Does Category Learning Alter Perception or Memory of Magnitudes?

Jessica M. Choplin
Vanderbilt University

Philip J. Kellman
University of California, Los Angeles

Address for correspondence:
Jessica M. Choplin
Vanderbilt University Psychology Department
111 21st Ave. S.
Nashville, TN 37203
Phone: (615) 322-5169
Fax: (615) 343-8449
jessica.choplin@vanderbilt.edu

Abstract

In categorical perception effects (CP), observers detect magnitude differences that cross category boundaries better than magnitude differences that do not. Several researchers have argued that these CP effects reflect changes in sensitivity that create category coherence during category learning (e.g., Livingston, Andrews, & Harnad, 1998). Others have argued that CP effects reflect how language acquisition affects perception (e.g., Davidoff, 2001). In three experiments, we examined these arguments by investigating the cognitive mechanisms responsible for CP. We found evidence that CP is an effect not of perception (information pick up) *per se*, but rather of memory (information storage) and we also explored the role of memory in these effects. These results suggest that CP does not reflect category coherence or the effects of language on perception.

Does Category Learning Alter Perception or Memory of Magnitudes?

Across an impressive variety of perceptual categories including phonemes, colors, faces, and novel stimuli, the ability to categorize items has been associated with categorical perception (CP) effects, in which observers are better able to detect a given magnitude difference when it crosses a category boundary than when it does not[1]. For example, observers are typically better able to detect a wavelength difference between a shade of green and a shade of blue than the same wavelength difference between two shades of green or two shades of blue (Bornstein & Korda, 1984). In this paper, we investigate the cognitive mechanisms responsible for these effects. In particular, we distinguish between mechanisms that involve the perceptual pick up of magnitude information from mechanisms that involve the storage of (memory for) information and determine which is most likely responsible for CP effects.

We were motivated to investigate the cognitive mechanisms responsible for CP by competing accounts of this phenomenon. One proposal has been that category learning produces changes in perceptual sensitivity, which thereby create category coherence (Goldstone, 1994; Harnad, 1987; Livingston, Andrews, & Harnad, 1998). Another proposal has been that CP provides an example of Whorf's (1956) hypothesis, that the language an observer learns to speaks (specifically the category terms an observer learns to use) affects how she or he perceives stimuli (Davidoff, 2001). These claims have in common the idea that perceptual sensitivity changes as a result of category learning. Both would be unlikely, however, if mechanisms involving the storage (rather than pick-up) of information were responsible for CP.

CP tasks usually come in two phases. In the first phase, the boundary between two categories on the relevant dimension is established. The nature of this phase depends upon whether or not observers already know the categories. If observers already know the categories – such as color categories (e.g., blue vs. green) or phoneme categories (e.g., /da/ vs. /ta/), then this phase only requires a determination of where the category boundary lies. Experimenters typically present stimuli to observers and have them categorize the stimuli. The point on the dimension at which observers become more likely to place stimuli in one category rather than the other represents the category boundary. If observers do not already know the categories – as in the case of experimenter-defined categories, then observers must be trained to categorize differently items on the two sides of the boundary. In the second phase, observers' sensitivity to within- and between-category magnitude differences is measured. Note that studies of this kind are ordinarily conducted to assess how category information affects sensitivity. When there are other influences on sensitivity (i.e., Weber's law), such studies typically control for these extraneous influences by scaling presented differences to keep non-categorical influences constant. Sensitivity in these studies is best measured using classical psychophysical measures such as just noticeable differences (JNDs) or signal detection measures such as d'. We will return to discuss why it is vital to use measures of this type in the discussion section of Experiment 1.

Typical results for these CP tasks are shown in Figure 1. Along the x-axis is plotted the magnitude of the presented stimuli typically scaled to control for non-categorical influences on sensitivity such as Weber's law. In Experiments 1 through 3 below we define categories by size. Size, therefore, would be plotted along the x-axis. Analogously, if investigating CP effects for colors, wavelengths in nanometers would most likely be plotted along the x-axis. If investigating phonemic categories, then the relevant phonemic dimension, e.g., voice onset time, would be plotted along the x-axis.

Along the y-axis is plotted sensitivity, indicated here as d'. CP effects are observed when sensitivity is highest across the category boundary and lower within each of the categories, as is the case in Figure 1.

Does CP reflect sensitivity differences that produce category coherence as proposed by Livingston, Andrews, & Harnad (1998) and others (Goldstone, 1994; Harnad, 1987)? Does it reflect how language affects observers' perceptual abilities as proposed by Davidoff (2001)? The answers will depend upon the cognitive mechanisms that create CP effects. We discuss four proposed models of CP effects. Two of these models assume that CP is best explained by mechanisms that involve the perceptual pick up of magnitude information and, in general, suggest affirmative answers about true changes in sensitivity. The other two models assume that CP is best explained by mechanisms involving storage of magnitude information; these models suggest negative answers about true changes in sensitivity.

Perceptual Pick-Up Models

The defining assumption of the perceptual pick-up models is that category membership affects magnitude perception. These models differ on the mechanisms that they assume are responsible for altered perception. Nevertheless, the common assumption that category membership affects the perception, rather than simply the assessment or judgment, of magnitudes implies a set of common predictions. For current purposes, the most central of these is that the categories an observer has ought to affect the way stimuli appear. Categorical perception effects, therefore, ought to be observable for simultaneously presented items. We discuss two perceptual pick-up models: the Feature detector model (Abbs & Sussman, 1971) and the Selective Attention Model (e.g., Goldstone, 1994).

Feature Detector Model. Abbs and Sussman (1971) proposed a model of speech perception in which human speech sounds activate phonemic feature detectors. These feature detectors, in turn, could produce CP effects by providing inputs to mechanisms that differentiate magnitudes. In particular, if one stimulus magnitude activates Feature Detector X and another activates Feature Detector Y, then the two magnitudes will be less likely to become confused. For stimuli on different sides of a category boundary, their different feature detector activations would enhance the perceived difference between them. The additional information provided by feature detectors would produce expansion effects such as that shown in Figure 2 in which sensitivity across the category boundary is greater than baseline sensitivity. According to the Feature Detector Model interpretation, the straight line represents observers' abilities to detect magnitude differences if there were no feature detectors; detection ability would be a simple function of stimulus magnitude differences. The curved lines represent observers' predicted abilities to detect magnitude differences given feature detectors. The feature detectors add information that aids observers' attempts to detect magnitude differences beyond that provided by stimulus magnitude differences alone.

Although Abbs and Sussman (1971) proposed this as a theory of speech perception and hypothesized that the phonemic feature detectors were innate, the theory can easily be generalized to provide an account of CP in various stimulus domains, including those in which categories are learned. Abbs and Sussman themselves pointed out that visual stimuli activate visual-feature detectors analogous to the phonemic feature detectors they proposed. It is also possible that feature detectors could develop through perceptual learning. In this way, the feature detector model could encompass recent findings that perceptual learning produces CP effects (Goldstone, 1994) and that some categories that produce CP effects are culture specific (Davidoff, 2001).

Selective Attention Model. An alternative perceptual pick-up model was proposed by Goldstone (1994). In his model of CP, observers learn to selectively focus their attention on local regions of a dimension. This selective attention could be focused on select regions of a dimension to pick up either more finely grained magnitude information (producing expansion effects; Figure 2) or less finely grained magnitude information (producing compression effects; Figure 3). Under this theory, the straight lines in these figures represent observers' abilities to detect magnitude differences before category learning and the curved lines represent observers' predicted abilities to detect magnitude differences after category learning. In the expansion effect (Figure 2), category learning produces a tuning of attention to pick up more finely grained magnitude information in the region of the category boundary. In the compression effect (Figure 3), category learning produces a tuning of attention to ignore finely grained magnitude information in regions that fall entirely within a category. Goldstone (1994) reported evidence for expansion effects. Using a different paradigm (similarity ratings), Livingston, Andrews, & Harnad (1998) reported evidence for compression effects.

Memory Storage Models

The defining assumption of memory storage models is that category membership does not affect magnitude perception *per se*; rather category membership affects the judgment or assessment of magnitudes. These models assume that perceptual sensitivity and noise are the same regardless of whether an observer is categorizing stimuli or attempting to detect differences. The observer's assessment of category membership adds no new information to aid her or his ability to detect magnitude differences beyond that already provided by perception. These models predict that no CP effects will be observed for tasks based on perceptual magnitude information available in the stimulus. Rather, CP effects arise from processes that compensate for the declining accuracy of the observer's memory for perceived magnitudes after some time has elapsed. As memory for perceived magnitudes degrades, an observer can compensate in certain tasks by relying on category information.

We discuss two storage models: the Category Label Storage Model (Massaro, 1976) and a class of models suggested by work in magnitude estimation, the Biased Value Storage Models (i.e., Huttenlocher, Hedges & Vevea, 2000; Choplin & Hummel, 2002).

Category Label Storage Model. Massaro (1976) noted that the ability to label stimuli would help an observer detect magnitude differences in certain kinds of tasks. Specifically, observers in a discrimination task will try to remember both the presented magnitude values of stimuli (i.e., phonemes) and their labels. Because memory is imperfect, however, observers forget the exact presented magnitude values. When this happens, observers rely upon stored labels to produce responses. These labels, however, will only aid an observer's ability to detect differences across the category boundary. If the observer has forgotten the exact magnitude of one or both stimuli, but knows that one item belonged in Category X and the second item belonged in Category Y, she or he will not think that the two magnitudes are identical. Within a category, however, category labels will not be useful. After the observer has forgotten the exact magnitude of the stimuli, she or he will have no recourse in the category labels.

This model is very similar to the Feature detector model. The primary difference, however, is that under the Feature detector model information provided by feature detectors is immediately available to aid observers' abilities to detect magnitude differences. By contrast, under the Category Label Storage Model the category labels only aid an observer's ability to detect magnitude differences *after* some time has elapsed and memories of presented magnitude values have faded. In particular, the Category Label Storage Model predicts a pattern of results resembling either expansion (Figure 2) or compression (Figure 3) effects depending upon what is chosen as the baseline. If the observer's ability to detect magnitude differences after memories of presented magnitude values have faded is chosen as the baseline, then the pattern of results would constitute an expansion effect. Category information would preserve the observer's ability to detect differences across the category boundary relative to the lost information within categories. By contrast, if the observer's ability to detect magnitude differences before memories of presented magnitude values have faded were chosen as the baseline, then the pattern of results would resemble a compression effect. Category information would be unable to preserve the observer's ability to detect differences within categories.

Goldstone (1994) argued against this model on the grounds that category learning often improves observers' abilities to detect differences in regions adjacent to the category boundary not just at the category boundary itself. If categorization were deterministic (that is, everything on one side of the boundary was placed in one category and everything on the other side was placed in the other), then the category Label Storage Model would only predict improvement for differences that crossed the category boundary. There should never be improvement in adjacent regions. However, Goldstone, Lippa, and Shiffrin (2001) pointed out that improved discrimination in regions adjacent to the category boundary would not eliminate the Category Label Storage Model as a viable model of CP if categorization were stochastic. That is, if observers sometimes miscategorize stimuli, they may at times associate different category labels with items that lie entirely within one category. The Category Label Storage Model could then account for this improved discrimination within categories.

To strengthen arguments against the Category Label Storage Model, Goldstone, Lippa, and Shiffrin (2001) presented evidence that category learning produces changes in the representations of stimuli that cannot be explained by the Category Label Storage Model. To make this point they performed an experiment in which their observers rated the similarity of Faces A, B, C, and D to a neutral Face E. Observers were then trained to categorize Faces A and B into one category and Faces C and D into another. Finally, observers once again rated the similarity of Faces A, B, C, and D to the neutral Face E. After category training, ratings of faces in the same category for similarity to the neutral face were closer together than they had been previously. Goldstone et al. argued that the Category Label Storage Model could not explain this effect on the grounds that a strategy of utilizing category labels would not be appropriate for situations in which one of the items had not been categorized.

In our view, however, this result still does not eliminate the Category Label Storage Model as a viable model of CP. In the first place, the tasks are very different. Just as performance on absolute judgment tasks differs from performance on discrimination tasks (see Shiffrin & Nosofsky, 1994), performance on similarity judgments may also be very different from performance on CP discrimination tasks. They could very well involve different processes and representations. Second, while some judgmental strategies that utilize category labels would not be applicable to situations in which one of the items had not been categorized, one can easily imagine strategies in which category labels would still have an effect. For example, utilizing the information provided by category labels could produce a pattern of similarity judgments like that observed by Goldstone et al. (2001), if observers were to fallaciously treat the relation *similar(X Y)* as a transitive relation. That is, if A is similar to B because they share a category label, and B is similar (or dissimilar) to the neutral face E, then – by fallacious transitive reasoning – A would also be similar (or dissimilar) to E. Observers may be particularly likely to use a strategy such as this in cases in which within-category similarity ratings are mixed with categorized-to-neutral similarity judgments as in Goldstone, et al.'s experiment.

Biased Value Storage Models. An alternative class of storage models is suggested by work on magnitude estimation. As in the Category Label Storage Model, this model assumes that an observer's representation of a particular stimulus' perceived magnitude fades with time. As observers forget specific magnitude values, conceptual information such as category membership (Huttenlocher, Hedges & Vevea, 2000) or magnitude comparisons (i.e., "larger," or "smaller;" Choplin & Hummel, 2002) produce biases in recall. This reliance on conceptual information produces systematic memory biases about magnitudes. If observers judge whether two magnitudes are different based upon these biased magnitude values rather than presented magnitude values, then their ability to detect differences might be proportional to the differences in biased values, rather than proportional to the differences in presented values.

Based on prior research, this model predicts that differences in memory will be exaggerated across category boundaries. Huttenlocher and her colleagues have found biases in recall of magnitude values toward the central tendency of categories (See Figure 4). For example, Huttenlocher, Hedges & Vevea (2000) presented pictures of fish of varying sizes to their observers. The fish were then taken away and the observer reproduced their sizes from memory, using a method of adjustment. The reproduced sizes were biased closer to the central tendency of the category than the presented sizes. That is, the smallest sizes were overestimated and the largest sizes were underestimated; both were recalled as being closer than they actually were to the central tendency of the category. The resulting distribution of recalled sizes had significantly higher kurtosis than the presented distribution of sizes. When two categories are next to each other (as in Figure 4, and in CP tasks), these biases in recall would increase the recalled difference between items crossing a category boundary. The difference between the categories would increase, because the largest items in a category of small items would be underestimated and the smallest items in a category of large items would be overestimated. Likewise, the differences between items within a category would decrease as all items moved toward the central tendency of the category.

Choplin and her colleagues have also found cases in which magnitude differences that cross category boundaries are exaggerated (See Figure 5; Choplin & Hummel, 2002). Specifically, they found that verbally expressible magnitude comparisons (e.g., "larger" and "smaller") affect the mapping of values to magnitude representations (c.f., Hummel & Holyoak, 2001). The presented differences that are less than the differences suggested by a magnitude comparison (such as in CP tasks in which differences are miniscule) will often become exaggerated to better match the larger comparison-suggested difference. In their experiments, Choplin & Hummel presented circles and triangles of varying sizes to their observers. The circles were always slightly larger than the triangles (or vice versa). Half of the observers compared the sizes of the circles to the sizes of the triangles and half did not. The observers who compared sizes recalled items in the large category (e.g., circles) as larger and items in the small category (e.g., triangles) as smaller than did their counterparts in the control condition who did not compare sizes. Importantly, the recalled differences in size across the category boundary were larger than the presented differences across the category boundary. If observers in the CP task were to make similar magnitude comparisons, analogous mappings to magnitude representations would likely occur in CP tasks thereby increasing the represented difference between items that cross a category boundary (See Figure 5).

Together these two models suggest a class of discrimination models that could account for CP effects. If at a delay, observers judge whether or not two magnitudes values are different using recalled and/or represented magnitude values, then the ability to detect differences would be proportional to the differences in recalled and/or represented values rather than the difference between presented values. Since the difference between two recalled and/or represented magnitude values would likely be exaggerated across a category boundary, the ability to distinguish a given presented magnitude difference that crosses a category boundary could very well be greater than the ability to detect the same presented magnitude difference that does not. These processes would produce a pattern of results resembling either expansion (Figure 2) or compression (Figure 3) effects depending upon the baseline used and the biases in recalled and/or represented magnitude values. Huttenlocher and colleagues' model of stimulus judgment predicts both an expansion effect between categories and a compression effect within categories (See Figure 4). By contrast, Choplin and colleagues' model of comparison-induced distortions predicts only an expansion effect between categories (See Figure 5).

Observers may also judge similarity using recalled values, rather than presented values. This class of models (Huttenlocher and colleagues' model of stimulus judgment in particular) could provide a novel account of Goldstone, et al.'s (2001) finding that similarity ratings to a non-categorized face are affected by category training. In particular, if the recalled values of same category faces are biased to be closer together, then their respective differences to a non-categorized face (whose recalled values would presumably not be biased) would resemble each other more after category training than before.

In three experiments, we explored the cognitive mechanisms responsible for CP effects. Observers were trained to categorize visual stimuli. After training to criterion, observers were asked to judge whether two items (pictures of fish; see Figure 6) were the same size or not using a forced-choice same/different task. Stimulus pairs either belonged to the same category or to different categories and we analyzed the accuracy of observers' judgments using signal detection methods. In Experiment 1, we tested a prediction of the perceptual pick-up models, namely, that CP ought to be observed in discrimination tasks in which items are presented simultaneously. In Experiment 2, we tested several predictions of the memory storage models, namely, that CP effects ought to be larger when 1) observers are required to hold category information in memory, and 2) memories of presented magnitude values have faded. In Experiment 3, we explored the role of memory in CP effects.

Experiment 1

Because perceptual pick-up models assume that categories affect perception of magnitudes, they predict that CP effects should be observable in discrimination tasks in which items are presented simultaneously. By contrast, memory storage models predict CP effects only when observers rely on category information to compensate for the declining accuracy of memory for magnitudes after some time has elapsed. Memory storage models therefore would not straightforwardly predict CP effects in discrimination tasks within which items are presented simultaneously. In short, a failure to find CP effects in discrimination tasks when items are presented simultaneously would be troublesome for perceptual pick-up models, but not for memory storage models.

With some exceptions (which we will discuss below), most research on CP has involved sequentially presented items. Phonemes, the stimuli most studied in CP research, have always been presented sequentially; although distinct phonemes could be presented simultaneously from different spatial locations, this approach would arguably introduce complications in terms of interference or emergent features. In any case, it has not been used in studies of CP effects with phonemes. The majority of studies using visual stimuli have also presented items sequentially. For example, Bornstein & Korda (1984) presented colors sequentially. They presented one color for 300 ms, followed by a 250 ms inter-stimulus interval, followed by a second color for 300 ms. Their observers judged whether the two colors were identical or not. In analogous size and brightness discrimination tasks, Goldstone (1994) presented squares sequentially. He presented one square for 1000 ms, followed by a 33 ms inter-stimulus interval, followed by a second square for 1000 ms. Etcoff & Magee (1992) presented facial expressions sequentially in an ABX task, presenting Face A for 750ms, followed by an inter-stimulus interval of 1000ms, followed by Face B for 750ms, then a second inter-stimulus interval for 1000ms, then finally Face X for 1000 ms. Observers judged whether Face X was identical to Face A or to Face B. Roberson and Davidoff (2000) used a similar task in which Color X was presented first; after a 5,000 ms or 10,000 ms inter-stimulus interval, Colors A and B were then presented simultaneously. The pervasive use of sequential presentation in these tasks raises the question of whether CP effects will generalize to conditions of simultaneous presentation.

Experiment 1 tested whether CP effects occur in discrimination tasks when items are presented simultaneously. The stimuli were pictures of fish such as those shown in Figure 6. We varied how "fat" these fish were, their width, which always appeared as the vertical dimension in the displays. The thin fish were called "g fish" and the fat fish were called "h fish." We trained observers to categorize these fish to criterion and then had them judge whether two simultaneously presented fish were identical in size or not. A separate control group was not trained to categorize the fish and we used their performance to measure the perceptible differences between our stimuli prior to category training.

Method

Observers.

The observers were 33 undergraduate students from the University of California, Los Angeles. Of these 33 observers, 17 were trained to categorize fish, and the other 16 served as a control group. The control group was used to measure the perceptible differences between our stimuli without category training. All observers participated to fulfill a course requirement.

Apparatus.

All displays were presented on a Macintosh computer and were programmed in C. The observer sat in a comfortable chair approximately 60 cm away from the monitor. Responses were entered on a standard computer keyboard.

Stimuli.

The fish were drawn horizontally as in Figure 6. The horizontal span from the tip of the nose to the back of the tail was always 260 pixels (approximately 8.22° visual angle): 200 pixels from the tip of the nose to the back of the body (approximately 6.31° visual angle) and 60 pixels from the front of the tail to the far back of the tail (approximately 1.91° visual angle). The tails of the fish were concave at the back. We manipulated the vertical span (fatness) of the fish. The smallest vertical span was 60 pixels (approximately 1.91° visual angle). To minimize the effects of Weber's law subsequent vertical spans, V_i, were given by $V_i = 60 * 1.15^i$, where i represented integer fatness values from 0 (approximately 1.91° visual angle) to 7 (approximately 5.01° visual angle). Throughout the remainder of this paper, we will refer to this scale as the fatness scale and values (0 through 7) on this scale as fatness values. The resulting approximate visual angles were 1.91°, 2.16°, 2.48°, 2.86°, 3.30°, 3.81°, 4.38°, and 5.01°. Fish that had fatness values less than 3.5 (approximately 3.11 ° visual angle) or thinner were classified as "g fish." Fish that had fatness values more than 3.5 were classified as "h fish." The tails of the fish were slightly thinner than the rest of the body having vertical spans 90% of the vertical body span. The eyes of the fish had a diameter of 5 pixels (approximately 0.16° visual angle).

Procedure.

The 17 observers in the experimental group performed two tasks: a categorization training task and a discrimination task. The 16 observers in the control group performed only the discrimination task.

Category Training Task. Observers were told that "h fish" were fat while "g fish" were thin and that they should learn to identify the "border" that divided the "h fish" from the "g fish." On each training trial, a fish with a randomized fatness value of 0 through 7 was presented on the screen for 1,000 ms and was taken away. After the fish was taken away, observers were asked whether they thought the fish was a "g fish" or an "h fish." They were to press the <h> key, if they thought the fish was an h fish or the <g> key, if they thought the fish was g fish. They were given immediate feedback before the next trial began. To continue to the second stage of the experiment, observers were required to categorize 120 fish correctly. Whenever they miscategorized a fish, they were required to categorize 10 more fish correctly in addition to the original 120 fish. For example, if an observer miscategorized 20 fish, she or he was required to correctly categorize 200 extra fish in addition to the original 120, for a total of 320.

Discrimination Task. Observers saw two fish. Their task was to judge whether the fish had identical fatness values (same) or not (different). The fish flashed simultaneously on the screen for 1,000 ms and were removed. One was presented on the left side of the monitor, and the other was presented on the right (counterbalanced) with 1.15° visual angle between the left side fish's tale and the right side fish's nose. The horizontal locations were kept constant and vertical locations were randomized. On some of the trials the fish had the same fatness value; and on the others they did not. When the fatness values of the fish were not identical, the values were adjacent. For example, a fish with a fatness value of 3 was only paired with fish having fatness values of 2, 3, or 4, but no others. Observers pressed the <s> key for same, if they thought the fish were identical and the <d> key for different, if not. The order of the discrimination trials was randomized. For each interval between a fatness value i and an adjacent value j, we collected 60 judgments -- 15 in which i was paired with itself, 15 where j was paired with itself, 15 where i was on the left and j was on the right, and 15 where j was on the left and I was on the right.

Dependent Measures and Data Analysis. The primary measure in this experiment was d', calculated from same/different judgments. To control for Weber effects, we assumed that the averaged d' scores of the observers who did not receive category training reflected the true perceptible differences between our stimuli before category training. Then for each observer who received training, we divided her or his d' score for each interval by the averaged d' scores of the observers who did not receive category training on that same interval. Previous research (e.g., Goldstone, 1994; Levin & Beale, 2000) has found that in addition to altering sensitivity across the interval that crosses the category boundary, categorical perception effects often alter sensitivity over intervals adjacent to this interval. To maximize our power to find CP effects, we adopted a very sensitive measure of CP effects a priori. Specifically, we assessed CP effects by calculating quadratic contrasts[2] on the Weber-controlled sensitivity scores across the 7 intervals. This measure provides a particularly broad and inclusive definition of CP effects, because it measures increased sensitivity in regions adjacent to the category boundary not just at the category boundary itself.

Results and Discussion

The Weber-adjusted discriminability scores are presented in Figure 7. Despite our sensitive measure of CP effects, the quadratic contrasts were not significantly greater than zero (M = -0.45; SD = 2.86), $t < 1$. We replicated these null results in two follow-up experiments that we will not report here. After extensive categorization training, we found no CP effects in a discrimination task within which items were presented simultaneously. These results are troublesome for perceptual pick-up models.

How do these results compare to the few studies in which CP effects have been sought with simultaneously presented items? CP effects have been reported in such studies; however, these studies have used alternative dependent measures (e.g., similarity judgments) rather than psychophysical measures of discrimination. Livingston, Andrews, and Harnad (1998), for example, trained their observers to categorize items and then asked their observers to judge the similarity of items. They found that items within a category were rated more similar than items that crossed a category boundary. This effect does not imply increased ability to discriminate magnitudes that cross category boundaries, however. Most likely within-category items are more similar because they share the property of belonging to the same category, i.e., they share the same category label. Other studies (e.g., Beale & Keil, 1995; Levin & Beale, 2000) have asked observers to identify which of two items most resembles an endpoint. Beale & Keil, for example, morphed President Clinton's face into President Kennedy's face, then asked observers to identify which of the two faces most resembled President Clinton or President Kennedy. They found that observers performed more accurately on this task in the middle of the range (where the faces could not easily be classified as Clinton or Kennedy faces) than at the ends of the range (where the faces were easily classified as Clinton or Kennedy faces). Our concern with this task is that Beale & Keil did not measure sensitivity to differences *per se*. The observers' task was not to judge whether two faces were different, rather the observers' task was to judge the relative likelihood of classifying a face as belonging to either President Clinton or President Kennedy. Unfortunately toward the endpoints of the morph the faces would be uniformly classified as belonging to either President Clinton or President Kennedy. The likelihood of classifying one face as belonging to Clinton, for example, would not significantly differ from the likelihood of classifying the second face as belonging to Clinton. As a result, the task may not be well defined toward the endpoints and the observed results could be the result of this task confusion, rather than the result of real differences in observers' abilities to detect magnitude differences.

The null results of Experiment 1 raise doubts about perceptual pick-up models, but (as is the nature of null results) are not conclusive. To address this concern, we pitted the predictions of the memory storage models against the predictions of the perceptual pick-up models in Experiment 2.

Experiment 2

Unlike the perceptual pick-up models, the memory storage models predict that CP effects will most likely be found when category information is held in memory across the time interval and observers use it to compensate for declining memory of magnitudes. In Experiment 2, we manipulated two factors: 1) the time of presentation in which stimuli were presented, simultaneously (as in Experiment 1) or sequentially, and 2) whether observers were required to hold category information in memory.

The memory storage models assume that category information must be held in memory across a time interval in order to be useful in the discrimination task. To test this assumption, we required half of the observers in Experiment 2 to report the category membership of the presented items. This requirement ensured that participants would hold category information in memory across the time interval for the seemingly irrelevant task of reporting category membership. Category information would then be available to aid discrimination. By contrast, perceptual pick-up models predict no effects *per se* of having observers hold category information in memory across the time interval. That is, because CP is supposed to be an effect of perceptual mechanisms, holding category information in memory should not aid an observer's ability to detect magnitude differences.

Unlike perceptual pick-up models, the memory storage models predict that CP effects ought to be larger for sequentially presented stimuli than for simultaneously presented stimuli. In Experiment 2, we directly tested the effect of inserting a time interval between the presentations of the two stimuli when the observer is attempting to detect size differences. The memory storage models predict that observers discriminating between stimuli that are presented sequentially would produce larger CP effects than would observers discriminating between stimuli that are presented simultaneously. The perceptual pick-up models do not.

Method

Observers.

One hundred undergraduate students from the University of California, Los Angeles (UCLA) participated to fulfill a course requirement.

Stimuli and Apparatus.

The stimuli were identical to those used in Experiment 1. As in Experiment 1, displays were presented by a Macintosh computer and programmed in C.

Design and Procedure.

The design and procedures in Experiment 2 were identical to those in Experiment 1 except in the following respects.

Categorization Training. The category training task was identical to that of Experiment 1 except that when observers miscategorized a fish, they were required to categorize 5 more fish correctly in addition to the original 120 fish. For example, if an observer miscategorized 20 fish she or he would be required to correctly categorize 100 extra fish in addition to the original 120, for a total of 220.[3]

Discrimination Testing. As in Experiment 1, observers saw two pictures of fish on each trial. Two groups of 50 observers were obtained by random assignment. In the Simultaneous Presentation Condition, the two fish were presented simultaneously for 1,000 ms. In the Sequential Presentation Condition, the two fish were presented sequentially. Either the left or the right fish (chosen at random) was presented on the screen for 1,000 ms, followed by a 1,000 ms inter-stimulus interval, followed by the presentation of the second fish for 1,000 ms. Half of the observers in each presentation condition (25 observers) reported the category membership of both the fish on the left and the right on each trial, whereas the other half of the observers did not report category membership. For each interval pair of stimuli with adjacent fatness values i an j, we collected 80 judgments -- 20 in which i was paired with itself, 20 where j was paired with itself, 20 where i preceded j, and 20 where j preceded i.

Data Analysis. As in Experiment 1, sensitivity (d') scores were calculated from the same/different judgments. Again, to control for Weber effects, we used the same procedure that we used in Experiment 1. That is, we divided the d' scores of each observer in Experiment 2 by the averaged d' scores of the control group in Experiment 1. Once again, we utilized the a priori measure of CP effects that we used in Experiment 1. That is, quadratic contrasts were calculated on these Weber-controlled sensitivity scores.

Results and Discussion

The Weber-adjusted discriminability scores are presented in Figure 8. Consistent with the memory storage models, this figure appears to show clear CP effects in the Sequential Presentation Condition in which observers reported category membership. It appears that observers in this condition were able to retain some of the discrimination ability lost over sequential observations by utilizing category information.

To test these apparent effects, we first analyzed the quadratic contrasts for each condition separately using a Bonferroni-adjusted α-level of .0125 ($\alpha = .05/4$). We first analyzed the results from the observers who did not report category membership. We found that neither the contrasts from observers in the Simultaneous Presentation Condition ($\underline{M} = 0.08$; $\underline{SD} = 1.9$, $\underline{t} < 1$) nor the contrasts from observers in the Sequential Presentation Condition ($\underline{M} = 0.94$; $\underline{SD} = 2.43$; $\underline{t}(24) = 1.94$, $\underline{p} > .0125$) were reliably greater than zero We next analyzed the results from observers who reported category membership. We found that the contrasts were not reliably greater than zero from observers in the Simultaneous Presentation Condition ($\underline{M} = 0.76$; $\underline{SD} = 2.41$; $\underline{t}(24) = 1.58$, $\underline{p} > .0125$) but were significantly greater than zero for observers from the Sequential Presentation Condition ($\underline{M} = 2.32$; $\underline{SD} = 2.86$; $\underline{t}(24) = 4.05$, $\underline{p} < .001$).

To test whether these groups were significantly different from each other, a 2 (presentation: simultaneous vs. sequential) x 2 (report category membership: yes or no) between-observer analysis of variance (ANOVA) was performed on these quadratic contrasts. This ANOVA revealed a main effect of presentation condition such that CP effects were larger in the Sequential Presentation Condition than in the Simultaneous Presentation Condition, $\underline{F}(1,96) = 6.24$, $\underline{MSE} = 5.87$, $\underline{p} < .05$. The ANOVA also revealed a main effect of category reporting such that those observers who reported category membership produced larger CP effects than those observers who did not, $\underline{F}(1,96) = 4.50$, $\underline{MSE} = 5.87$, $\underline{p} < .05$. There was no reliable interaction, $\underline{F} < 1$. A Scheffe post-hoc analysis only revealed a reliable difference between the observers in the Sequential Presentation Condition who reported category membership and observers in the

Simultaneous Presentation Condition who did not report category membership. None of the other groups were reliably different from each other. We replicated the finding that observers who report category membership in sequential presentation discrimination tasks exhibit CP effects in an experiment that we will not report here.

Although the results of Experiment 2 were not completely consistent with the predictions of the memory storage models (an interaction was predicted), they were generally more consistent with the memory storage models than they were with the perceptual pick-up models. CP effects were larger in the Sequential Presentation Condition than they were in the Simultaneous Presentation Condition. Furthermore, CP effects were larger among observers who reported category membership than those who did not. Perceptual pick-up models are unable to account for either of these main effects.

The results of this experiment support the memory storage models of CP. However, the role of memory in CP effects is not yet clear. Which memory storage model – the Category Label Storage Model or the Biased Value Storage Models -- provides the best explanation of CP effects? In Experiments 3, we explored the role of memory in CP effects.

Experiment 3

The Biased Value Storage Models predict that observers who show CP effects will also show biases in recall such that recalled differences across the category boundary will be larger than the presented difference across the category boundary. In Experiment 3, we tested this prediction by asking participants to recall sizes.

Assuming that observers categorize "g fish" separately from "h fish," Huttenlocher and colleagues' model of stimulus judgment predicts that recalled differences ought to be larger than presented differences across the category boundary.

In particular, the recalled fatness of the fish with presented fatness values of 0 and 1 would be biased upward and the fish with presented fatness values of 2 and 3 would be biased downward both toward the central tendency of the "g fish" category. Also, the fish with presented fatness values of 4 and 5 would be biased upward and the fish with presented fatness values of 6 and 7 would be biased downward both toward the central tendency of the "h fish" category. Because the fish with the presented fatness value of 3 would be biased downward and the fish with presented fatness values of 4 would be biased upward, the recalled difference between the categories ought to be larger than the presented difference.

Choplin and Hummel's (2002) model of comparison-induced distortions predicts a second pattern of possible results. Assuming that participants are comparing the sizes of the two categories of fish (i.e., "g fish are smaller than h fish" or "h fish are larger than g fish") and the presented difference between the categories is smaller than the comparison-suggested difference (as would most certainly be the case with these stimuli in which differences are miniscule), their model predicts that represented differences would be larger than presented differences across the category boundary. In particular, to match the larger comparison-suggested difference the mapping of presented sizes to magnitude representations would be biased to make the represented difference larger, thereby making all of the "g fish" smaller and all of the "h fish" larger. If observers who show CP effects were to show either of these patterns of bias, this finding would lend support to the Biased Value Storage Models.

By contrast, several patterns of results would be troublesome for the Biased Value Storage Models. Huttenlocher and colleagues' model of stimulus judgment makes very different predictions than those described above, if observers place all fish in the category "fish presented in this experiment." In such a case, their model predicts that fish with presented fatness values of 0, 1, 2, and 3 would be biased upward and fish with presented fatness values of 4, 5, 6, and 7 would be biased downward both toward the central tendency of the category "fish presented in this experiment." In such a case, recalled differences would be smaller across the category boundary than elsewhere. If observers who show CP effects were to show this pattern of bias or no bias at all, the results would be troublesome for the Biased Value Storage Models. Either of these patterns would leave the Category Label Storage Model as the most parsimonious explanation of CP effects.

We tested these predictions in Experiment 3. After completing category training and discrimination tasks such as those in the previous experiments, observers recalled the sizes of fish.

Method

The observers were 13 undergraduate and graduate students at the University of Chicago. They were paid $15 for their participation. The stimuli were fish identical to those used in Experiments 1 and 2, except the smallest vertical span was 70 pixels (approximately 2.23° visual angle) and this height was increased by 25%, rather than 15%, exponentially from 0 to 7 times. That is, vertical spans, V_i, were given by $V_i = 70 * 1.25^i$, where i represented integer fatness values from 0 (approximately 2.23° visual angle) to 7 (approximately 10.45° visual angle). The resulting approximate visual angles were 2.23°, 2.74°, 3.43°, 4.32°, 5.40°, 6.72°, 8.41°, and 10.45°. Fish that had fatness values less than 3.5 (approximately 4.85 ° visual angle) or thinner were classified as "g fish." Fish that had fatness values more than 3.5 were classified as "h fish." The training task was identical to the training task in Experiment 2. In the same/different task stimuli were presented sequentially, the inter-stimulus interval was always 1,000ms, and observers were required to report category membership. For each interval between a fatness value i and an adjacent value j, we collected 40 same/different judgments -- 10 in which i was paired with itself, 10 where j was paired with itself, 10 where i preceded j, and 10 where j preceded i.

Size Recall Task. After each quarter of the discrimination trials (i.e., after 25%, 50%, 75%, and 100% of the trials), observers were asked to recall the sizes of each of the eight fish (fatness values 0 through 7). In a random order, one of the eight fish flashed on the screen for 1,000ms and was taken away. A second fish of random sized was then presented on the screen and observers adjusted the size of this fish to match their memory of the size of the first fish.

Data Analysis. As in the previous two experiments, sensitivity (d') scores were calculated from the same/different judgments. Unlike the previous two experiments, however, we had no control group from which to calculate Weber-controlled sensitivity scores. As a result, quadratic contrasts were calculated directly on the d' scores. We calculated the recalled size of each fish by averaging the responses from the four times each observer had recalled the size of that fish. We then calculated the amount of bias by subtracting presented sizes from recalled sizes.

Results and Discussion

We first tested for CP effects. Replicating the results of Experiment 2, the contrasts on the d’ scores were marginally greater than zero [$M = 3.24$; $SD = 6.63$; $t(12) = 1.76$, $p < .1$]. This effect may have been marginal (rather than completely reliable) in this experiment due to reduced power, i.e., fewer observers, fewer judgments per observer (more noise), and no control for Weber effects. We next analyzed the recalled sizes of fish. The average recalled sizes of the fish with presented fatness values of 0, 1, and 2 were all significantly larger than presented [M = 13.23 pixels; SD = 13.74; t(12) = 3.47, p < .0005 for fatness values of 0; M = 10.98 pixels; SD = 10.02; t(12) = 3.95, p < .0005 for fatness values of 1; and M = 9.98 pixels; SD = 10.74; t(12) = 3.35, p < .005 for fatness values of 2 respectively]. The average recalled sizes of the fish with presented fatness values of 3 and 4 were not significantly different than presented [M = 2.35 pixels; SD = 9.71; t < 1 for fatness values of 3; and M = -0.83 pixels; SD = 7.19; t < 1 for fatness values of 4 respectively]. Finally, the average recalled sizes of the fish with presented fatness values of 5, 6, and 7 were all significantly smaller than presented [M = -8.5 pixels; SD = 15.7; t(12) = 1.95, p < .05 for fatness values of 5; M = -10.7 pixels; SD = 14.51; t(12) = 2.65, p < .01 for fatness values of 6; and M = -19.9 pixels; SD = 30.74; t(12) = 2.33, p < .05 for fatness values of 7 respectively]. Most importantly, the difference in the recalled values of fish with presented fatness values of 3 and 4 (M = 20.17 pixels; SD = 9.34; the difference that spans the category boundary) was not significantly different from the presented difference (17 pixels).

These results are inconsistent with the predictions of Huttenlocher et. al's (1991, 2000) model of stimulus judgment, given the added assumption that observers would treat all fish as if they belonged to the "g fish" and "h fish" categories. They are also inconsistent with the pattern of results predicted by Choplin and Hummel's (2002) comparison-induced distortions given the added assumption that observers would compare categories. They are, however, consistent with Huttenlocher's model of stimulus judgment, given the added assumption that observers treated all of the fish as if they belonged to the category "fish presented in the experiment." This pattern of results is, therefore, inconsistent with the Biased Value Storage Models leaving the Category Label Storage Model as the most parsimonious explanation of CP effects.

General Discussion

In three experiments, we investigated the cognitive mechanisms responsible for CP effects. In Experiment 1, we tested a prediction of the view that perceptual mechanisms are responsible for CP effects, namely that CP effects ought to be observable for items presented simultaneously in the discrimination task. However, when items were presented simultaneously, no CP effects were observed. This null result was troublesome for the view that perceptual mechanisms are responsible for CP effects.

In Experiment 2, we tested a prediction of the view that memory mechanisms are responsible for CP effects. Namely, we tested the prediction that category information aids detection of magnitude differences only if it is used to compensate for declining memory of exact presented magnitudes. To test this prediction, we manipulated two factors in the CP discrimination task. The first factor was whether observers were required to report the category membership of the two items. This manipulation ensured that observers would have to hold category information in memory for this seemingly irrelevant task and it would then be available to aid the detection of magnitude differences that cross the category boundary. The second factor was presentation time. Items in the discrimination task were presented either simultaneously or sequentially. We found that those observers who reported category membership showed greater CP effects than did those who did not. We also found that observers who attempted to detect differences between sequentially presented stimuli showed greater CP effects than observers who attempted to detect differences between simultaneously presented stimuli. These findings are consistent with the view that memory mechanisms are responsible for

CP effects and inconsistent with the view that perceptual mechanisms are responsible for CP effects.

In Experiment 3, we explored the specifics of memory's role in CP effects. In particular, we investigated whether memory of category labels or biased memory of magnitude values was responsible for CP effects. To investigate this question, we asked observers to recall magnitude values. If biases in the recalled and/or represented magnitude values were responsible for CP effects, then observers who exhibited CP effects ought to have recalled values such that the recalled differences across the category boundary should have been larger than the presented differences, but they were not. These results suggest that memory of category labels is most likely responsible for CP effects.

Recent work by Roberson and Davidoff (2000) provides converging evidence for this conclusion. In a series of experiments, they investigated the effects of verbal and visual interference on categorical perception of colors and facial expressions. The basic idea behind their experiments was that introducing modality-specific interference could isolate the codes that observers use in performing CP discrimination tasks. If visual codes were responsible for CP effects, then visual interference ought to have disrupted the effect. If, by contrast, verbal codes were responsible for CP effects, then verbal interference ought to have disrupted the effect. In their CP discrimination effect, Roberson and Davidoff asked their observers to view a target item (either a color or a face). This target item was followed by a long inter-stimulus interval (5,000 or 10,000 ms) and visual interference (e.g., colored dot patterns or visual facial features), verbal interference (i.e., words), or no interference (control condition) was inserted during the inter-stimulus interval. After the interval observers viewed two items one of which was identical to the target and the observers identified the identical item. Roberson and Davidoff found robust CP effects under conditions of visual and no interference, but they failed to find CP effects under conditions of verbal interference. They concluded that verbal codes are responsible for CP effects.

These findings have direct implications for models of CP effects. Specifically, because the Perceptual Pick-up Models have traditionally held that the locus of CP effects hinges on visual or perceptual algorithms, they have not specified roles for verbal codes in creating CP effects. To date these models have offered no reasons why verbal interference ought to disrupt CP effects. By contrast, the Category Label Storage Model has already specified a role for verbal codes in creating CP effects. Specifically, category labels presumably held as verbal codes are used to preserve an observer's ability to detect differences across the category boundary. The results of Roberson and Davidoff, therefore, provide converging evidence for the view that verbally held category labels are responsible for CP effects.

In this article, we have cited research on CP effects over a variety of category types. However, the focus of our research has been on models of CP effects in learned categories. We leave analysis of the question of whether the Category Label Storage Model can account for CP effects in innate categories to future research. In particular, some color categories may be innate and, if so, our analysis of CP effects in learned categories may not be directly applicable to these color categories. There are several reasons to think that our analysis of learned categories may not apply to color categories. For example, Bornstein, Kessen &Weiskopf (1976) found that 4-month-old infants showed slower dishabituation to color changes that crossed adult category boundaries than to color changes that did not. Because this stage of development is prior to language acquisition it is difficult to argue that infants are using verbal labels to preserve their ability to detect differences that cross the category boundary. Another explanation is needed to accommodate these data.

Other color categories seem to be learned. For example, whereas English speakers have a single color word for blue, Russian speakers have one color word for light blue and a separate color word for dark blue. Because these color categories are language and culture-specific, they must be learned (see our discussion of Whorf's hypothesis below). Therefore, our analysis ought to be applicable to these categories. Recent work by Frank and Boroditsky (2001) supports the hypothesis that CP effects across the light blue/dark blue category boundary arise among Russians when they use category labels to preserve their abilities to detect differences after memories of exact magnitudes have faded. Paralleling our analysis, Frank and Boroditsky found that Russian speakers were better able to discriminate between two sequentially presented blues when one blue was light and the other dark than when both were light or both were dark. However, these same speakers showed no such cross-category advantage for simultaneously presented blues. By contrast, English speakers were no better able to discriminate differences that crossed the category boundary than differences that did not in either condition. These results are consistent with the view that some CP effects involving color categories (i.e., those involving learned color categories) occur when category labels are used to preserve the ability to detect differences that cross category boundaries after memory of exact magnitudes has faded.

Our conclusion that the Category Label Storage Model sufficiently explains CP effects involving learned categories has a number of implications. We discuss the implications of this conclusion for category coherence, Whorf's hypothesis, and the applicability of models of magnitude estimation to perceptual discrimination.

Category Coherence

A central issue for research on categorization is the problem of category coherence. What allows all of the diverse members of a category to be put together in one group? Why aren't other items included in this group? A partial answer to these very complex questions suggests that preferred categories will maximize the similarity of items in the category to each other (maximize within-group similarity) and minimize the similarity of these items to items not in the category (minimize between-group similarity). However, if the similarity between items is defined as the objective distance between items on continuous dimensions of measurement, then this answer will be untenable. Any item close to a would-be category boundary will necessarily be just as similar to an item x units across the category boundary as to an item x units toward the central tendency of its own category. There would, therefore, be no reason to place a category boundary at one location on the dimension as opposed to another. (See Davidoff, 2001, for an extended discussion of this problem.)

Several researchers (notably, Livingston, Andrews, & Harnad, 1998) have pointed to CP effects as a possible solution to this problem. In their view, category learning causes a warping of distances on the continuous dimension that artificially increases within-category similarity and/or decreases between-category similarity. That is, they argue that the relevant similarity between items is not the objective distances between items on objective dimensions of measurement, rather the relevant similarity is the subjective distances between items on artificially warped dimensions. CP effects are, then, a measure of the warping of similarity space that creates category coherence during category learning.

By contrast, our results suggest that observers must be able to categorize – and apparently be able to appreciate the coherence of categories – before CP effects can arise. If observers did not have this ability and they were unable to associate category labels with items, then the labels would not be available to assist them in the CP discrimination task. Our results, therefore, suggest that CP effects have very little to do with category coherence. Notice, however, that simple associations with category labels may provide a small degree of category coherence especially in artificial category learning experiments such as those often used to study CP effects.

Whorf's Hypothesis

Davidoff (2001) argued that CP effects demonstrate a case in which the language an observer speaks (the categories an observer uses) affects how she or he perceives stimuli (Whorf, 1956). To make this argument, he reviewed research on the differences in color categorization behavior between English speakers in the United Kingdom and Berinmo Speakers in Papua, New Guinea. Berinmo speakers have a color category named "wor" that includes some oranges, yellow, some browns, and yellowish green. They also have another category named "nol" that includes most greens and blues. In three different studies involving similarity judgments, the ease of category learning, and recognition memory, Berinmo speakers showed CP-like effects across the wor-nol category boundary, but not the blue-green category boundary. By contrast, English speakers showed CP-like effects across the blue-green category boundary, but not the wor-nol category boundary. Davidoff concluded that these effects reflect true differences in perceptual sensitivity caused by the language an observer speaks.

Our results, along with the results of Franks and Boroditsky (2001) suggest a different interpretation of the studies that Davidoff (2001) reviewed. The differences between Berinmo and English speakers could arise simply because the speakers of different languages have access to different category labels. These category labels could, in themselves, bias similarity judgments, make some categories easier to learn than others, and aid recognition memory. That is, the differences between Berinmo and English speakers may be differences in the judgments speakers are likely to make, rather than true perceptual differences. These differences, then, are consistent with a growing literature demonstrating the effects of language on judgments of magnitude (e.g., Boroditsky, 2001; Choplin & Hummel, 2002). They are not, however, a definitive demonstration that the language an observer speaks affects how she or he perceives stimuli.

Perceptual Discrimination

A long line of research has investigated systematic biases in magnitude judgments. This research has documented numerous ways in which judged magnitudes are not isomorphic with objective magnitudes. For example, the observer's adaptation level (Helson, 196?), the range and frequency of stimuli (Parducci, Knobel & Thomas, 1976), the central tendency of categories (Huttenlocher, Hedges & Vevea, 2000), and magnitude comparison (Choplin & Hummel, 2002) all create nonisomorphisms between judged and objective magnitudes. One of our primary motivations for pursuing the line of research we report here was to investigate the effects of these nonisomorphisms on perceptual discrimination. The basic idea was that observers might rely on judged magnitudes, rather than objective magnitudes, when engaged in perceptual discrimination tasks. If so, perceptual discrimination would have been best in areas where judged magnitude differences were largest.

The results of Experiment 3 suggest that observers' perceptual discrimination abilities will not necessarily mirror their judged magnitude differences. In Experiment 3, observers who exhibited CP effects – better discrimination across the category boundary than within each of the categories – recalled the sizes of items surrounding the category boundary as being closer together than they were presented. This result is the opposite of what would be predicted if observers made discriminations based upon judged magnitude differences rather than presented magnitude differences. This dissociation between perceptual discrimination and judged magnitude differences makes it unlikely that research on judgments of magnitude will inform research on perceptual discrimination.

Conclusion

The results of the experiments we report suggest that CP simply reflects observers' abilities to label stimuli. They suggest that CP effects most likely do not reflect conceptual coherence created during category learning. They probably do not demonstrate an influence of language on the perception of stimuli. Nor do they involve the effects of magnitude judgments on perceptual discrimination.

CP effects are still interesting, however, in that they demonstrate the role of labels in magnitude judgments. Apparently, CP effects reflect an observer's ability to use labels to preserve the accuracy of magnitude judgments despite declining accuracy of memory for magnitudes. There are many such examples of the role of labels in magnitude judgments. For example, standard algorithms used to apply labels (e.g., "slightly," "medium," "very") to magnitudes have been found to affect a large variety of magnitude judgments including the judged sizes of geometric shapes (Parducci, Knobel & Thomas, 1976), judgments of fair grading (Wedell, Parducci & Roman, 1989), and hedonic judgments such as the pleasantness of tastes (Riskey, Parducci & Beauchamp, 1979; Parducci, 1995). Labels, therefore, play an extremely important role in providing information that observers can then use to make magnitude estimates. Separating the effects of labels on magnitude judgments from the effects of other heuristic estimation strategies and/or changes in magnitude representation will most likely have to be done on a case by case basis (Wedell, 1995).

References

Abbs, J. H. & Sussman, H. M. (1971). Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. *Journal of Speech and Hearing Research*, 14, 23-36.

Beale, J. M. & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, 57(3), 217-239.

Bornstein, M. H. & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, 46(3), 207-222.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1-22.

Choplin, J. M. & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General*.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, 5(9), 382-387.

Etcoff, N. L. & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44, 227-240.

Frank & Boroditsky, L. (May 19, 2001). *Language affects memory, but does it affect perception?* First Annual Stanford Undergraduate Research Conference.

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

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

Graham, C. H. (1965). Discriminations that depend on wavelength. In C.H. Graham (Ed.). Vision and visual perception. New York: Wiley.

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

Harnad, S. (1987). Category induction and representation. In S. Harnad, (Ed.), *Categorical perception: The groundwork of cognition* (535-565). New York: Cambridge University Press.

Hummel, J. E. & Holyoak, K. J. (2001). A process model of human transitive inference. In M. Gattis (Ed.). *Spatial Schemas and Abstract Thought*, Cambridge, MA:MIT Press.

Huttenlocher, J., Hedges, L. V. & Vevea, J. L. (2000). Why do categories affect stimulus judgments. *Journal of Experimental Psychology: General*, 129(2), 220-241.

Judd, D. B. (1932.) Chromaticity sensibility to stimulus differences. Journal of the Optical Society of America, 22, 72-108.

Levin, D. T. & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception & Psychophysics*, 62(2), 386-401.

Livingston, K. R., Andrews, J. K. & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 24 (3), 732-753.

Massaro, D. W. (1976). Auditory information processing. In W. K. Estes (Ed.), Handbook of learning and cognitive processes, Vol. 4: Attention and memory. Hillsdale, NY: Erlbaum.

Paap, K. R. (1975). Theories of Speech Perception. In D. W. Massaro (Ed.), Understanding language: An information-processing analysis of speech perception, reading, and psycholinguistics. (pp. 151-204.) New York: Academic Press.

Parducci, A. (1995). Happiness, pleasure, and judgment: The contextual theory and its applications. Hillsdale, NJ: Lawrence Erlbaum Associates.

Parducci, A., Knobel, S. & Thomas, C. (1976). Independent Contexts for Category Ratings: A Range-Frequency Analysis. *Perception & Psychophysics*, 20(5), 360-366.

Riskey, D. R., Parducci, A. & Beauchamp, G. K. (1979). Effects of context in judgments of sweetness and pleasantness. *Perception & Psychophysics*, 26(3), 171-176.

Roberson, D. and Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, 28(6), 977-986.

Shiffrin, R. M. & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, 101, 357-361.

Smith, D. P. (1971). Derivation of wavelength discrimination from colour-naming data. Vision Research, 11, 739-742.

Wedell, D. H. (1995). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception & Performance*, 21, 1158-1173.

Wedell, D. H. Parducci, A. & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology*, 102(2), 233-248.

Whorf, B. L., (1956). *Language, thought, and reality; selected writings*. Cambridge, MA: Technology Press of Massachusetts Institute of Technology.

Author Notes

Preparation of this paper was supported by NSF (SBR-9720410) and NEI research grants (EY13518-01) to Philip J. Kellman. We are grateful to Janellen Huttenlocher, John E. Hummel, and Elizabeth L. Crawford for helpful comments and to Sean Duffy, Chad S. Sampanes, Tannaz Sassooni, Stanley Chiao, Deanna Arutyunyan and Shlomit Shlomof for their help running experiments. Correspondence should be directed to Jessica M. Choplin Vanderbilt University Psychology Department; 111 21st. Ave. S.; Nashville, TN 37203. E-mail: jessica.choplin@vanderbilt.edu.

Figure Captions

Figure 1. The Categorical Perception (CP) Effect. Sensitivity is greater across the category boundary than within categories.

Figure 2. Expansion Effect. The straight line represents observers' abilities to detect magnitude differences before category learning and the curved line represents observers' abilities to detect magnitude differences after category learning. Category learning increases sensitivity across the category boundary.

Figure 3. Compression Effect. The straight line represents observers' abilities to detect magnitude differences before category learning and the curved line represents observers' abilities to detect magnitude differences after category learning. Category learning decreases sensitivity within categories.

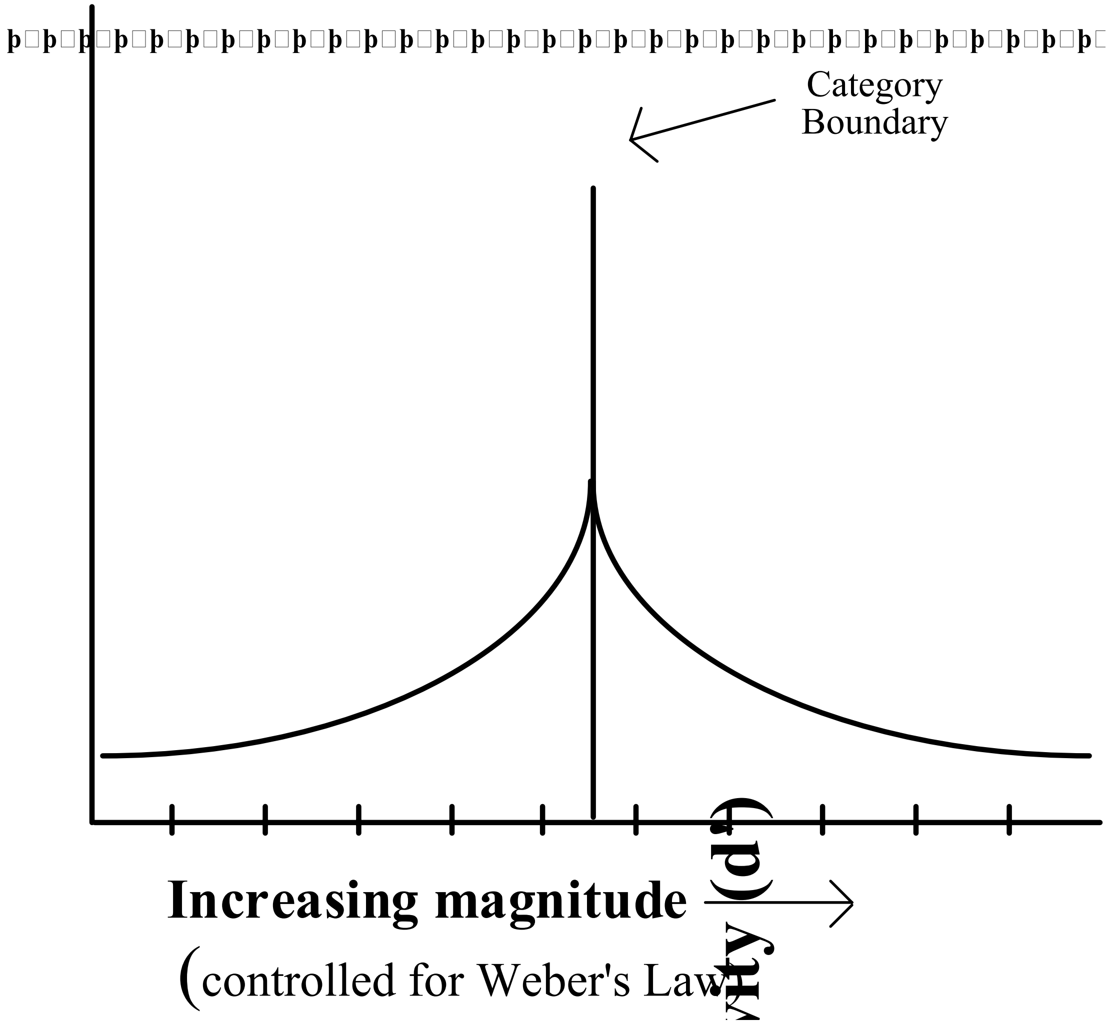
Figure 4. Huttenlocher and colleague's model of stimulus judgment predicts that recalled differences could be larger than presented differences across the category boundary and smaller than presented differences within categories.

Figure 5. Choplin and colleague's model of comparison-induced distortions predicts that if observers compare categories, then represented differences could become larger than presented differences across the category boundary.

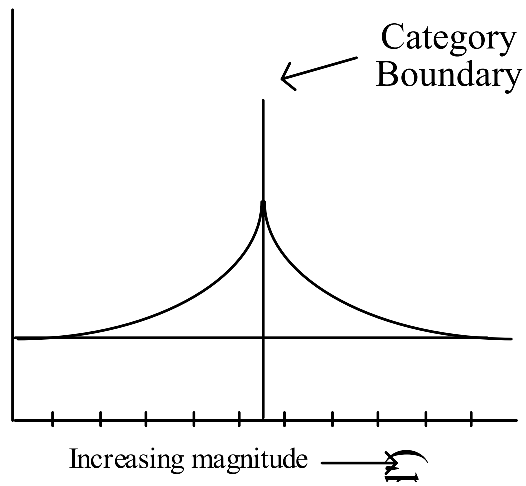
Figure 6. Fish Stimuli.

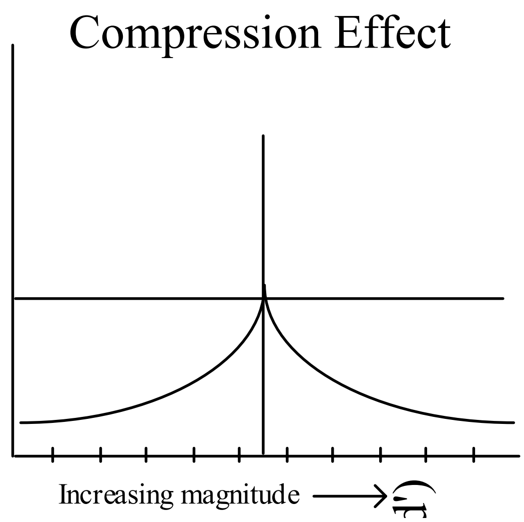
Figure 7. Results of Experiment 1. No CP effects were observed.

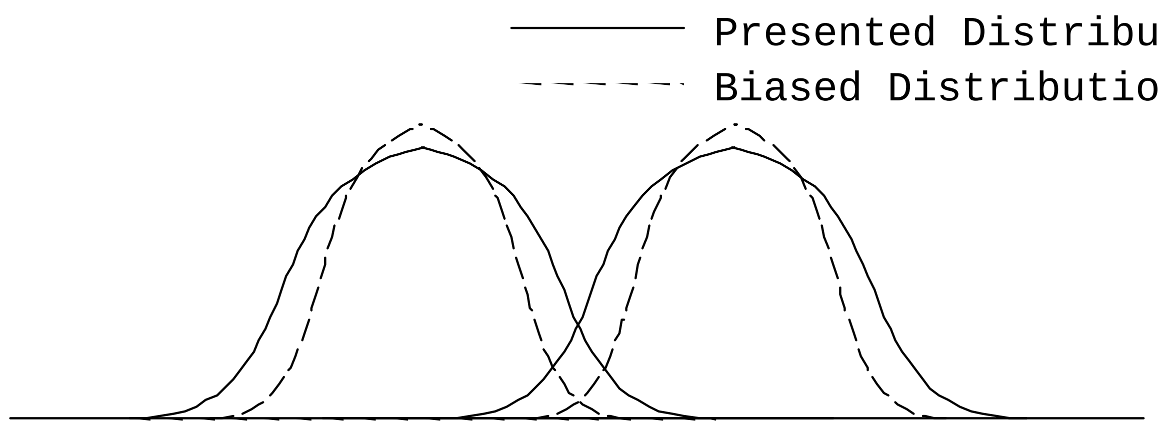
Figure 8. Results of Experiment 2. Reliable CP effects were observed only when items were presented sequentially and observers reported category information.

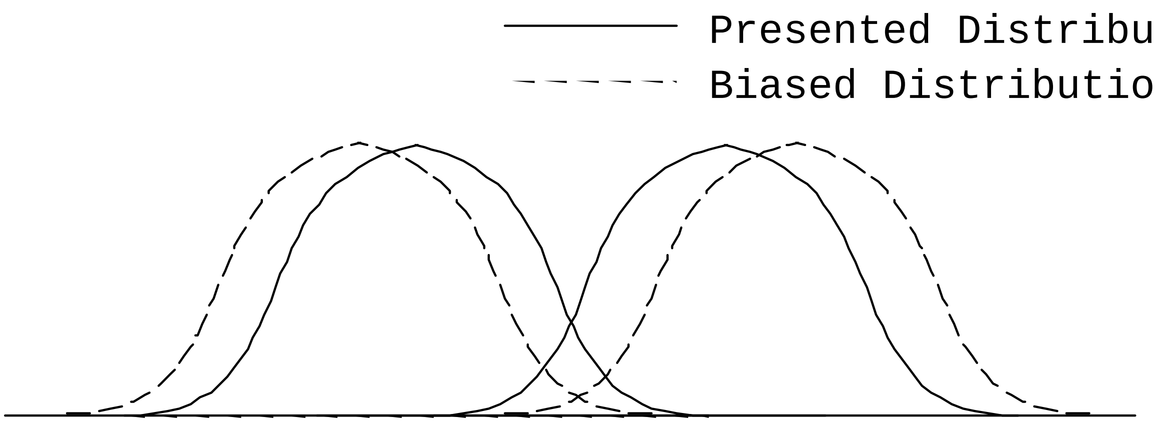

Category
Boundary
Increasing magnitude
(controlled for Weber's Law

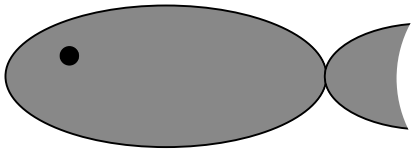
Expansion Effect

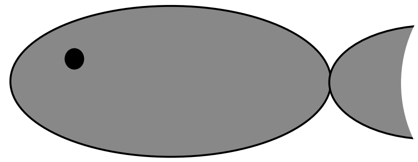


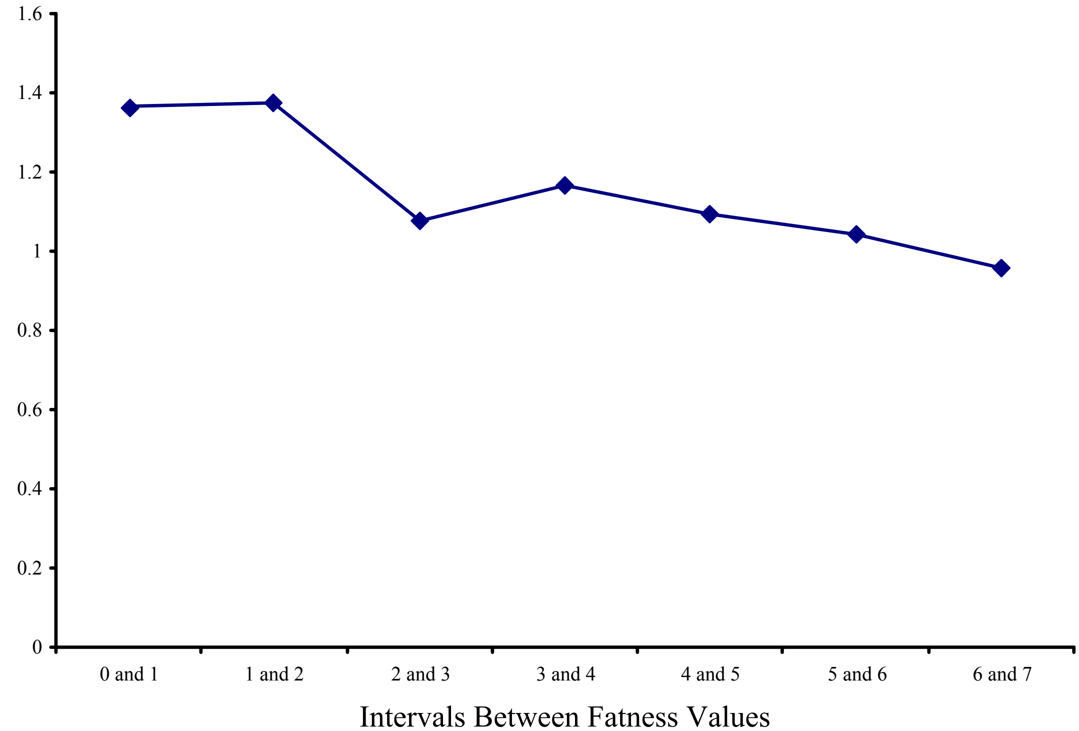

Compression Effect
Increasing magnitude

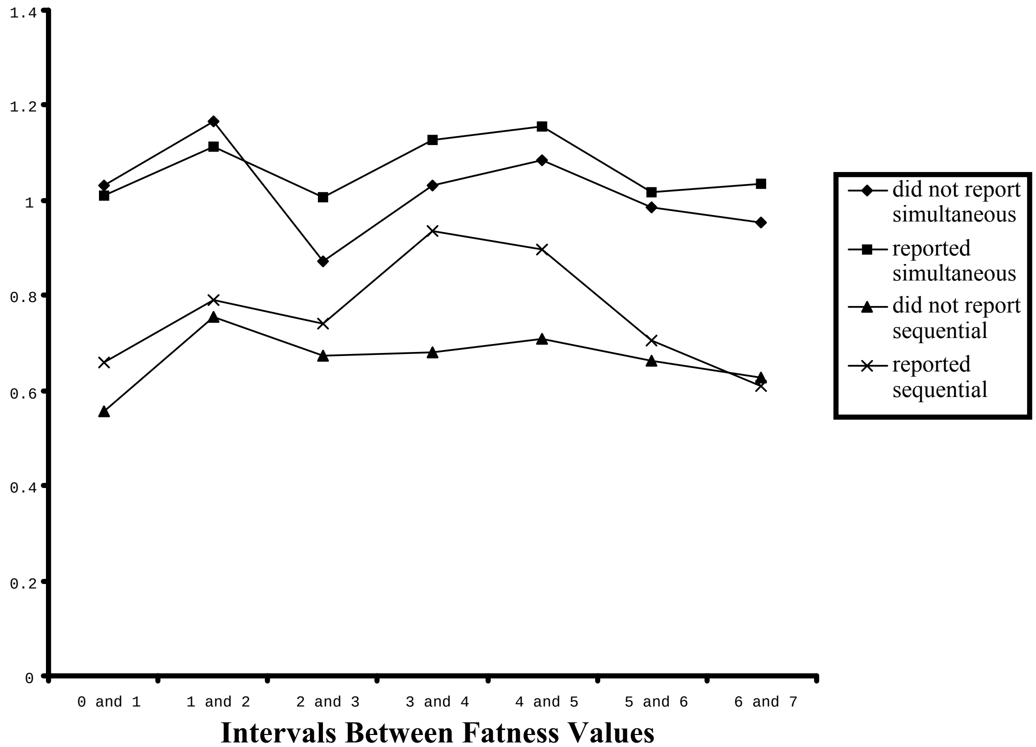

Presented Distribu
Biased Distributio


Presented Distribu
Biased Distributio






1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 and 1
1 and 2
2 and 3
3 and 4
4 and 5
5 and 6
6 and 7
Intervals Between Fatness Values


1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 and 1
1 and 2
2 and 3
3 and 4
4 and 5
5 and 6
6 and 7
Intervals Between Fatness Values
did not report simultaneous
reported simultaneous
did not report sequential
reported sequential

[1] We use the term "categorical perception" to refer to this empirical phenomenon. However, the phenomenon may or may not rely upon perceptual mechanisms. Note also that this definition of CP is very different from the definition historically used by researchers working on phoneme perception (Paap, 1975). This definition comes from more recent work looking at analogous effects in learned categories of visual stimuli (Goldstone, 1994).

[2] The coefficients were {-5, 0, 3, 4, 3, 0, -5}.

[3] The reason for this design change was that Experiment 1 was performed chronologically after Experiment 2. We added more training to Experiment 1 to see whether more training would produce CP effects given simultaneous presentation, it did not.