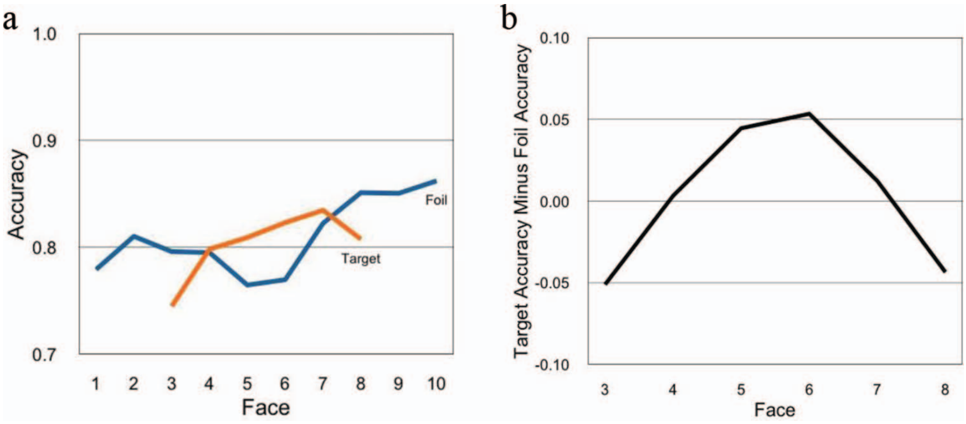

Figure 6. Panel a: Accuracy for each face on trials where that face was the target (orange [light gray] line) and when it was the foil (blue [dark gray] line) in Experiment 2. Panel b: For each face in Experiment 2 that was ever a target, difference in accuracy between trials where that face was the target and trials where that face was the foil. See the online article for the color version of this figure.

explain why participants in these experiments were more accurate when the foil was more extreme than the target, which is the opposite of what previous experiments have shown. Experiment 3 was designed to test the hypothesis that the reason for this pattern is the stimulus distribution. If the choice distribution is what determines whether accuracy is higher when foils or targets are more extreme, then an XAB task that uses these same faces but in a distribution like the one used in Hanley and Roberson's (2011) experiments should result in higher accuracy when targets, rather than foils, are more extreme.

Experiment 3

Method

Participants. Forty-two Indiana University students completed the experiment for partial course credit. The racial makeup of the sample was 67% White and 2% Black. This experiment was run in the lab rather than on Amazon's Mechanical Turk to help meet department quotas on the number of opportunities for undergraduate students to participate in psychology experiments.

Materials. The face continuum in Experiment 3 was the same as in Experiments 1 and 2. The only difference was that the distribution of targets and foils was such that only the middle six faces were used. The four most extreme faces in the first two experiments, which had only been foils but never targets in those experiments, were never presented in Experiment 3.

Procedure. The procedure was the same as in Experiment 1, except for the number of trials and the stimulus distribution. Here there were 300 trials over five blocks. The stimulus distribution (see Figure 1b) was representative of previous XAB experiments that have shown CP effects, including those in Hanley and Roberson (2011). Each target was presented 10 times per block, and there were no faces that appeared as foils that did not also appear as targets. The categorization task was the same as in Experiment 2.

Results

The results of this experiment were consistent with the hypothesis that the direction of the difference in accuracy between trials in which the target was more extreme than the foil and trials in which the foil was more extreme than the target is determined by the stimulus distribution. Again, Figure 7 shows that participants saw the midpoint of the continuum as the category boundary between Black and White. There was a CP effect: participants were more accurate on between-category trials than within-category trials (95% HDI of difference: [0.14, 0.32]). Replicating the pattern from Hanley and Roberson (2011) and Hendrickson et al. (2012), accuracy was higher when the target was more extreme than the foil (95% HDI of difference: [−0.53, −0.36]). There was again an interaction such that extremeness had a greater effect on within-category than between-category trials (95% HDI of double-difference: [−0.15, −0.06]). Figure 8 shows accuracy on within-

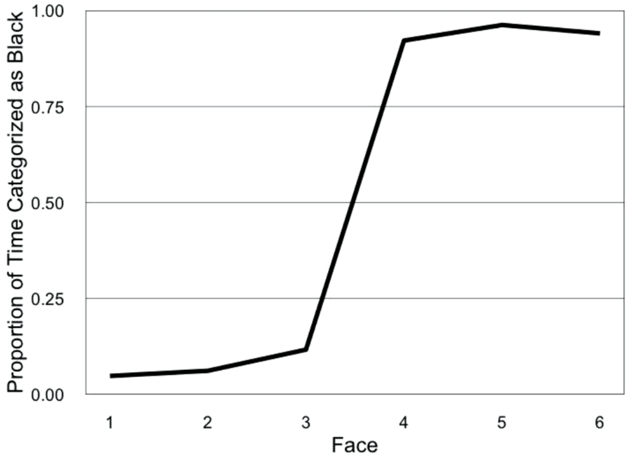

Figure 7. Proportion of categorization trials in which each face in Experiment 3 was categorized as Black.

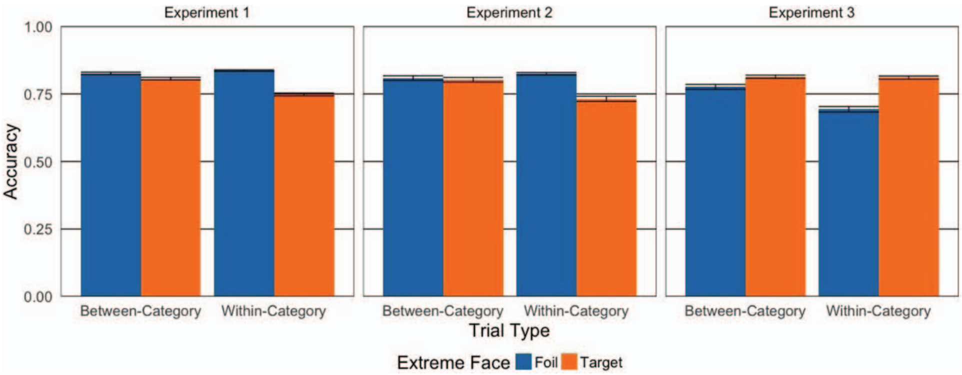

Figure 8. Results from Experiments 1, 2, and 3. Error bars represent the bounds on the 95% HDI. See the online article for the color version of this figure.

and between-category trials when either the target or foil is more extreme in all three experiments.

Figures 9a and 9b show the accuracy for individual faces. As with the accuracy by trial type, these graphs show that the pattern of results in Experiment 3 is essentially the mirror image of the results in the first two experiments. Participants were biased toward calling the more extreme faces targets and against calling the more central faces targets.

Discussion

These three experiments show that accuracy differences based on extremeness (target more extreme vs. foil more extreme) in an XAB task can be manipulated by changing the distribution of stimuli in the experiment. Experiments 1 and 2 used a stimulus distribution in which each target face was equally likely to be presented with a foil that was closer to either the Black or White end of the stimulus continuum. Experiment 3 used the same target faces, but was set up in such a way that the most extreme target faces were only ever presented with a foil that was closer to the opposite end of the continuum (e.g., the blackest target was only ever presented with a whiter foil). All three experiments showed CP effects, with higher accuracy on between-category trials than on within-category trials. In the first two experiments, participants were more accurate on trials where the foil was more extreme than the target, compared with trials where the target was more extreme than the foil. This is the opposite of what has been found in previous experiments (Hanley & Roberson, 2011; Hendrickson et al., 2012). In Experiment 3, which used a choice distribution similar to those from previous experiments, this pattern was reversed and was consistent with earlier studies. That is, participants were more accurate on trials where the target was more extreme than the foil.

One potential explanation for these results is that participants in the experiments learned which choices tend to be correct. In Experiments 1 and 2, there was a 0.33 probability that the more extreme face from a pair was the target face, and indeed the most extreme faces participants saw were never targets. Participants could have learned that more extreme faces were more likely to be foils than targets, and used this to inform their decisions by biasing their choices toward saying the less extreme face of a pair was the target. In Experiment 3, there was a 0.66 probability that the more extreme face from a pair was the target. Participants in Experiment 3 could have learned that more extreme faces were more likely to

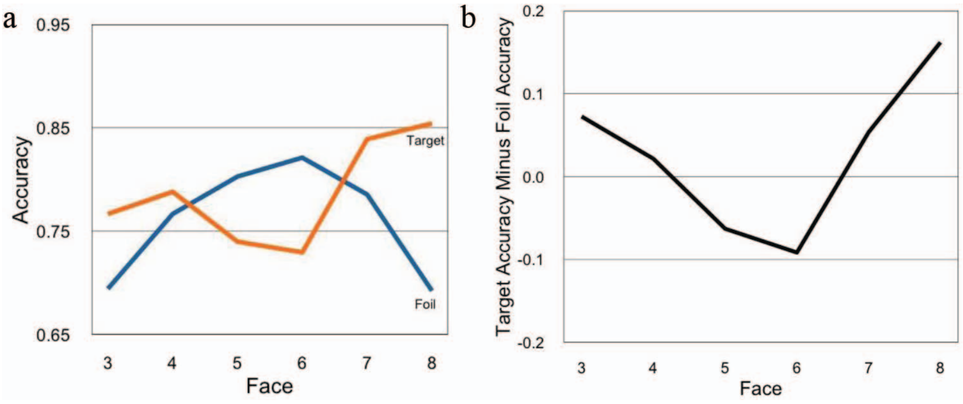

Figure 9. Panel a: Accuracy for each face on trials where that face was the target (orange [light gray] line) and when it was the foil (blue [dark gray] line) in Experiment 3. Panel b: For each face in Experiment 3 that was ever a target, difference in accuracy between trials where that face was the target versus foil. See the online article for the color version of this figure.

be targets than foils, leading to a bias toward selecting the more extreme face. This explanation assumes that participants were able to learn statistical regularities present in the experiments, possibly through a kind of reinforcement learning mechanism. Participants did not receive feedback on their choices, so such a mechanism may require internal feedback.

In the next section, we present simulation results from two models. The first model performs the XAB task using perceptual evidence, and the second uses category labels. Both models also use information about what types of faces (extreme vs. not extreme) have been correct on previous trials, and which faces come from the same category as the target face. The simulations show that both PE and CL models can produce the pattern of results shown in our experiments, given that they also use information about the stimulus distribution.

Modeling

Model 1: Category Labels

According to Hanley and Roberson's (2011) category label explanation of CP effects, participants in an XAB task do not have an accurate perceptual representation of the target stimulus at test. Instead, they rely on the category labels of the original stimulus and the two options to make their decision. If the participant knows that the original stimulus was from Category A, and one of the options at test is from Category A and the other is from Category B, then the decision is an easy one. If people consistently categorized each stimulus as an A or B, then we would expect uniformly lower accuracy on within-category trials compared with between-category trials. On every within-category trial, the stimuli would all have the same category label, so participants would be forced to rely on their inaccurate perceptual representation of the original stimulus. On between-category trials, on the other hand, the target would always have a category label that matched the original stimulus and the foil would have a label that did not match.

Categorization, however, is not consistent for all stimuli. After the XAB task in each experiment, participants completed a categorization task in which they categorized each face as either Black or White multiple times. Figure 7 shows the proportion of the time that participants in Experiment 3 categorized each target face as Black in the categorization task. As the curve shows, central faces (those near the category boundary) are categorized less consistently than the extreme faces. That is, participants' uncertainty about stimulus categorization is greatest near the category boundary. This, according to Hanley and Roberson (2011), is what produces the difference in accuracy for trials in which the target face was more extreme versus trials in which the foil was more extreme.

In an XAB task, participants will make the correct choice some percentage of the time just by using the perceptual evidence available. They will be more accurate, though, when they assign the original face (X) and the target (A) to the same category but the foil (B) to a different category, and less accurate when they assign the original face and the foil to the same category but the target differently. As a concrete example, consider the choices between Face 1 and Face 3, where Face 1 is close to the White end of the continuum, and Face 3 is just on the White side of the category boundary. Call the event that Face i is assigned to the White category W_i and the event that Face i is assigned to the Black category is B_i. Suppose that $p(W_1) = 0.90$, $p(B_1) = 0.10$, $p(W_3) = 0.55$ and $p(B_3) = 0.45$. When Face 1 is the target, the probability that the original and target faces are assigned to the same category and that this category is different from the one the foil is assigned to is

$$(p(W_1) \bullet p(W_1) \bullet p(B_3)) + (p(B_1) \bullet p(B_1) \bullet p(W_3)) = 0.37, \tag{1}$$

whereas the probability that the original and foil faces are assigned to the same category and that this is different from the one the target is assigned to is

$$(p(W_1) \bullet p(B_1) \bullet p(W_3)) + (p(B_1) \bullet p(W_1) \bullet p(B_3)) = 0.09. \tag{2}$$

(The rest of the time, the target and foil are assigned to the same category.) When Face 3 is the target, the probability that the original and target faces are assigned to the same category and that this category is different from the one the foil is assigned to is

$$(p(W_3) \bullet p(W_3) \bullet p(B_1)) + (p(B_3) \bullet p(B_3) \bullet p(W_1)) = 0.2125, \tag{3}$$

whereas the probability that the original and foil faces are assigned to the same category and that this is different from the one the target is assigned to is

$$(p(W_3) \bullet p(B_3) \bullet p(W_1)) + (p(B_3) \bullet p(W_3) \bullet p(B_1)) = 0.2475. \tag{4}$$

Thus, even though the choice between Face 1 and Face 3 is a within-category decision, the probability that category labels will help or hinder the participant in making the correct choice depends on which face is the target. When Face 1 (the more extreme face) is the target, there is a 0.37 probability that the faces will be assigned to categories in a way that points to the correct choice, and a 0.09 probability that the assigned categories will point to the incorrect choice. When Face 3 (the less extreme face) is the target, there is a 0.2125 probability that the categories help make the right choice, and a 0.2475 probability that they point toward the wrong choice.

But what about trials in which the assigned category labels are inconclusive? Consider the case where the original face is categorized as White, and both the target and the foil are categorized as Black. A participant could choose between the target and foil randomly. However, if the choice were made in the context of the stimulus distribution in Experiment 3 where the more extreme face is correct 66% of the time, a more adaptive strategy would be to guess that the more extreme face was correct in situations where the category labels are inconclusive. Similarly, if less extreme faces tend to be correct, one should choose the less extreme face when labels are inconclusive.

We simulated this model performing the XAB task from Experiment 3 by having it categorize the X (original), A (target) and B (foil) faces on each trial as either White or Black with the same probability as participants in Experiment 3. Then, the model chose the correct face with the following probabilities: probability 1 if X and A were categorized differently than B, probability 0 if X and

B were categorized differently than A, probability 1 if the labels were inconclusive and the target was more extreme, and probability 0 if the labels were inconclusive and the foil was more extreme. Each of the 12 pairs of choices was made 10,000 times.

We also simulated the model performing Experiment 1. The simulation was the set up the same as the previous simulation, except that the model was correct with probability 0 if the labels were inconclusive and the target was more extreme, and with probability 1 if the labels were inconclusive and the foil was more extreme.

Figures 10 and 11 show that this model reproduces the pattern of results from Experiment 3 and Experiment 1, respectively. The model shows the CP effect, and it is biased toward the more extreme face when extreme faces are more likely to be correct and toward the less extreme face when extreme faces are less likely to be correct. Thus, Hanley and Roberson's (2011) CL explanation can account for the results shown in their article, and when information about the stimulus distribution is incorporated in the model it can also account for the pattern shown in our Experiments 1 and 2.

Model 2: Perceptual Effects

The previous section showed that a CL model can account for the results in our experiments. This section will show that a PE model can also reproduce the correct patterns. In this model, each of the two faces on each trial has some weight of evidence in favor of that face being the correct choice. This weight is a combination of perceptual evidence and evidence accumulated over the course of the experiment about how often extreme or less extreme faces are the targets. On each trial, the weight for the target face, w_t, is drawn from a beta(a_t, b_t) distribution, and the weight for the foil, w_f, is drawn from a beta(a_f, b_f) distribution. The beta distribution is a distribution over the interval [0, 1], and can be interpreted as degree of belief that a face has a particular probability of being the correct choice. The a and b parameters control the shape of the distribution: when a > b, more weight will be on the right side of the distribution; when b > a, more weight will be on the left side of the distribution; and the greater a and b are, the more concentrated the distribution will be about its mean. For the target face, the equations for a_t and b_t are

$$a_t = (p(\gamma) \bullet trials) + (\alpha \bullet trials \bullet g(\beta)), \quad (5)$$

$$b_t = (1 - p(\gamma)) \bullet trials, \quad (6)$$

where $p(\gamma) = \gamma$ if the target is more extreme than the foil, or $1 - \gamma$ if the target is less extreme than the foil; γ is the probability that the more extreme face is the target for the experiment being simulated; *trials* is the number of trials that have occurred; α is a parameter representing the perceptual evidence in favor of the target; $g(\beta)$ is β if the trial is a between-categories trial (i.e., A and B of the XAB test come from different categories) and 1 if it is a within-categories trial; and β is a parameter representing the additional perceptual advantage the target gains on between-category trials. For the foil face, the equations for a_f and b_f are

$$a_f = (1 - p(\gamma)) \bullet trials, \quad (7)$$

$$b_f = p(\gamma) \bullet trials. \quad (8)$$

Once the weights are drawn from these distributions, the softmax function is applied to the two weights and the model chooses the correct face with the resulting probability. For Experiments 1 and 3, we simulated each choice 100,000 times, with the number of trials fixed at 360 to represent the model's state at the end of the experiment. We used the optim function in R to fit the α and β parameters for each experiment separately (the γ parameter is fixed by the probability of the more extreme face being the target in each experiment). Figures 12 and 13 show that this model manages to produce the correct pattern of results for both experiments. It shows a CP effect for both. For the simulation Experiment 3 the model is more accurate when the target is more extreme than the foil. For the simulation of Experiment 1 the model is more accurate when the foil is more extreme than the target. For Experiment 3 the best fitting parameters were $\alpha = .94$ and $\beta = 20$. The best fitting parameters for Experiment 1 were $\alpha = .495$ and $\beta = 40.1$. That β was greater than 1 implies that categories affected perception.

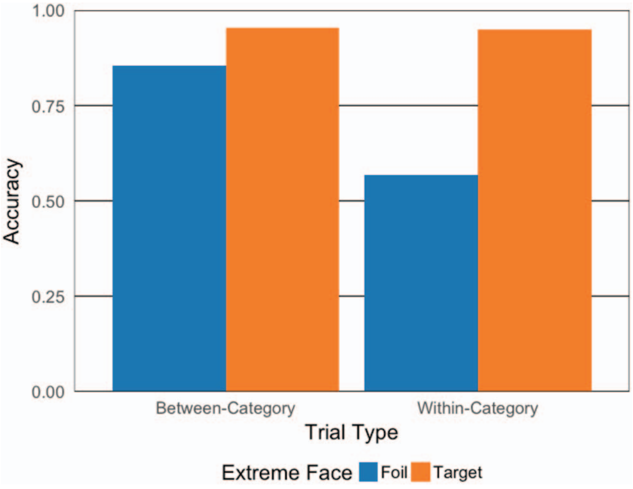

Figure 10. Results from category label model simulation of Experiment 3. See the online article for the color version of this figure.

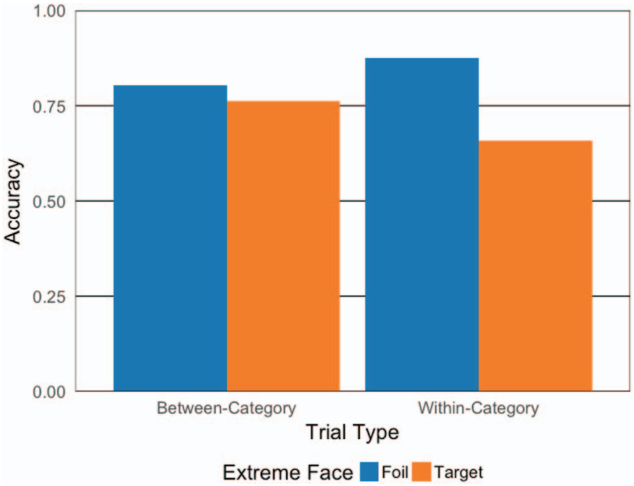

Figure 11. Results from category label model simulation of Experiment 1. See the online article for the color version of this figure.

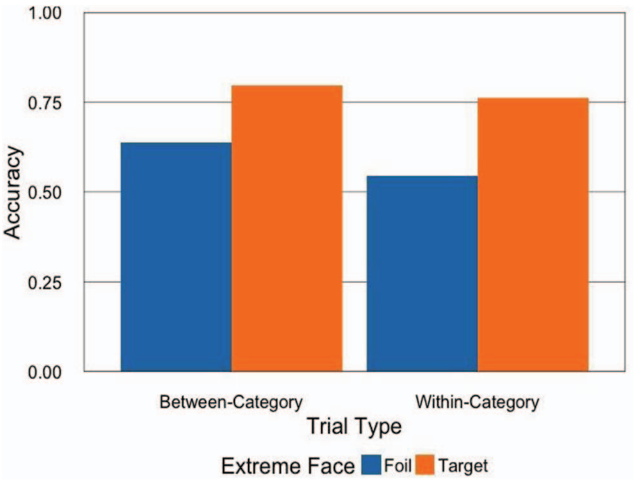

Figure 12. Results from perceptual effect model simulation of Experiment 3. See the online article for the color version of this figure.

Both the CL and PE models can account for the CP effect, as well as the influence of the distribution of foils seen in the current studies regarding the extremity bias. These models use both information about what kinds of stimuli (extreme vs. nonextreme) tend to be targets and category information to make decisions in the XAB task. Thus, although the results here cannot settle the question of whether CL or PE models are the better explanation for CP effects, they do suggest that learning about the stimulus distribution is necessary to explain the full pattern of results found in studies of CP using the XAB task.

General Discussion

Perceptual effect models of CP explain the phenomenon as resulting from the influence of categories on perceptual representations (e.g., Goldstone et al., 2000). Hanley and Roberson (2011) argued for a category label explanation instead, citing as evidence an asymmetry in results from CP experiments, wherein accuracy was higher on trials in which the target face is more extreme than the foil. The three experiments presented here provide evidence that this asymmetry can go in the opposite direction, depending on the distribution of targets and foils used. In Experiments 1 and 2, in which the distribution was such that the extreme face from any given pair had a 0.33 probability of being the target face, participants were more accurate when the foil was more extreme than the target. In Experiment 3, in which all faces appeared as both targets and foils (similar to the distributions used in Hanley and Roberson's article) and the more extreme face had a 0.66 probability of being the target, participants were more accurate when the target was more extreme than the foil.

Simulation results showed that, when augmented with information about probability of extreme versus nonextreme stimuli being the target, both CL and PE models can account for the results of all three experiments. Thus, although the results we present show that the effect Hanley and Roberson (2011) sought to explain was caused by a confound in the experiments they reanalyzed, they do not provide evidence against CL models in favor of a PE model. Instead, these results suggest there exists some mechanism by which participants are able to learn about which stimuli along the continua are more likely to be targets and use this information in the experiment, leading to a bias either toward or away from extreme faces. This is remarkable because we gave participants no feedback about their choices and no indication that extremeness of a face would be correlated with that face's probability of being the target. Although we do not take any strong positions on how participants learned in the present studies, we do assume in the models that participants learned bias for or against extreme faces in general, rather than learning about the probabilities that specific faces are targets. It is possible that evidence is accumulated separately for each face, but the current studies suggest this is not the case. In Experiments 1 and 2, the most extreme faces that were ever targets were more likely to be targets than foils. If evidence was being accumulated for each face separately, we should expect participants to be more accurate when these faces are targets than when they are foils. However, this was not the case—participants were more accurate when these faces were foils than targets, suggesting that the bias learned in this task is a general bias for or against extreme faces, rather than a specific bias for individual faces. Future work should attempt to answer the question of how best to model the mechanism that allows people to learn this bias without being given feedback.

Although these experiments do not provide evidence in one direction or the other in the debate of CL models versus PE models, they have implications for future CP research. In a CP experiment it is important to control the magnitude of the physical differences between stimuli, but our results show that experimenters should also be careful about how often each stimulus or category is shown and how often it is the correct choice or the incorrect choice. This applies no matter what the stimulus space looks like, but bisected continua like the ones focused on in this article are particularly problematic. Some arrangements of the stimulus space, like the circular one used by Roberson, Damjanovic, and Pilling (2007), avoid the problem by not having extreme stimuli. In addition, although categories that sit on either side of a one-dimensional continuum are in some ways convenient to use in research, perceiving and thinking about them surely represents very little of how categories affect the way we interact

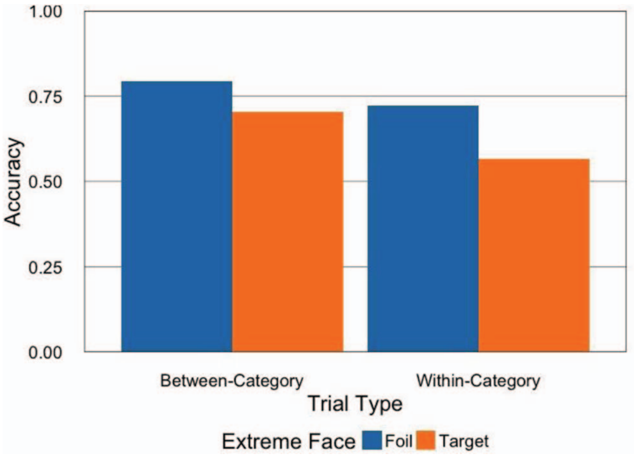

Figure 13. Results from perceptual effect model simulation of Experiment 1. See the online article for the color version of this figure.

with our environment. Future researchers should continue to look for stimuli and experiment designs that can teach us about how CP works across the broad range of categories and challenges humans encounter.

References

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47,* 1–12. http://dx.doi.org/10.3758/s13428-014-0458-y

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 69–78. http://dx.doi.org/10.1002/wcs.26

Goldstone, R. L., Steyvers, M., Spencer-Smith, J., & Kersten, A. (2000). Interactions between perceptual and conceptual learning. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 191–228). Mahwah, NJ: Lawrence Erlbaum.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition, 31,* 169–180. http://dx.doi.org/10.3758/BF03194377

Hanley, J. R., & Roberson, D. (2011). Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review, 18,* 355–363. http://dx.doi.org/10.3758/s13423-010-0043-z

Harnad, S., Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modelling: Steps toward principled integration* (pp. 191–206). Ames, IA: Iowa State University.

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 Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 1662–1667). Austin, TX: Cognitive Science Society.

Kayser, A. (1985). *Heads by Alex Kayser*. New York, NY: Abbeville Press.

Kikutani, M., Roberson, D., & Hanley, J. R. (2010). Categorical perception for unfamiliar faces: The effect of covert and overt face learning. *Psychological Science, 21,* 865–872. http://dx.doi.org/10.1177/0956797610371964

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General, 129,* 559–574. http://dx.doi.org/10.1037/0096-3445.129.4.559

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 732–753. http://dx.doi.org/10.1037/0278-7393.24.3.732

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Attention, Perception & Psychophysics, 15,* 285–290. http://dx.doi.org/10.3758/BF03213946

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, p. 125). Wien, Austria: Technische Universitat Wien.

R Core Team. (2015). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org

Roberson, D., Damjanovic, L., & Pilling, M. (2007). Categorical perception of facial expressions: Evidence for a "category adjustment" model. *Memory & Cognition, 35,* 1814–1829. http://dx.doi.org/10.3758/BF03193512

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28,* 977–986. http://dx.doi.org/10.3758/BF03209345

Steyvers, M. (1999). Morphing techniques for manipulating face images. *Behavior Research Methods, Instruments, & Computers, 31,* 359–369. http://dx.doi.org/10.3758/BF03207733

Received May 29, 2017
Revision received March 5, 2018
Accepted March 8, 2018 ■