Journal of Experimental Psychology: General
2002, Vol. 131, No. 4, 477–493

Copyright 2002 by the American Psychological Association, Inc.
0096-3445/02/$5.00 DOI: 10.1037//0096-3445.131.4.477

Acquisition of Categorical Color Perception: A Perceptual Learning Approach to the Linguistic Relativity Hypothesis

Emre Özgen and Ian R. L. Davies
University of Surrey

Color perception can be categorical: Between-category discriminations are more accurate than equivalent within-category discriminations. The effects could be inherited, learned, or both. The authors provide evidence that supports the possibility of learned categorical perception (CP). Experiment 1 demonstrated that observers' color discrimination is flexible and improves through repeated practice. Experiment 2 demonstrated that category learning simulates effects of "natural" color categories on color discrimination. Experiment 3 investigated the time course of acquired CP. Experiment 4 found that CP effects are acquired through hue- and lightness-based category learning and obtained interesting data on the dimensional perception of color. The data are consistent with the possibility that language may shape color perception and suggest a plausible mechanism for the linguistic relativity hypothesis.

This article explores the nature and origin of color categorization by considering how categorical perception (CP; e.g., Harnad, 1987), linguistic relativity (e.g., Whorf, 1940/1956), and perceptual learning (e.g., Goldstone, 1998) are related. These three areas of inquiry overlap because, first, differences in CP across languages are one manifestation of linguistic relativity (e.g., Kay & Kempton, 1984), and, second, perceptual learning is a likely route through which language influences color cognition, and more specifically through which CP is induced. Our experiments explored whether adult color discrimination can be improved through training and whether CP is induced by learning novel color categories. We argue that if they can be, then this increases the plausibility that similar mechanisms may be involved during language learning, giving rise to relativistic effects. Before explicating this sketch of our argument further, we briefly review each of the three areas.

CP of Color

Color perception is categorical. The wavelength continuum of the spectrum is perceived as qualitatively different categories designated in English by terms such as *red, green, yellow,* and *blue*. Moreover, discrimination of stimuli separated by a category boundary is faster and more reliable than that of equivalently spaced within-category stimuli (e.g., Bornstein & Korda, 1984; Boynton, Fargo, Olson, & Smallman, 1989; Roberson & Davidoff, 2000).

Such categorical effects could be innate and universal, or they could be learned, channeled by linguistic categories. Bornstein, Kessen, and Weiskopf (1976) found apparent CP in 4-month-old infants. After habituating to a given color, infants looked more at a novel stimulus if it came from a different adult color category than if it came from the same category as the habituation stimulus, although both types of stimulus were *physically* (in wavelength terms) equally distant from the habituation stimulus. Thus, hue discriminability does not map uniformly onto wavelength. Color order systems such as Munsell are intended to be perceptually uniform. Equal size steps in Munsell hue have the same perceptual distance across color space. Nevertheless, adults still show CP even when the interstimulus distances are equated in Munsell units (Bornstein & Korda, 1984; Roberson & Davidoff, 2000). Gerhardstein, Renner, and Rovee-Collier (1999) equated distances between the test and the habituation stimuli perceptually (in Munsell space) and found no CP effects for infants of 4 months. These findings suggest that along the physical continuum of light wavelength, infants' color perception shows similar categorical effects to that of adults, but there may still be scope for language to influence color perception, including acquired CP across the borders of a language's main color terms.

The plausibility of acquired CP for color is strengthened by the prevalence of acquired CP for other stimulus domains. For example, acquired CP effects were found for size and lightness (Goldstone, 1994) and for quasi microorganisms (Livingston, Andrews, & Harnad, 1998). However, we know of no evidence that color CP can be acquired, and one of the aims of our experiments was to explore this issue.

Linguistic Relativity

The linguistic relativity hypothesis (LRH) proposes that languages differ greatly in the way they "break down" the natural world and that the mental processes, or thoughts, of the speakers of a language will be shaped accordingly (Whorf, 1940/1956). Color

---

Emre Özgen and Ian R. L. Davies, Department of Psychology, University of Surrey, Surrey, England.

This work is part of Emre Özgen's doctoral studies (Özgen, 2000). The study was funded in part by a scholarship to Emre Özgen by the Psychology Department at the University of Surrey and in part by Economic and Social Research Council Grant R000238310. We also owe many thanks to Paul Sowden, Stevan Harnad, Debi Roberson, and an anonymous reviewer for their useful comments on this work and on the manuscript.

Correspondence concerning this article should be addressed to Emre Özgen, Department of Psychology, University of Surrey, Guildford, Surrey GU2 7XH, England. E-mail: e.ozgen@surrey.ac.uk

categorization has been a suitable testing ground for the LRH. This is mainly because, at first glance, it seemed that languages divided and encoded the continuous color space in very different ways. Studying color perception and cognition of speakers of different languages seemed therefore an ideal way of testing the predictions of the LRH.

Berlin and Kay's (1969) and Rosch–Heider's classic studies (Heider, 1971, 1972) seemed to have irrevocably established that there was little if any effect of language on color perception and cognition. There were dissenting voices from the general consensus, most notably from Lucy and colleagues (Lucy & Schweder, 1979), and more recently, Rosch's interpretation of some of her findings have been questioned (e.g., Ratner, 1989; Saunders & van Brakel, 1997). In addition, a growing number of field studies have found cross-language differences consistent with language influencing color cognition to some degree. These effects include differences in CP consistent with language differences (Kay & Kempton, 1984; Roberson, Davies, & Davidoff, 2000), as well as more general effects that may be partly due to differences in CP (e.g., Davies & Corbett, 1997, 1998; Özgen, 2000). In the former cases, speakers of languages that categorize blue and green together showed no CP effect across the blue–green boundary, whereas English speakers did. In the latter cases, if colors were distinguished linguistically, they were less likely to be grouped together in perceptual grouping tasks than if they were not so distinguished. This may be in part because cross-category perceptual distances have been stretched, creating a dislocation in perceptual color space that is used as the basis for grouping. Blues look less like greens if the distinction is marked linguistically.

All of the above studies are vulnerable to the criticism that the observed differences may merely be due to a direct naming strategy. For instance, in a triad task—which color is least like the other two—Kay and Kempton (1984) found CP effects around the blue–green boundary for English speakers but not for Tarahumara speakers who do not distinguish between blue and green linguistically. This may occur because participants choose the color with a different label from the other two. For instance, with a Blue 1–Blue 2–green triad, in which the Blue 1–Blue 2 perceptual distance is similar to the Blue 2–green distance, English speakers may use the labels to resolve the perceptual problem. This option was not available to Tarahumara speakers. Roberson and Davidoff (2000) found evidence consistent with CP being a direct language effect. CP was eliminated if verbal labeling was prevented, by getting participants to read a list of words (related or unrelated to color) out loud. Similarly, in a perceptual grouping task, participants may again use language to resolve difficult perceptual decisions. In the groups formed, almost inevitably, some colors in one group may be more similar to some in another group than to some of those in their own group. (For example, a blue stimulus near the blue–green boundary may be more similar to a green one near but on the other side of the boundary than it is to a blue stimulus far from it.) Having different labels could attenuate the effect of their perceptual closeness and provide a basis for grouping them separately. Participants claim that they are doing these tasks purely perceptually, but nevertheless, there may be some implicit direct language effect. There is evidence that if color names are not available, color grouping can be very difficult. Roberson, Davidoff, and Braisby (1999) found that a male anomic had great difficulty with a color-grouping task, although he still showed CP in a two-alternative-forced-choice (2-AFC) discrimination task.

To explore whether color perception is affected by the acquisition of color categories, we used perceptual and category learning paradigms in the present study and tested for induced CP effects. Such a finding would support the possibility that cross-language differences in behavior on color tasks similarly arise as a result of the influence of learning linguistic categories on perception. These CP effects could arise through a process of perceptual learning.

Perceptual Learning

Perceptual learning is typically characterized by an improvement in stimulus discrimination or detection after repeated practice. It occurs for a variety of visual phenomena, such as grating discrimination (Fine & Jacobs, 2000), hyperacuity (Fahle, 1997), motion discrimination (Zanker, 1999), orientation identification (Schoups, Vogels, Qian, & Orban, 2001), contrast detection (Sowden, Rose, & Davies, 2002), and stereo acuity (Sowden, Davies, Rose, & Kaye, 1996; for reviews, see Karni & Bertini, 1997; Goldstone, 1998).

In most cases learning requires that the relevant stimulus property be attended to. For instance, improvement in orientation discrimination does not occur if the brightness of the lines rather than their orientation is attended to (Shiu & Pashler, 1992). Similarly, improved discrimination of target location does not occur if the background shape is attended to or vice versa (Ahissar & Hochstein, 1993). There is some evidence suggesting that learning vernier acuity (sensitivity in detecting whether two parts of a broken line are aligned) can take place without explicit attention; training on the horizontal component of a vernier offset stimulus leads to performance gains on both the horizontal and vertical components (Weiss, Edelman, & Fahle, 1993).

Acquired CP is a form of perceptual learning. It could result from several kinds of perceptual change, including within-category compression and cross-category expansion. Within-category compression (or acquired equivalence) is characterized by a "shrinking" of perceptual space for stimuli that are categorized together. Cross-category expansion results from "stretching" perceptual space spanning the category boundary. Livingston et al. (1998) reported within-category compression resulting from categorization of complex stimuli, whereas Goldstone (1994) found cross-category expansion for simple stimuli varying two-dimensionally. It is also possible that sensitization to differences may not be restricted to local, categorization-relevant regions but can spread to the whole range of stimulus variation (Goldstone, 1994). Such sensitization may also occur dimensionally. Differential attention to a category-relevant dimension may result in acquired equivalence or distinctiveness for entire dimensions (Nosofsky, 1986). In the experiments that follow, we explore whether color discriminability can be improved through training, the kind of perceptual changes underlying acquired CP, and the role of attention in acquired CP.

Dimensions of Color and Color Categories

Color perception varies in many ways, but much of the variation is captured by three dimensions—hue, lightness, and saturation—corresponding approximately to Munsell hue, value, and chroma.

Although color categories are often thought of as primarily hue based, specification of their range requires all three dimensions. For instance, *yellow* is restricted to relatively light regions of color space, and *pink* to desaturated regions. The dimensions of color are predominantly integral or unanalyzable (Burns & Shepp, 1988; Garner, 1974). Thus nonexperts, at least, find it difficult to attend to any one dimension alone or to describe colors precisely in terms of the three dimensions.

Experiments on color CP have usually used stimuli varying only in hue (e.g., Bornstein & Korda, 1984; Roberson et al., 2000). In our experiments, stimuli also varied in lightness. This was done partly to more closely emulate natural categories and partly to investigate what, if anything, was learned about the dimension irrelevant to categorization. Thus, in Experiment 4, stimuli varied in hue and lightness during categorization training, but only one dimension was relevant to categorization. In the subsequent test phase, discrimination was measured on both dimensions. As attention is usually required for perceptual learning to occur, there may be no change in discriminability on the irrelevant dimension. However, as hue and lightness are integral dimensions, variations in the irrelevant dimension may have been detected, and learning might occur.

Linguistic Relativity and CP

We now return to our core argument. If color perception is susceptible to perceptual learning, and if such learning includes producing CP, then language learning should influence where in color space changes occur. For a child to learn to use color terms acceptably, he or she has to discover implicitly what the defining parameters of the categories are. Typically, the child will overextend the reference of a color term but will be guided to gradually restrict its use to an acceptable range. During this process, more attention to boundary regions than category centers will be required to work out where the boundaries are. This differential exposure should enhance discriminability for boundary regions relative to central regions and gradually produce acquired distinctiveness between adjacent categories and possibly acquired equivalence within categories. And, of course, as the number of categories and the location of category boundaries vary across languages (Berlin & Kay, 1969), so should the locations of heightened or reduced sensitivity vary.

We are not suggesting that color categories are entirely learned or that languages segment color space "without constraint." The evidence for CP before language acquisition (Bornstein et al., 1976) and for similarities across languages (Berlin & Kay, 1969) rules out such extreme positions. Similarly, the neuropsychological and psychophysical evidence is consistent with color perception being largely hardwired and largely universal, although the neurological basis of color categories is not yet known (Valberg, 2001). Rather, our suggestion is more modest. It suggests a plausible route through which language, by directing attention to boundary regions, could affect CP. This route may also bridge the gap between whatever is hardwired and the categories one is obliged to learn as a member of a language community. Languages exploit hardwired perceptual discontinuities in their formation of color categories. But other, presumably social, forces drive how many and to some degree which categories are formed. For instance, it is relatively common for languages not to mark the blue–green distinction, despite infants having CP across the blue–green boundary (Bornstein et al., 1976). Moreover, adult speakers of such languages show different patterns of CP from speakers of languages that do mark this distinction (Roberson et al., 2000). And, even in languages with similar category structures, the position of category boundaries can vary (Berlin & Kay, 1969; see, for instance, differences in the purple–blue boundary in English and Russian and in English and Catalan). Finally, note that Bornstein et al. (1976) found CP across primary category boundaries (red–yellow, blue–green, and green–yellow) but did not test secondary category boundaries such as red–pink, red–orange, or blue–purple. And, it is just these secondary categories that are most commonly "missing" from many languages as encapsulated in their low position on Berlin and Kay's (1969) hierarchy.

The foregoing conjecture does not imply that labeling is necessary for CP or for the acquisition of CP. In the latter case, its role is indirect, guiding attention to critical regions. Labeling may be involved in CP, as Roberson and Davidoff (2000) suggested, but it is not necessary. They found that verbal interference in the interval between target and test stimuli eliminated CP. However, the color anomic reported by Roberson et al. (1999) had difficulty with color grouping but still showed CP in a simpler task. Moreover, CP can survive verbal interference, and it occurs in visual search in which decision rates are too fast for labels to be available to influence performance (Pilling, Wiggett, & Davies, 2002).

As things currently stand in the color categorization literature, all of this rests on an assumption: New color (specifically, hue) boundaries can be learned. As summarized earlier, there is considerable evidence that CP can be acquired. However, no direct evidence exists for learned color categories. It is quite possible that color is a "special case" in which all perceptual boundaries are innate and hence the convergence of many languages on quite similar color term repertoires (Berlin & Kay, 1969; Kay & McDaniel, 1978). Finding evidence to the contrary is thus essential to support the arguments outlined in the above paragraphs.

Present Experiments

The four experiments reported here investigated, first, whether color perception improved with training and, second, whether learning new color categories produced CP. Experiment 1 used a same–different task with fast successive presentation of color stimuli varying in hue and lightness. Participants trained for three sessions and were then tested with new stimuli to assess the specificity of learning. In Experiments 2, 3, and 4, participants trained to categorize colors across a novel boundary, and then discrimination was assessed using a same–different task. Acquisition of CP and its time course were assessed, and transfer of learning to new stimulus regions or unattended dimensions was explored. Category training required participants to learn to correctly classify exemplars into two categories. Training included two aspects of natural category learning not usually included in laboratory studies. First, training stimuli were chosen at random from within the continua of each category rather than being drawn from a fixed set. Second, categories were not defined explicitly. Rather, they had to be discovered through trial-and-error with feedback.

Experiment 1: Color Discrimination Learning

This experiment tested whether color discrimination improved with training and assessed whether any learning was restricted to the training stimuli. Participants practiced fast, successive same–different judgments of color over 3 days. On the 4th day, they were tested in a contiguous region of color space to explore transfer effects. Both hue and lightness judgments were performed.

Method

Participants

There were 14 participants (8 men and 6 women); they were all students at the University of Surrey, Surrey, England. Mean age was 23 years, with a range of 18 to 26 years. All had normal color vision as assessed by the City University Color Vision Test (Fletcher, 1980).

Apparatus

A personal computer running experimental software and a Samsung SyncMaster 15gle CRT monitor were used. Monitor calibration and color measurements were made using a Minolta CS-100 chromameter. In addition, buttons on a PC game-pad were used for responses.

Stimuli

Stimuli were drawn from two sets of 16 emulated Munsell colors displayed on the monitor; one set was blue and the other green. Each set formed a 4 × 4 matrix, varying on the hue and lightness dimensions. Saturation was kept constant at Munsell Chroma 6 for both sets. The four lightness levels were Munsell Value 5, 6, 7, and 8 for both sets. Hue levels were Munsell Hue 2.5B, 5B, 7.5B, and 10B for the blue set, and 2.5G, 5G, 7.5G, and 10G for the green set. Stimuli were displayed as 5 cm × 5 cm squares on a black background, whose centers fell on the equator of the display horizontally and 5 cm away either to the right or to the left of the central meridian vertically (total separation = 10 cm).

Procedure

Participants were allocated randomly into two training groups: the blue group ($n = 8$) and the green group ($n = 6$). Each group trained on their color set on 3 consecutive days and were then tested with the other set on the 4th day. The task was a same–different judgment task that required participants to decide whether pairs of color presented successively were identical or not. The first stimulus was displayed for 500 ms, followed by a blank screen interstimulus interval (ISI) for 500 ms.[1] The second stimulus was then displayed until the response was made. The interval between trials was 1,500 ms. One stimulus was displayed on the left-hand portion, whereas the other was displayed on the right-hand portion of the screen as described earlier. Screen locations were randomized. Responses were made by pressing appropriate buttons on the game-pad.

When two stimuli of a given same–different pair differed in lightness, hue was constant, and when they differed in hue, lightness was constant. The size of the difference was either one hue level (2.5 Munsell units) or one lightness level (1 value unit). There were 12 such pairs for both dimensions. For same trials, each of the 16 stimuli in a set was used an equal number of times. In each session, all possible pairs were repeated 10 times, resulting in 400 trials (10 × [12 + 12 + 16]). The order of presentation within and across trials was randomized. A short pause was allowed after each block of 50 trials.

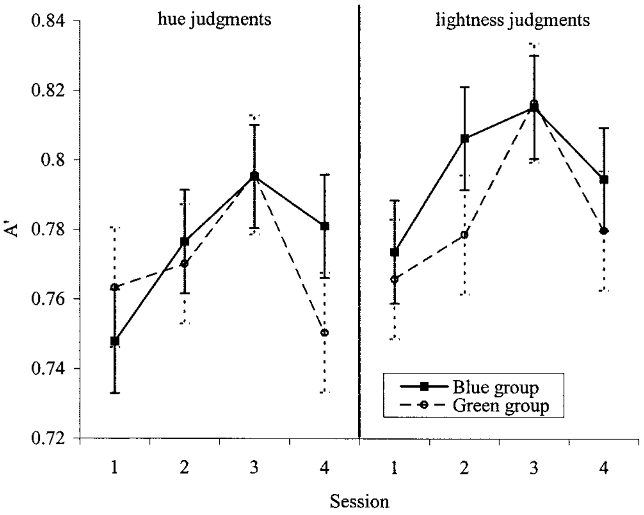

Figure 1. Mean A' across four sessions on same–different judgments of hue and lightness in Experiment 1. Note that the stimuli changed on the fourth session (blue to green; green to blue). Error bars represent ± 95% within-subject confidence intervals (see Footnote 2).

Results

Accuracy was assessed by combining hits and false alarms to produce Grier's index, A' (Grier, 1971). This nonparametric measure was preferred to the usual (parametric) d' because it makes fewer assumptions about the distribution of the data. Figure 1 shows mean A' scores across the 4 days on hue and lightness judgments for the two training groups.[2] Accuracy on both dimensions increased on the first 3 days. An analysis of variance (ANOVA) with two within-subjects variables of *day* (Days 1 to 3) and *dimension* (hue or lightness) and a between-subjects variable of *group* (blue or green) supported this observation. The main effect of day was statistically significant, $F(2, 24) = 6.93$, $MSE = 0.0018$, $p < .01$, $\eta^2 = 0.37$, whereas no other main effects or interactions were.

It can also be seen in Figure 1 that when the stimulus set was changed to a neighboring range on the final day, there was a drop in accuracy. An ANOVA with the same variables as above but only comparing Day 3 and Day 4 revealed a significant main effect of day, $F(1, 12) = 4.87$, $MSE = 0.0024$, $p < .05$, $\eta^2 = 0.29$.

---

[1] The ISI for Experiment 1 was based on Bornstein and Korda (1984) and Goldstone (1994). In the remaining experiments we used a longer ISI, to allow the possibility of interpolating interference tasks in the ISI. This was not done in the present experiments but was in comparable experiments that we report elsewhere (Pilling et al., 2002).

[2] Within-subject and between-subject confidence intervals (CIs) presented in this article were computed in the way described by Loftus and Masson (1994) and Loftus (2002). For within-subject CIs in our multifactor mixed designs, we use the error term relating to the highest number of interactions (i.e., for a single between-subjects and two within-subject factor design, we used the error term for the three-way interaction). Between-subjects CIs are shown in our figures (see later) around a single floating point. However, in some cases, more than one interval can be computed. In such cases, we used the most conservative (largest) between-subject CI.

Comparison of Day 1 and Day 4 performances supported this lack of transfer; accuracy was equivalent for the first and the final day performances ($F < 1$).

Discussion

Discriminability of hue and lightness improved across the training period. This improvement did not seem to result from high-level learning such as an awareness of the difficulty levels involved or stimulus range; no transfer to a novel stimulus set was observed.

Having established that color discriminability could be changed, the next experiment investigated whether such improvement can be channeled differentially for within- and cross-category discrimination, simulating CP effects observed for linguistic categories.

Experiment 2: Category Learning and Acquired CP

Participants were trained to divide a basic color category (blue or green) into two new categories. After training, same–different judgments were used to see if discriminability had been enhanced, as in Experiment 1, and whether CP had been acquired across the new boundary. Acquiring CP required reversing the perceptual structure of the old categories: Discriminability peaks in boundary regions and is lowest in the focal regions of perceptual categories (Harnad, 1987; Pastore, 1987; Rosch, 1975). In this experiment the new boundary lies on the focal point of an existing category (blue or green). Therefore, in the test phase, across/new-category[3] judgments were made on pairs from the focal area of each existing category, whereas within/new-category pairs were from closer to the preexisting category boundaries (e.g., yellowish-greens or purplish-blues).

One group of participants trained on *green* and one on *blue*. In addition, two control groups received no training but underwent the same–different judgments. One control group was able to view the test stimuli and some example same–different pairs before the same–different task, whereas another group was not (see *Procedure*). Comparison of these two control groups provided a measure of the extent to which changes in the experimental groups were due to general familiarity with the stimuli rather than category training per se. In addition, they provided a way of assessing whether the perceptual structure of the preexisting categories was as we suggested earlier. The two training groups were also tested on both color regions. Comparisons of their performances on their trained and untrained regions allowed assessment of whether learning was restricted to the training stimuli or whether it generalized to some degree to a neighboring region.

Comparisons between category learners (CLs) and untrained controls, as well as those between two groups of participants who categorized in different stimulus regions, are reported to explore the specific issues outline. These include overall improvement on same–different judgments of hue as a result of categorization training, transfer of learning to a different stimulus region, and induced CP effects in the form of a differential improvement for pairs of hues straddling a category boundary. Induced category effects were explored further to clarify whether compression or expansion effects were responsible for these changes.

Method

Participants

There were 40 participants (14 men and 26 women). They were mostly psychology undergraduates and were paid a fee or course credits for taking part. Their age ranged from 18 to 41 years, with a mean age of approximately 22 years. All of the participants had normal color vision as assessed by the City University Color Vision Test (Fletcher, 1980).

Stimuli and Apparatus

A personal computer running experimental software and a 15-inch Sony Trinitron 100fs monitor were used. The Color Science Library 2.0 supplied by CGSD (Computer Graphics Systems Development Corporation) was used to reproduce colors on the monitor. Stimuli in both training and test phases of the experiment were colored squares measuring 5 cm × 5 cm and were displayed at the center of the computer screen, against a background of neutral gray.

Stimuli in the training phase were generated randomly in two regions of color space: blue and green. Figure 2 shows two training areas in Munsell space, in which hue and lightness vary along the *x*- and *y*-axes, respectively. Hue boundaries, indicated by a vertical dashed line in each area (enclosed by a square), fell roughly in the center of the linguistic categories occupying each region (7.5B and 7.5G for the blue and green regions, respectively). The maximum possible chroma that could be realized for each training region was used. This was 7.9 and 7.7 for green and blue, respectively. Stimuli within 0.2 Munsell hue units of the boundary were avoided. On each trial in training, a random point within the relevant (blue for blue group, green for green group) training area was selected, and the corresponding color was displayed on the monitor.

Stimuli for the same–different task were drawn from each training region to form a 4 × 4 matrix of 16 colors (see the diamond-shaped markers in Figure 2). Hue varied between 3.75G and 1.25 BG for green and between 3.75B and 1.25PB for blue in steps of 2.5. Value varied between 5.25 and 6.75 in steps of 0.5 for both regions. Chroma was constant at the same value as for the training stimuli (7.9 and 7.7 for green and blue). Stimuli adjacent to the boundary were 1.25 hue steps away from it on either side.

Procedure

Participants were allocated into one of four groups at random ($n = 10$): blue, green, Control 1, and Control 2. Participants in the two experimental groups (blue and green) performed categorization training in three daily sessions only in their assigned training region (blue or green). They then completed a same–different task, in both regions, immediately after the last training session. Controls did not do the training task but completed the same–different task, also in both regions; Control 2 participants were shown the matrices of 16 stimuli before same–different judgments.

Training phase. During training, participants in the blue and green groups learned to categorize along the hue dimension around the boundary lines described above in *Stimuli and Apparatus*. Training consisted of two stages. In the first, *context training* phase, participants made category judgments while being able to see examples of stimuli they had previously categorized correctly. A randomly chosen color was presented in the center of the screen, together with 16 “slots” to be filled with the incoming colors, divided into two groups: eight slots (two columns, four rows) on each side

---

[3] In referring to the stimuli in the present study, we use the terms *within/new-category* and *across/new-category* to indicate stimulus differences falling within or across our novel, learned categories (e.g., Blue 1 and Blue 2) and not to preexisting, “natural” categories (e.g., blue and green).

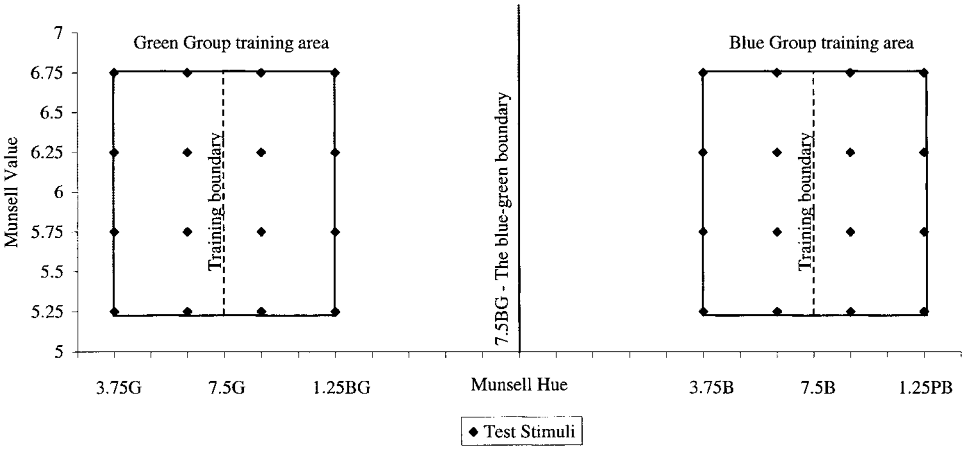

Figure 2. Random color generation areas for training in the blue and green regions in Experiment 2. Hue and lightness vary along the *x*- and *y*-axes, respectively. Chroma was constant at 7.9 for the green and 7.7 for the blue training area. Each training area is shown by a square, in which the dashed line represents the respective hue boundary that participants learned during training. Training stimuli fell on randomly selected points within an area, avoiding points very close to the boundary (within 0.2 hue units). The solid vertical line at the center of the figure shows the "natural" blue–green boundary (Bornstein & Korda, 1984). The diamond-shaped markers show the fixed set of stimuli used in same–different judgments that followed categorization training.

of the test stimulus. The random test color could be at any point within a given training area, as described earlier and shown in Figure 2, and thus could fall on either side of the training boundary (avoiding points very close to it). (Note that stimuli varied on hue and lightness but the training-relevant variation was the former.) The participants sorted the test stimuli into two groups around this boundary by "dragging and dropping" the color in the center into an empty slot on the left or right, using the computer mouse. The judgment they had to make was binary: left-hand side (Category 1) or right-hand side (Category 2). The specific slot they filled had no importance, as long as they got this judgment right.

Participants received immediate feedback. If a response was correct, the color remained in the slot; if the response was incorrect, then the color disappeared accompanied by a sound indicating "incorrect." When they correctly filled all 16 slots, a "set" was complete and a new one began until the criteria for successful category learning were met.

The participant could place the first color in a set on either side of the screen. Once it was placed on the left or right side, colors from the same experimental category (those from the same side of the relevant training boundary) as the first color had to be placed on the same side. Similarly, colors from the other experimental category (other side of the boundary line) had to be placed on the other side of the screen. (That is, it was up to the participant which side of the display Categories 1 and 2 were assigned to.) Participants were not given specific information on what the categorization was based on, and they had to learn from the feedback to complete the training phase. As the correct responses accumulated in each set, the number of colors for comparison increased (up to the maximum number of slots available: 16).

The instructions given to participants at this stage were as follows:

> You will see a range of colors which you should sort into two groups; it is your task to discover and successfully learn the rule that separates these two groups. On every trial, you will see a color in the center of the screen. Using the mouse, place this color in an empty slot on the left-hand side if you think it belongs to one group, and right-hand side for the other group. It does not matter which slot you place it in within the left- or right-hand side groups of slots. You complete a *set* when you correctly fill all slots, and start a new one until you get sufficiently good at the task to finish. At the beginning of each set, when all slots are empty, you can place the first color on either side. From the next color onward, the question is, "Does this go with the first color, or does it belong to the opposite group?" If your answer is correct, the color stays where you put it. If you are incorrect, the color disappears and you hear a sound.

The criteria for completing the first stage of training were that at least 10 sets were completed, there were at least 3 error-free sets, and this stage lasted at least 20 min. This first stage of training typically took about 30 min.

In the second stage of training, single test colors appeared in the center of the screen, and the participant had to decide which of the two categories it belonged to by responding "left" or "right" with the mouse buttons. Incorrect choices were signaled by the word "incorrect" replacing the test color, accompanied by a sound. The criteria for completing this stage were completion of at least 50 trials and 25 consecutive correct responses. This stage typically took about 5 min.

The instructions given to the participants in the second stage were as follows:

> The task here is the same as before. But this time you see only one color at a time. You now indicate which group it belongs to by pressing the mouse button: left button for one group, right button for the other (as in the first part, it does not matter which button you press for the first color). If your choice is incorrect you will see the word "incorrect" in place of the color and hear a sound. You will be able to finish this part of the experiment when you get sufficiently good at the task.

The rationale behind this two-stage training regime is as follows. Typically, a category-learning experiment requires a participant to make an absolute judgment (Category 1 vs. Category 2) on a single test stimulus. We argue, however, that in real life, learning new categories takes place in this isolated way ("Is this color blue or green?") as well as within a context ("Does this go with these or those?"). This contextual nature of the category judgment makes it a more discrimination-based one. Importantly, however, the judgment is still absolute, or binary: One of two possible answers is correct. In an attempt to emulate such a process, we started our participants off with the context-training stage. The second stage, however, was used to assess whether training had been successful in conventional category-learning terms as well as to reinforce it, hence the considerably shorter duration of the second stage. Thus, we make no indirect claims about the significance of each type of learning in real life by our choice of the weights of each of our stages.

Discrimination test phase. All of the participants did the same–different judgment task for both color sets. Pairs of colors were displayed successively in the center of the computer screen, with 5 s between stimuli, 1 s between trials, and the duration of 1 s of the first stimulus. The second stimulus was displayed until a response was made, using the left mouse button for *same* and the right mouse button for *different*.

For different trials, adjacent hue pairs of the same value were used. There were 128 trials for each color region: 32 same trials and 96 different trials; each of the 16 possible same pairs was repeated twice, whereas each of the 12 possible different pairs was repeated eight times. There were thus a total of 256 trials in the discrimination task. The order of presentation of the trials was randomized.

Participants in the two training conditions (blue and green) completed this phase immediately after their last training session on Day 3. Participants in the control groups only completed this phase of the experiment. Those in Control 1 were shown the test phase stimuli in the 4 × 4 matrix layout as described in Figure 2 (diamond markers) and were given examples of *same* and *different* pairs. Participants in the Control 2 group did not receive exposure to the stimulus sets. All participants did 20 practice trials with randomly selected experimental stimuli (equally from both sets) before beginning the task. The discrimination task typically took about 40 min.

Results

During the training phase, participants performed between about 300 and 400 context-training trials on each day and about 50 to 150 single-color trials. The number of single-color trials reduced somewhat by the 3rd day, but there was no reliable reduction in the number of context-training trials.

The dependent variable in the analyses below was Grier's (1971) A' index of accuracy. All analyses are mixed-design ANOVA with one or both of two within-subjects variables: *judgment region* (blue or green) and *categorical relationship* (within/new-category or across/new-category). The between-subjects variable(s) varied across analyses. Tukey's honestly significance difference was used for post hoc tests. Exhaustive reporting of the results would be extensive because of the complexity of the design. Rather than do so, we concentrate on addressing the questions raised in the introduction and report only the relevant parts of the ANOVAs.[4] The overall pattern of results was consistent across analyses and converges coherently.

Baseline Performance

There were no significant differences between the two control groups ($F < 1$) and no interactions involving them. The two control groups were therefore combined for subsequent analyses.

Figure 3 shows the control groups' scores in both regions and the trained region scores of the CLs (left panel, blue group in blue; right panel, green group in green). It can be seen that there were two main patterns evident in the controls' performance. First, unexpectedly, judgments for the blue region were more accurate than those for the green region, $F(1, 18) = 8.80$, $MSE = 0.012$, $p < .01$, $\eta^2 = 0.33$. Second, there seems to be an effect of the preexisting category structure for both regions: Within/new-category pairs were easier to discriminate than across/new-category pairs, $F(1, 18) = 15.6$, $MSE = 0.0039$, $p < .01$, $\eta^2 = 0.46$. This was also partly the case for the training groups for their respective untrained regions. Figure 4 plots same–different performances by both training groups across the two regions. It can be seen that within/new-category scores tended to be higher than across/new-category scores for the green group in the blue region and for the blue group in the green region. An ANOVA on just the untrained region judgments of CLs revealed a significant main effect of categorical description: Prior to training, participants were better at our within/new-category pairs than those that straddled our experimental boundary, $F(1, 18) = 19.15$, $MSE = 0.006$, $p < .0005$, $\eta^2 = 0.52$.

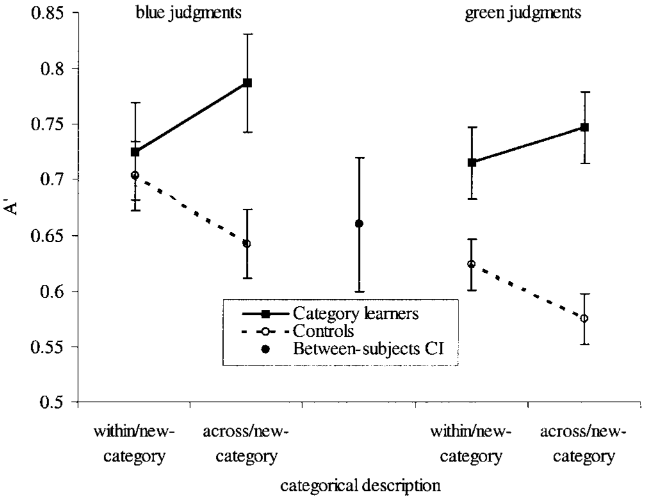

Figure 3. Mean within/new-category and across/new-category A' for category learners and controls in two regions in Experiment 2. Performance by category learners in only the trained region is shown. For each training group, $n = 10$; for controls, $n = 20$. Error bars represent ± 95% within-subject confidence intervals (CIs). The free floating point and its error bars represent ± 95% between-subjects CI.

Overall Improvement

It can be seen in Figure 4 that accuracy was higher for the trained regions than for the untrained regions. The blue group was better at blue region judgments than the green group, whereas the reverse was true for green region judgments. This was supported by a significant interaction between training region and judgment region, $F(1, 18) = 26.49$, $MSE = 0.015$, $p < .0005$, $\eta^2 = 0.60$.

Comparison of the training groups with the controls also indicates that region-specific overall improvement occurred. It can be seen in Figure 3 that for the blue region, the blue group is better overall than controls, and for the green region, the green group has higher scores than controls. These differences between the training groups and controls were significant for both blue: main effect, $F(1, 28) = 5.01$, $MSE = 0.019$, $p < .05$, $\eta^2 = 0.15$; and green: main effect, $F(1, 28) = 15.85$, $MSE = 0.015$, $p < .0005$, $\eta^2 = 0.36$.

---

[4] More complete analyses are available from Emre Özgen.

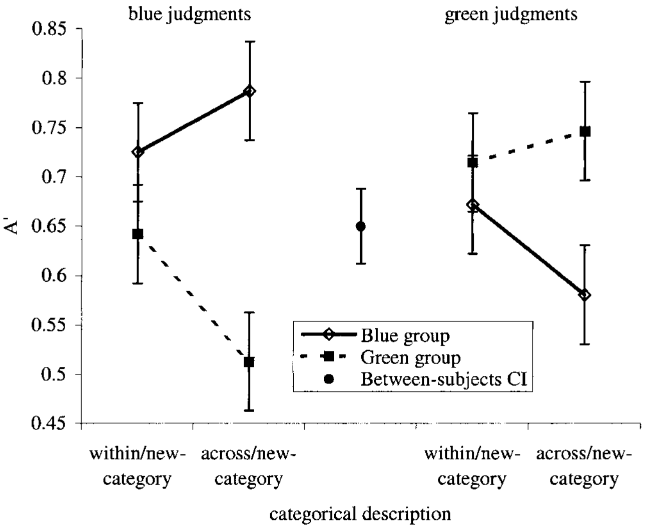

Figure 4. Mean discrimination test scores for the blue and green groups in two regions in Experiment 2. For each group, $n = 10$. Error bars represent ± 95% within-subject confidence intervals (CIs). The free floating point and its error bars represent ± 95% between-subjects CI.

Transfer of Improvement

Mean scores for controls and untrained region scores for the CLs were compared to test whether any transfer of improvement occurred. The blue group performed at about the same level as controls in the green region ($F < 1$); the green group was less accurate than controls on blue, $F(1, 28) = 6.80$, $MSE = 0.018$, $p < .05$, $\eta^2 = 0.2$. No positive transfer was observed, and in the case of the green group, some negative transfer was found.

Induced CP Effects

It can be seen in Figure 4 that both training groups scored higher on across/new-category than on within/new-category discrimination for their respective training regions, whereas the reverse was the case for the untrained region, $F(1, 18) = 21.43$, $MSE = 0.0058$, $p < .0005$, $\eta^2 = 0.54$. In addition, the interactions between training group and categorical relationship were significant when the two regions were analyzed separately—blue region, $F(1, 18) = 21.43$, $MSE = 0.0043$, $p < .0005$, $\eta^2 = 0.54$; green region, $F(1, 18) = 7.86$, $MSE = 0.0048$, $p < .0005$, $\eta^2 = 0.3$—and across/new-category judgments were significantly better than within/new-category judgments when just the trained regions were analyzed, $F(1, 18) = 8.05$, $MSE = 0.0028$, $p < .05$, $\eta^2 = 0.31$. Post hoc tests showed that the blue group was better than the green group at blue region across/new-category judgments ($p < .0005$), and the green group was better than the blue group at green region across/new-category judgments ($p < .0005$).

The above pattern was supported by comparisons between training groups and controls. Across/new-category judgments were better than within/new-category for the training groups in the trained region, whereas the opposite was true for controls (see Figure 3). This trend was supported by a significant two-way interaction of group (controls vs. training) and categorical relationship for both regions: blue region, $F(1, 28) = 11.06$, $MSE = 0.0046$, $p < .005$, $\eta^2 = 0.28$; green region, $F(1, 28) = 8.72$, $MSE = 0.0025$, $p < .01$, $\eta^2 = 0.24$.

Cross-Category Expansion or Within-Category Compression?

Although there were differences between the training groups on across/new-category judgments (see above), there were no significant differences between the groups on within/new-category scores in either region ($p = .28$ for blue, $p = .9$ for green). Comparison with the controls showed that the training groups' within/new-category scores were at least no worse than controls for both regions, for both training groups (minimum $p = .6$). These findings show that there was across/new-category expansion but not within/new-category compression.

Discussion

It was predicted that the preexisting category structure should result in greater accuracy for within/new-category judgments than for across/new-category judgments. This was supported. Within/new-category pairs (from areas further toward the boundary of preexisting categories) were easier to discriminate than across/new-category pairs (focal regions of preexisting categories). This effect was so for all participants without training in the region.

The main finding of the present experiment is that CLs reversed this pattern. Their across/new-category performance in the trained region tended to be more accurate than their within/new-category performance, unlike controls and unlike the training group that trained in the alternate region. Combined analysis of the trained region judgments of CLs showed that across/new-category pairs were judged more accurately than within/new-category pairs. It appears that participants can "unlearn" the effects of preexisting categories while learning the novel categories.[5] Thus the effects of linguistic categories on discrimination and memory (Bornstein & Korda, 1984; Boynton et al., 1989) may not be immutable but may be subject to tuning.

Consistent evidence was found for cross-category expansion (acquired distinctiveness). In none of the comparisons was there evidence for desensitized within/new-category judgments as a result of category learning, whereas across/new-category judgments were found to be reliably superior for training-relevant region performances of CLs.

Participants also improved overall on judgments in the region that they trained to categorize. Those who trained in a given region were more accurate overall on judgments in that region than their own judgments in the adjacent region and than those of the group who trained in the adjacent region. They also judged differences more accurately in the familiar region than control participants. Post hoc analyses of comparisons with controls revealed somewhat

---

[5] It is important to note that our use of the term *unlearning* or *reversal* does not refer to a reduction in (otherwise higher) sensitivity around an existing boundary (e.g., blue–green). Rather, it means an enhancement in sensitivity in the focal areas of an existing category, which is "naturally" low in comparison with less focal areas (as shown by the pretraining superiority of within/new-category over across/new-category judgments). All subsequent uses of these terms in the present study should be understood in this way.

mixed findings on the issue of overall sensitization. In the green region, CLs (the green group) improved on both within/new-category and across/new-category judgments compared with controls. However, improvement tended to be restricted to across/new-category judgments in the blue region, although the main effect of group (blue group vs. controls) was significant. On balance, this suggests that categorization can lead to sensitization over an entire dimension (Nosofsky, 1986). Consistent with this, Özgen (2000) found that speakers of Turkish, which encodes the blue region with two terms, judged distances in similarity space to be greater overall in this region than did English speakers.

Finally, there was no evidence of transfer of learning to a novel stimulus region. Overall improvement and induced CP were restricted to the training region. Untrained region judgments of category learners were no different from the controls in the untrained region, and they showed the same effect of preexisting category structure. It seems that learning was stimulus specific and not merely task related. However, it is possible that some transfer would have occurred if the transfer set was closer to the training set than that used here. For instance, if the new set had the same hues as the training set but differed in saturation, transfer may have occurred.

Experiment 3: Speed of Acquisition of CP

Experiment 3 sought to establish the time course of the acquisition of CP. The previous experiment showed that three training sessions of about 1,500 categorization trials in total was sufficient to establish CP but did not assess when it was first detectable. The purpose of the present experiment was mainly practical. If CP could be established with significantly less than 1,500 training trials, the nature of CP could be explored without having to spend as many participant hours.

In Experiment 3, the same training method as in the previous experiment was used over two sessions, and a test phase was added before and after each training session. Tests of discrimination improvement and CP effects were carried out as in Experiment 2. Training and testing used the green stimulus set, and no test of transfer was made. Improvement was assessed by comparison of the training group with a control group that did only the test phase.

Method

The method was based on the previous experiment using the same apparatus and a subset of the stimuli (the green set) and is only described in outline. Sixteen participants were allocated randomly into two groups: CLs and controls. On each of 2 consecutive days, CLs did the same training as the green group in Experiment 2. They also did the same–different task for the green region, immediately before and after training. Controls did the discrimination task once following familiarization with the stimuli as for Control 1 in the previous experiment.

Results

Table 1 shows mean within/new-category and across/new-category A′ scores in pre- and posttraining discrimination test phases across 2 days, along with control participants' performance on the same task. It can be seen that for CLs, on the 1st day, within/new-category and across/new-category scores were similar in both pre- and posttraining test phases, whereas on the 2nd day, across/new-category scores tended to be higher than within/new-category scores.

Day 1

For Day 1 pretraining performance, in which both CLs and controls were equally unfamiliar with the discrimination test task, there were no significant main effects or interactions (all *F*s < 2, all *p*s $> .18$). However, the CLs' posttraining performance was almost significantly more accurate than controls: main effect $F(1, 14) = 3.14$, $MSE = 0.014$, $p = .098$, $\eta^2 = 0.18$; all other *F*s < 1.7, *p*s $> .22$.

Day 2: Pretraining

On Day 2, CP effects were observed for CLs even in the pretraining phase. Across/new-category judgments were more accurate than within/new-category judgments for CLs but not for controls: interaction $F(1, 14) = 14.78$, $MSE = 0.0018$, $p < .005$, $\eta^2 = 0.51$. Post hoc tests showed that CLs' across/new-category judgments were more accurate than their own within/new-category judgments, and more accurate than within/new-category and across/new-category judgments by controls ($p < .01$ for all comparisons). In addition, CLs were more accurate than controls overall: main effect, $F(1, 14) = 5.51$, $MSE = 0.0094$, $p < .05$, $\eta^2 = 0.28$. The emergence of these effects before training on Day 2 as opposed to after training on Day 1 should yield a significant interaction between categorical description (within/across) and session (Day 1–posttraining/Day2–pretraining). Although the observed pattern was in the right direction, the inter-

Table 1
Mean Within/New-Category and Across/New-Category A′ on Same–Different Judgments for Category Learners Across 2 Days and for Controls in Experiment 3

	Day 1		Day 2	
Group and categorical description	Pretraining	Posttraining	Pretraining	Posttraining
Category learners				
Within	0.69 (0.04)	0.70 (0.03)	0.68 (0.02)	0.71 (0.05)
Across	0.69 (0.03)	0.73 (0.05)	0.76 (0.03)	0.75 (0.05)
Controls				
Within	0.66 (0.03)			
Across	0.62 (0.02)			

Note. For each group, $n = 8$. Numbers in parentheses show ±1 standard error.

action failed to reach significance, $F(1, 7) = 2.16$, $MSE = 0.003$, $p = .19$, $\eta^2 = 0.24$.

Day 2: Posttraining

CLs still showed CP: interaction $F(1, 14) = 5.50$, $MSE = 0.002$, $p < .05$, $\eta^2 = 0.28$; although overall the groups did not differ: main effect $F(1, 14) = 2.69$, $MSE = 0.023$, $p = .12$, $\eta^2 = 0.16$. Despite the latter finding, post hoc tests revealed that CLs' across/new-category judgments were better than controls for both across/new-category and within/new-category judgments, and CLs' within/new-category judgments were better than controls' across/new-category judgments ($p < .01$ for all cases).

Discussion

Category effects seemed to have emerged after just 1 day (around 500 training trials). Induced CP effects were evident at the start of the 2nd day, and again at the end of it. However, they were not found at the end of the 1st day, perhaps due to fatigue masking the effects. On the other hand, consolidation of learning a perceptual skill can take place during sleep (Karni & Sagi, 1993; Karni, Tanne, Rubenstein, Askenasy, & Sagi, 1994). Karni et al. (1994) found that training on a visual discrimination task leads to improvement on the day following training provided REM sleep periods were not disrupted. Fatigue effects can also account for the disappearance of the overall superiority of the CL performance in comparison with the controls.

In the light of these results, we used a shorter training period in the next study than that in Experiment 2. An intensive training regime was used on the 1st day, followed by a shorter refresher session on the 2nd day, followed by the test phase.

Experiment 4: Induced CP Effects on Hue and Lightness

Linguistic color categories encompass a volume of color space varying in hue, lightness, and saturation. For some categories, a restricted lightness range is a key property. For example, *yellow* denotes only light colors. Similarly, the distinction between the two *blue* terms in Russian (Davies & Corbett, 1994) and Turkish (Özgen & Davies, 1998) is primarily based on lightness. In fact, the only basic color terms in English that have examples across all lightness levels are *blue* and *green* (Boynton & Olson, 1987). One aim of this experiment was to see whether CP could be learned across a lightness boundary and whether it took the same form as learned CP for a hue boundary (Experiment 2)

In Experiment 2, there was no transfer of learning to a neighboring region of color space. In that case, participants had not been exposed to the transfer stimuli during the training phase. They had, however, been exposed to variations in lightness, but these were not relevant to categorization. We did not, however, assess what if anything had been learned about the irrelevant lightness dimension. One aim of this experiment was to see if anything was learned about the dimension irrelevant to classification. The training stimuli varied in hue and lightness, and one group learned to categorize on the basis of a hue boundary, whereas a second group learned to categorize across a lightness boundary. Lightness and hue discrimination were then measured for both groups in the training region of color space.

In perceptual learning experiments, usually little is learned about an unattended variation (e.g., Ahissar & Hochstein, 1993; Shiu & Pashler, 1992). However, if perceptual learning is induced by categorization, then it is possible that the discriminability of an irrelevant dimension could get worse. As Goldstone (1994) argued, CP could be based on either acquired distinctiveness of the relevant dimension or acquired equivalence of the irrelevant dimension. In the present case, hue categorizers' lightness judgments may become less accurate after training and, conversely, lightness categorizers' hue judgments would become less accurate. On the other hand, as hue and lightness are integral (Burns & Shepp, 1988; Shepard, 1964), then it may not be possible to ignore the irrelevant dimension. In that case, dimensional sensitization might occur for both dimensions.

In the experiments so far, we have used the Munsell metric as our measure of perceptual distance. However, although the Munsell system is intended to be perceptually uniform, we found in Experiment 2 that discrimination of pairs close to the boundary of a preexisting category was more accurate than equally spaced pairs from the focal region of the category. An alternative measure of perceptual distance is the Euclidean distance in the CIE L* u* v* perceptually uniform color space (Hunt, 1987). (*Note*. CIE stands for Comission Internationale de l'Eclairage [International Commission on Illumination].) Table 2 shows these distances for each stimulus pair used in the test phase of Experiment 2. It can be seen that CIE distances were smaller for across/new-category pairs than for the two within/new-category pairs of the same lightness

Table 2
Euclidean Distances in the Perceptually Uniform CIE L u* v* Space for Hue Pairs Used in Experiment 2*

Stimulus pair	Categorical description in Experiment 2	CIE distance (L* u* v*)
3.75B–6.25B at Value = 5.25	Within	7.44
6.25B–8.75B at Value = 5.25	Across	6.15
8.75B–1.25PB at Value = 5.25	Within	6.33
3.75B–6.25B at Value = 5.75	Within	7.71
6.25B–8.75B at Value = 5.75	Across	6.45
8.75B–1.25PB at Value = 5.75	Within	6.65
3.75B–6.25B at Value = 6.25	Within	7.90
6.25B–8.75B at Value = 6.25	Across	6.69
8.75B–1.25PB at Value = 6.25	Within	6.94
3.75B–6.25B at Value = 6.75	Within	8.00
6.25B–8.75B at Value = 6.75	Across	6.87
8.75B–1.25PB at Value = 6.75	Within	7.25
3.75G–6.25G at Value = 5.25	Within	8.29
6.25G–8.75G at Value = 5.25	Across	6.63
8.75G–1.25BG at Value = 5.25	Within	6.79
3.75G–6.25G at Value = 5.75	Within	8.69
6.25G–8.75G at Value = 5.75	Across	6.72
8.75G–1.25BG at Value = 5.75	Within	6.85
3.75G–6.25G at Value = 6.25	Within	9.03
6.25G–8.75G at Value = 6.25	Across	6.81
8.75G–1.25BG at Value = 6.25	Within	6.85
3.75G–6.25G at Value = 6.75	Within	9.31
6.25G–8.75G at Value = 6.75	Across	6.89
8.75G–1.25BG at Value = 6.75	Within	6.82

Note. Each hue pair has a Munsell hue difference of 2.5. The top 12 rows show the blue region pairs, whereas the rest are in the green region. It can be seen that across/new-category distances are always smaller than within/new-category distances.

(value). The mean CIE distances were 6.54 and 7.28 for blue across/new-category and within/new-category pairs, respectively, and 6.76 and 7.83 for green across/new-category and within/new-category pairs. According to the CIE metric, the Munsell system has built-in CP. Equivalent Munsell steps close to borders are smaller in CIE than at the center of categories.

In the present experiment we used CIE distance as our measure of perceptual distance to try to eliminate the effect of the preexisting category structure on learning new categories. Pilot work suggested that the use of equal CIE L* u* v* distances between pairs of hue produces equivalent discriminability for the focal and boundary areas of existing categories. The present experiment therefore used a new stimulus set for the test phase in the green region constructed using CIE distances (see Table 3). Pairs from the same or different experimental hue categories are thus likely to be judged equivalently by controls.

Comparisons between CLs and untrained controls, as well as between participants who categorized on different dimensions (hue vs. lightness), were made to investigate the specific issues outlined earlier. These include overall improvement on judgments along the trained dimension, transfer of improvement to a separate stimulus dimension, and induced CP effects in the form of a differential improvement for pairs of hue or lightness straddling the experimental category boundary. This final issue is further explored to assess whether compression or expansion effects are responsible for the changes.

Method

Participants

Thirty participants with an age range of 18 to 41 years and a mean age of 21 years took part. They were mostly psychology undergraduates at the University of Surrey and received course credits for their participation. All of the participants had normal color vision as assessed by the City University Color Vision Test (Fletcher, 1980).

Table 3
Matrix of 16 Stimuli Used in the Test Phase of Experiment 4

Lightness (value)	Hue			
Munsell	4.48G	6.40G	8.65G	0.9BG
5.40				
CIE L*	55.62	55.62	55.62	55.62
CIE u*	−43.50	−45.25	−46.62	−47.69
CIE v*	27.27	21.66	15.83	9.84
5.80				
CIE L*	59.67	59.67	59.67	59.67
CIE u*	−44.16	−45.96	−47.40	−48.62
CIE v*	28.17	22.41	16.52	10.52
6.20				
CIE L*	63.68	63.68	63.68	63.68
CIE u*	−44.70	−46.56	−48.07	−49.37
CIE v*	29.00	23.10	17.17	11.17
6.60				
CIE L*	67.64	67.64	67.64	67.64
CIE u*	−45.14	−47.07	−48.64	−49.97
CIE v*	29.75	23.74	17.76	11.78

Note. Hue and lightness varied uniformly as described in the text. Saturation was constant across the set at Munsell Chroma 7.9. Each cell represents the color for a given hue and lightness level and shows its CIE uniform chromaticity coordinates. The category boundaries introduced in the training phase fell between hue and lightness Levels 2 and 3 (7.5G for hue and value = 6 for lightness).

Apparatus and Stimuli

The apparatus was the same as for Experiment 2. The same range of green stimuli as in Experiment 2 were used for the training phase, and stimulus generation followed the same rules with the exception that here there was also a lightness boundary. One training group classified stimuli using the hue boundary with lightness irrelevant, as in Experiment 2, whereas the other group did the reverse. The lightness boundary was at Munsell Value 6 and lightness varied between Value 5.25 and 6.75 (a range of 0.75 on either side of the boundary[6]). The range of lightness values avoided spanned 0.04 Munsell value points on the two sides of the boundary (between Munsell Value 5.98 and 6.02).

For the discrimination test phase, we constructed a 4 × 4 matrix of stimuli following the same principles as in Experiment 2, except perceptual distances between hues were equated at 6 CIE distance units, equivalent to about two Munsell hue steps. Pairs either side of the boundary were three units from the boundary. Pairs varying in lightness were 0.4 value steps apart. Table 3 shows CIE L* u* v* coordinates and the corresponding Munsell codes for the stimulus matrix. Horizontal and vertical axes represent hue and lightness dimensions, respectively. The boundaries introduced in training fell on the midway point between the second and third columns for hue (7.5G) and rows (value = 6) for lightness. Thus, stimuli in columns 1 and 2 belonged to a different hue category than those in columns 3 and 4; similarly, stimuli in rows 1 and 2 were from a different lightness category than those in rows 3 and 4.

Procedure

Participants were allocated randomly into three groups of 10: hue categorizers (HCs), lightness categorizers (LCs), and controls. Participants in the two training groups (HCs and LCs) did categorization training for two sessions, followed by the discrimination task immediately after the last training session. Controls only did the discrimination task.

Training phase. The training task was the same as that used in Experiment 2. HCs and LCs both trained to categorize randomly generated colors from the green region into two groups of colors. Categories were defined by the hue boundary line for HCs and by the lightness boundary line for LCs. On Day 1, participants in the HC and LC groups completed the intensive training phase lasting about 50 min, consisting of a session of training as in Experiment 2, a 5-min rest, and then a further session of training. On Day 2, participants did refresher training, in which they repeated the training they received on Day 1 but with less stringent criteria. For the context-training stage, only two perfect sets were required; the minimum number of sets to be completed was five, and the minimum length of time to be spent was 10 min. The criteria for the second stage were not relaxed. This phase took about 15 min.

Discrimination phase. All of the participants did the discrimination task as in Experiment 2, except that only the green region was tested and pairs differed either on lightness or hue. Adjacent horizontal (hue) and vertical (lightness) pairs shown in Table 3 were used. For both hue and lightness dimensions, there were 8 within/new-category (columns/rows 1–2 and 3–4) and 4 across/new-category (columns/rows 2–3) pairs. Trials were presented in blocks, within which all different pairs varied on the same dimension. There were four blocks of trials, two for each dimension.

---

[6] The range of lightness variation used in Experiment 4 (range in Munsell value = 0.75) is narrower than in previous experiments. This was chosen to equate discriminability for hue and lightness pairs on the basis of pilot work findings. Indeed, discrimination performance by controls on hue and lightness was equivalent in Experiment 4.

Each of the 12 different pairs was repeated twice, and each same pair was shown once within each block of trials. This resulted in 80 (48 different and 32 same) trials per dimension. The order of presentation of blocks, and for trials within blocks, was randomized.

Control participants were shown the stimulus matrix in the same way as in Experiment 2 before the test phase. All of the participants received a set of 20 practice trials with the experimental stimuli before beginning the task. Participants typically took around 30 min to complete the same–different judgment task.

Results

A′ scores were calculated from same–different data as in previous experiments.

Perceptual Structure of the Stimulus Set

Figure 5 plots the same–different performance by controls, HCs (left), and LCs (right) on both dimensions, for within/new-category and across/new-category pairs. It appears that for controls, judgments on the two dimensions were about equally difficult ($F < 1$), as were within/new-category and across/new-category judgments ($F < 1$). Thus the use of the CIE distance measure seems to have compensated for the effects of the preexisting category structure found in Experiment 2.

Overall (Dimensional) Improvement

Figure 6 plots mean within/new-category and across/new-category scores on hue (left half) and lightness (right half) judgments for HCs and LCs only. It can be seen that lightness judgments tended to be better overall than hue judgments regardless of participants' training dimension. This was supported by a significant main effect of dimension, $F(1, 18) = 7.31$, $MSE = 0.011$, $p < .05$, $\eta^2 = 0.29$, and a nonsignificant interaction with training group, $F(1, 18) = 2.57$, $MSE = 0.011$, $p = .13$, $\eta^2 = 0.13$.

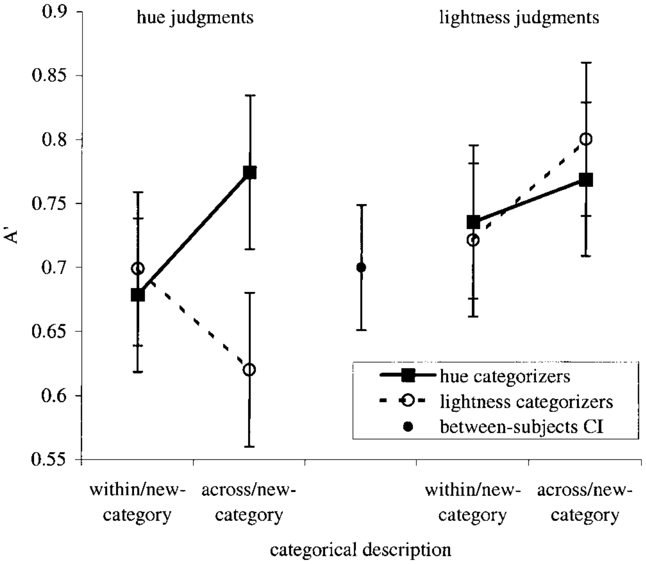

Figure 6. Mean within/new-category and across/new-category *A′* for hue and lightness categorizers on same–different judgments along two dimensions in Experiment 4. For each group, $n = 10$. Error bars represent ± 95% within-subject confidence intervals (CIs). The free floating point and its error bars represent ± 95% between-subjects CI.

Given that controls were no more accurate for lightness than hue, this result cannot be due to inherent properties of the stimulus set (hue and lightness judgments were equally difficult to begin with). It appears from the left panel of Figure 5 that HCs have improved discrimination for both dimensions, whereas LCs have been sensitized only for the lightness dimension (see right panel of Figure 5). This was supported by separate two-way ANOVAs on HCs and LCs. The main effect of judgment dimension was sig-

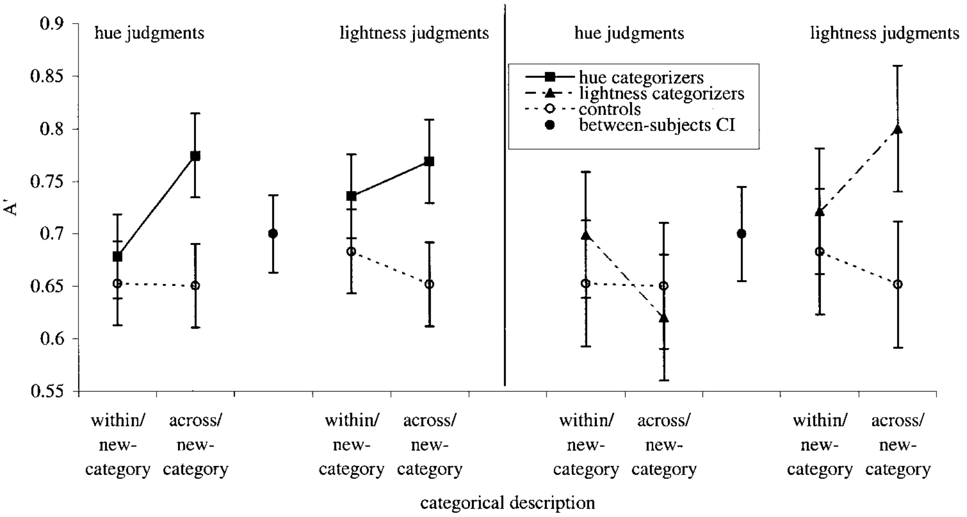

Figure 5. Comparisons of mean within/new-category and across/new-category *A′* scores between hue categorizers and controls (left) and between lightness categorizers and controls (right) on same–different judgments along two dimensions in Experiment 4. For each group, $n = 10$. Error bars represent ± 95% within-subject confidence intervals (CIs). The free floating points and their error bars represent ± 95% between-subjects CIs.

nificant for LCs, $F(1, 9) = 12.37$, $MSE = 0.0083$, $p < .01$, $\eta^2 = 0.58$, but not for HCs ($F < 1$). LCs were better at lightness judgments than hue judgments, whereas HCs were equally good at both types of judgment

If the training groups have been sensitized to their training dimension, they should also be more accurate for this dimension than controls. The left panel of Figure 5 shows a comparison of HCs with controls on hue (left half) and lightness (right half) judgments. It can be seen that HCs had higher scores overall than controls, $F(1, 18) = 10.29$, $MSE = 0.012$, $p < .01$, $\eta^2 = 0.36$, and this was not restricted only to hue (interaction of dimension and group was not significant; $F < 1$). A similar comparison between LCs and controls is summarized in the right panel of Figure 5. It can be seen that LCs tended to have higher scores for lightness judgments than controls, whereas this was not the case for hue judgments. This was supported by a nonsignificant main effect of group, $F(1, 18) = 2.84$, $MSE = 0.0018$, $p = .11$, $\eta^2 = 0.14$, and a significant interaction between group and judgment dimension, $F(1, 18) = 5.57$, $MSE = 0.0066$, $p < .05$, $\eta^2 = 0.24$. For lightness judgments alone, LCs were significantly better than controls, $F(1, 18) = 8.42$, $MSE = 0.01$, $p < .01$, $\eta^2 = 0.32$, whereas on hue judgments, they were no better than controls (no main effects or interaction).

In summary, both training groups showed enhanced accuracy on their respective training dimension. However, only HCs showed enhanced accuracy to the irrelevant training dimension (lightness). There was no evidence of acquired equivalence; judgments were not worse than controls for any comparison.

Induced Dimension-Specific CP Effects

It can be seen in Figure 6 that for hue judgments, HCs had higher across/new-category than within/new-category scores, whereas the reverse was true for LCs. Similarly, LCs had higher across/new-category than within/new-category scores for lightness judgments. The same was true, however, for HCs, but the pattern was less marked than for LCs. These impressions were supported by a significant three-way interaction (within vs. across, dimension, and HCs–LCs), $F(1, 18) = 7.52$, $MSE = 0.0081$, $p < .05$, $\eta^2 = 0.3$. This interaction was explored with separate two-way analyses on hue and lightness judgments. For hue judgments, the group by within-across interaction was significant, $F(1, 18) = 12.18$, $MSE = 0.0063$, $p < .01$, $\eta^2 = 0.4$; HCs had higher across/new-category than within/new-category scores, whereas the reverse was true for LCs. Post hoc comparisons for this analysis showed that across/new-category judgments by HCs were significantly better than within/new-category ones ($p < .05$), and they were significantly better than both within/new-category ($p < .05$) and across/new-category ($p < .01$) judgments by LCs. No difference between HCs and LCs was observed on within/new-category judgments of hue. These findings are consistent with cross-category expansion being responsible for induced CP effects.

For lightness judgments, however, the equivalent interaction was not significant ($F < 1$); although LCs did have higher across/new-category scores than within/new-category, HCs displayed the same pattern. In this case both groups had higher across than within scores: main effect, $F(1, 18) = 4.78$, $MSE = 0.0066$, $p < .05$, $\eta^2 = 0.21$. This once again suggests that cross-category expansion resulted from lightness categorization.

Figure 7 shows mean within/new-category and across/new-category scores on hue judgments for HCs and controls (left half), and those on lightness judgments for LCs and controls (right half). It can be seen that both training groups have higher across/new-category than within/new-category scores for judgments on their respective training dimension, whereas across/new-category judgments are no better than within/new-category judgments for controls.

Separate ANOVAs on each judgment type and follow-up post hoc tests supported these impressions. First, for hue judgments, although the group by within-across interaction was not quite significant, $F(1, 18) = 4.14$, $MSE = 0.0058$, $p = .057$, $\eta^2 = 0.19$, the pattern of post hoc test results shows that there was a CP effect for HCs but not for controls. HCs were more accurate at across/new-category judgments than their own within/new-category judgments and than controls' within/new-category and across/new-category judgments (maximum $p < .05$). Within/new-category scores of the two groups did not differ significantly. Again, CP seems to have resulted from cross-category expansion (acquired distinctiveness) rather than within-category compression (acquired equivalence).

Second, for lightness judgments, the group by within-across interaction, $F(1, 18) = 4.60$, $MSE = 0.0066$, $p < .05$, $\eta^2 = 0.2$, was significant; LCs were better at across/new-category than within/new-category judgments, whereas controls were not. Post hoc tests showed that LCs' across/new-category judgments were more accurate than their own within/new-category judgments and than both controls' type of judgment (maximum $p < .05$). Moreover, within/new-category performances by the two groups did not differ; again, cross-category expansion seem to underlie CP.

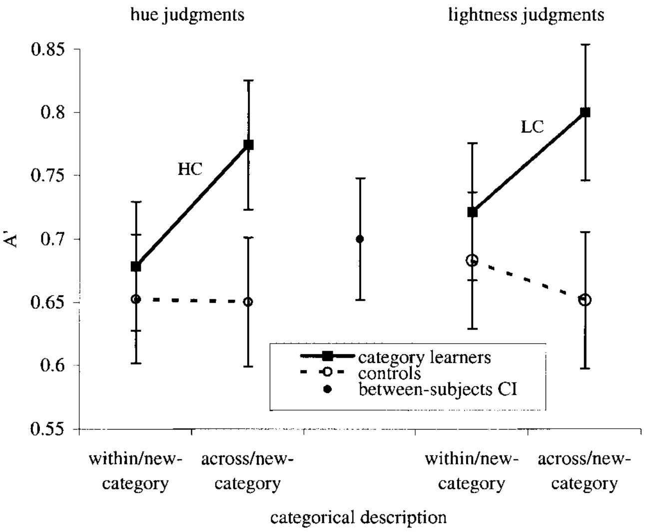

Figure 7. Mean within/new-category and across/new-category A' for category learners (HC = hue categorizer group; LC = lightness categorizer group) and controls on same–different judgments along two dimensions. Only the trained dimension scores for category learners are shown. For each group, $n = 10$. Error bars represent ± 95% within-subject confidence intervals (CIs). The free floating point and its error bars represent ± 95% between-subjects CI.

Discussion

Using stimuli spaced according to CIE distance created an even playing field. Across/new-category judgments were no more accurate than within/new-category judgments for the control group. This starting point made the effects of acquired CP clearer. No special pleading of having to unlearn the effects of preexisting categories was needed.

Learning either hue- or lightness-based categories induced CP. Judgments across the newly learned borders were more accurate than within/new-category judgments. And the size of this effect was of the same order for the two training groups (see Figures 5–7). This finding is consistent with the unusual division of blue into two basic color terms distinguished by lightness, affecting the structure of color similarity space of its speakers (Özgen & Davies, 1998). In the plausibility argument we proposed in the introduction, we suggested that, in effect, language learning functioned as a more haphazard version of the category training we have used in our experiments, but with a similar end result.

The mechanism underlying acquired CP seems to be cross-category expansion (acquired distinctiveness). Categorization learners on each dimension had higher across/new-category scores than other groups, as well as higher scores than their own within/new-category scores. At the same time, there were no differences on within/new-category scores; in other words, there was no evidence of within-category compression (acquired equivalence). This pattern is consistent with Goldstone's (1994) findings on category learning of size and brightness.

Interesting findings emerged from Experiment 4 in relation to the issue of dimensional sensitization. No evidence for the acquired equivalence of an entire dimension was found. CLs were not less accurate on the category-irrelevant dimension than were controls. However, acquired distinctiveness for the category-relevant dimension was observed for both HCs and LCs. Most interestingly, HCs had improved judgments on both hue and lightness dimensions, whereas LCs improved only on lightness judgments. The latter finding is consistent with perceptual learning findings, which suggest that attention plays an important role in perceptual learning (Ahissar & Hochstein, 1993; Shiu & Pashler, 1992). The discrimination improvement shown by LCs did not transfer to differences on the hue dimension, which they did not attend. However, this attentional explanation does not account for the finding that hue learning transferred to lightness. It is possible that this could be due to an asymmetrical relationship of *integrality* (Garner, 1974; Lockhead, 1972) between hue and lightness dimensions of color. Attending to (category-relevant) hue variation in the presence of (category-irrelevant) lightness variation results in sensitization to lightness. This may be because participants were not able to extract hue information independently of lightness variation. However, the reverse was not true. Participants who attended to lightness in the presence of an irrelevant hue variation did not seem to have improved hue sensitivity. It is possible to argue that they were able to extract lightness information without having to attend to hue. These results are consistent with the findings reported by Burns and Shepp (1988), who instructed their participants on dimensional relations by using clearly analyzable (height and width of a rectangle) as well as integral stimuli (hue, lightness, and saturation dimensions of color). After making sure that their participants understood dimensional relations, they used a triad classification task in which participants were asked to pick the two stimuli that shared a dimensional value. They found that although participants were able to do this for lightness and saturation against hue, the reverse was not true. They were not able to extract hue information against lightness or chroma. This asymmetrical pattern was also true for professionals who knew the Munsell system very well. It is therefore possible that transfer of learning to a stimulus dimension different from the one trained occurs only if the two dimensions are nonanalyzable. In the case of lightness, attentional mechanisms seemed to play a role in producing perceptual learning as suggested by earlier research (Ahissar & Hochstein, 1993; Shiu & Pashler, 1992).

General Discussion

The evidence reported in the present study is consistent with the possibility that category effects in color discrimination and memory, like those reported by Bornstein and Korda (1984) and Boynton et al. (1989), are acquired through learning. Experiment 1 provided evidence suggesting that color discrimination is flexible and there is room for improvement through relatively short laboratory training regimes. Experiments 2 to 4 showed that it is possible to induce CP effects on same–different judgments through a relatively fast learning regime based on hue as well as lightness categorization. Further, the effects of existing color categories on same–different judgments can be reversed through training. For untrained controls, equal Munsell hue differences across a given region resulted consistently in superior judgments for boundary over focal areas of the linguistic category occupying that region. After training on a new category boundary in a previously focal area, however, this pattern was reversed. In other words, participants displayed the effects of new categories at the expense of the effects of existing categories. Stimulus differences previously "played down" or altogether ignored became important and sensitized.

These findings are consistent with the LRH. Color term acquisition could be viewed as a similar process of categorization learning that takes place over a much longer time. If so, it might be that boundary effects are acquired through this process as they were in the present set of experiments. This mechanism can explain cross-linguistic effects on color memory, discrimination, and similarity space reported in earlier studies (Kay & Kempton, 1984; Özgen, 2000; Roberson et al., 2000).

Goldstone (1994) proposed that category learning results in dimensional sensitization specifically at category-relevant locations. This argument provides a possible mechanism underlying the present results. In comparison with a within/new-category pair, the difference between an across/new-category pair of colors becomes salient or important because of its relevance to the training task demands. Identifying the correct category of stimuli during training can be achieved by comparison with a representation of the trained category boundary. Receiving feedback on each trial adds to the stability of the location of this representation along the category-relevant dimension or shifts its position accordingly, particularly in the initial stages of training. As training progresses, identification errors are increasingly restricted to stimuli that are close to the boundary representation. For colors that are very close to the boundary, correct categorical identification occurs when a difference is detected, between the point that the stimulus occupies and the point at which the boundary representation is formed. This difference thus becomes important. Such emphasis along with repeated practice and exposure results in sensitization to differences in the boundary area. This can be detected on a subsequent same–different judgment task, in which accuracy in judging the differences between pairs of colors crossing this boundary is improved. This mechanism is consistent with the effect of category learning on perceptual similarity space proposed by Livingston et al. (1998).

Perception of Color Dimensions

Interesting results were found in Experiment 4 concerning the dimensional structure of color. It is generally accepted that color stimuli are integral (Burns & Shepp, 1988). Dimensions of a stimulus that are integral cannot be processed independently of each other (Garner, 1974; Lockhead, 1972). Goldstone (1994) argued that if two dimensions are integral, category learning on the basis of one dimension should thus lead to discrimination improvement on both dimensions. The present study found evidence suggesting an asymmetrical relationship between hue and lightness. It seems that attending to lightness differences can be independent of hue variation, whereas the reverse is not true. That is, categorizing on the basis of hue led to heightened sensitization for hue and lightness, whereas lightness categorization only caused sensitization for lightness differences. This asymmetry between hue and lightness (as well as hue and saturation) has also been reported in restricted classification tasks (Burns & Shepp, 1988).

Perceptual Change, or Labeling?

Although the evidence presented here is consistent with a learning-based account of CP effects for linguistic color categories, it is far from conclusive. Explanations of the present findings in terms of acquiring high-level skills such as an efficient labeling strategy are possible. Roberson and Davidoff (2000) reported that linguistic CP of colors can be interfered with (eliminated) by using an articulatory suppression task to prevent rehearsal of verbal labels during the ISI in a 2-AFC discrimination paradigm.

The present findings provide some insight into this issue. In Experiment 1, improvement in hue and lightness discrimination performance was observed after repeated practice. There were no category labels available to the participants, and the only labeling could have involved labeling each individual stimulus in the set. Further, in Experiments 2 to 4, overall improvement in discrimination for pairs of colors varying along the category-relevant dimension was found. A category labeling strategy should only enhance the cross-category judgments. Explanations in terms of labeling each individual stimulus in the category-learning experiments are problematic because participants trained on randomly generated colors spanning an area and did not necessarily see the same stimuli in the test phase.

Goldstone, Lippa, and Shiffrin (2001) addressed this issue with an experiment designed explicitly to unravel the effects of category learning on internal object representation from strategic effects. They found that the perceived similarity of each of two same–category faces to a neutral (noncategorized) face becomes more similar to each other as a result of category learning. If Face A and B are in the same category, and Face E is not categorized, the similarity ratings between A and E become similar to those between B and E. Goldstone et al. (2001) argued that because the neutral face (E) was not categorized prior to the ratings, this finding cannot be explained by a labeling strategy account. They proposed that the common features of objects in the same category are emphasized, which leads to their object descriptions (representations) becoming more similar through categorization.

Bornstein and Korda (1984) presented evidence from a successive-presentation discrimination task with an ISI of 250 ms using reaction time as their measure. They argued that cross-category judgments are faster because they are easier to process. They suggested that although the “pure visual” input is sufficient to judge differences, “category information” results in speed gains. They concluded therefore that visual and categorical information are processed in parallel. Such additional aid in processing stimulus differences may underlie the effects reported in the present study. Goldstone (1994) used a 33-ms ISI in the test phase of his category learning experiments and concluded that the effects were on perceptual discrimination. Indeed, the effectiveness of attaching previously unavailable labels to stimuli in such a short time is questionable. Furthermore, Garner (1974) reviewed findings that suggest that discrimination and memory are highly similar processes. He argued that in both cases the problem consists of encoding and comparison, with the latter based on similarity.

The problem then is perhaps not whether category effects are simply the result of consciously using or remembering names through verbal rehearsal but one that concerns a deeper question: To what extent do percepts and concepts interact?

Cognitive Penetrability of Perception

Goldstone (1995) asked participants to simulate the hue of a symbol by adjusting that of an identical symbol presented simultaneously. The test shape belonged to two categories (letters and numerals), which also differed from each other in terms of their dominant hue (more red or more violet). Goldstone found that the “redness” or “violetness” of a symbol was underestimated or overestimated in the hue-matching task toward the dominant hue of its shape category. Further, this was not restricted to “preset” categories like letter and numerals. When the stimulus sets consisted of shapes that formed two categories (polygons and conjoined lines) defined on the basis of the experimental context, systematic mistakes in hue estimates were made in a similar way. The hue judgment task did not require any shape-based categorization. The process of categorization that influenced hue perception was automatic.

Evidence for the penetrability of the low-level perceptual system can also be found in contrast-detection paradigms in psychophysics. Visual input is analyzed at multiple spatial scales using separate filters tuned to specific frequency bands (De Valois & De Valois 1990; Marr, 1982). Evidence suggests that spatial filtering of the visual input can be selectively used, depending on task demands. Observers are worse at detecting sinusoidal gratings when spatial frequency varies unpredictably than when it is the same across a block of trials (Davis & Graham, 1981; Davis, Kramer, & Graham, 1983). However, when observers are precued to the spatial frequency of a stimulus in a mixed block, these *uncertainty effects* are eliminated (Hübner, 1996a, 1996b). Further, it has been recently suggested that such cuing effects show tuning functions that are similar to those of the spatial frequency filters (Sowden, Özgen, & Schyns, 2001). Sowden et al. showed that cuing sets attention to a band of spatial frequencies that corresponds to the typical tuning bandwidth of spatial frequency channels. Such findings may provide an explanation for evidence suggesting that the processing of spatial scale can be selective and flexible depending on the categorization task being performed (Schyns & Oliva, 1999).

Another way in which perception can interact with conception is through perceptual learning. Goldstone (1998) reviewed literature suggesting that mechanisms such as attentional weighting, CP, stimulus imprinting, differentiation, and unitization can lead to perceptual change. In the light of such phenomena, an analogy is made for the interaction between perception and cognition that assumes perception to be less rigid than previously thought (Goldstone, Schyns, & Medin, 1997). According to this view, rather than being a "foundation" that cognitive processes act upon, perception is viewed as a "bridge" in that it is flexible enough to "sway under the weight" of cognitive processes (p. 1). For example, feature repertoires do not have to be fixed or finite; evidence suggests that functional, distinctive features can be *created* as a result of category learning (Schyns, Goldstone, & Thibaut, 1998; Schyns & Rodet, 1997). Such findings are supported by computational models of category learning (Schyns & Rodet, 1997).

There is evidence suggesting that perceptual learning may result in cortical change. For example, training to discriminate between sound frequencies leads to larger cortical representation for the familiar frequencies in monkeys (Recanzone, Schreiner, & Merzenich, 1993). There is also evidence suggesting that category learning may be supported by cortical plasticity. For example, using functional MRI, Reber, Stark, and Squire (1998) identified cortical areas correlated with category learning, showing significant increase or decrease in activity. They also found that as a result of category learning, activation in the visual cortex seemed to drop during the processing of categorical patterns, suggesting that such stimuli were processed rapidly and effortlessly as a result of prototype learning.

These examples point to the possibility that perception may be flexible enough to interact with cognitive processes, even though mechanisms of early vision may be impenetrable by cognition (Pylyshyn, 1999). It is possible that color category effects of the kind under scrutiny here may be viewed as part of such an interactive process. This type of mechanism is consistent with the findings of the present study pointing to a learning explanation of the perceptual effects of language categories. Although it is not possible to completely rule out the idea that such effects may result merely from high-level processes like "name strategies" (Kay & Kempton, 1984), the evidence presented here warrants further consideration of a relativist approach to color perception and cognition.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences, USA, 90,* 5718–5722.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 115–129.

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,* 207–222.

Boynton, R. M., Fargo, L., Olson, C. X., & Smallman, H. S. (1989). Category effects in color memory. *Color Research and Application, 14,* 229–234.

Boynton, R. M., & Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research and Application, 12,* 107–123.

Burns, B., & Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation and brightness. *Perception & Psychophysics, 43,* 494–507.

Davies, I. R., & Corbett, G. G. (1994). The basic color terms of Russian. *Linguistics, 32,* 63–89.

Davies, I. R., & Corbett, G. G. (1997). A cross-cultural study of color grouping: Evidence for weak linguistic relativity. *British Journal of Psychology, 88,* 493–517.

Davies, I., & Corbett, G. (1998). A cross-cultural study of color grouping: Tests of the perceptual–physiology account of color universals. *Ethos, 26,* 338–360.

Davis, E. T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision Research, 21,* 705–712.

Davis, E. T., Kramer, P., & Graham, N. (1983). Uncertainty about spatial frequency, spatial position, or contrast of visual patterns. *Perception & Psychophysics, 33,* 20–28.

De Valois, R. L., & De Valois, K. K. (1990). *Spatial vision.* Oxford, England: Oxford University Press.

Fahle, M. (1997). Specificity of learning curvature, orientation, and vernier discriminations. *Vision Research, 37,* 1885–1895.

Fine, I., & Jacobs, R. A. (2000). Perceptual learning for a pattern discrimination task. *Vision Research, 40,* 3209–3230.

Fletcher, R. (1980). *The City University Color Vision Test* (2nd ed.). London: Keeler.

Garner, W. R. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Gerhardstein, P., Renner, P., & Rovee-Collier, C. (1999). The roles of perceptual and categorical similarity in color pop-out in infants. *British Journal of Developmental Psychology, 17,* 403–420.

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

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science, 6,* 298–304.

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

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

Goldstone, R. L., Schyns, P. G., & Medin, D. L. (1997). Learning to bridge between perception and cognition. In R. L. Goldstone, P. G. Schyns, & D. L. Medin (Eds.), *Psychology of learning and motivation: Perceptual learning* (Vol. 36, pp. 1–14). San Diego, CA: Academic Press.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75,* 424–429.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: a critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 29–52). New York: Cambridge University Press.

Heider, E. R. (1971). "Focal" color areas and the development of color names. *Developmental Psychology, 4,* 447–455.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology, 93,* 10–20.

Hübner, R. (1996a). The efficiency of different cue types for reducing spatial-frequency uncertainty. *Vision Research, 35,* 401–408.

Hübner, R. (1996b). Specific effects of spatial-frequency uncertainty and different cue types on contrast detection: Data and models. *Vision Research, 36,* 3429–3439.

Hunt, R. W. G. (1987). *Measuring color.* Chichester, England: Horwood.

Karni, A., & Bertini, G. (1997). Learning perceptual skills: Behavioral probes into adult cortical plasticity. *Current Opinion in Neurobiology, 7,* 530–535.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature, 365,* 250–252.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M., & Sagi, D. (1994). Dependence on REM-SLEEP of overnight improvement of a perceptual skill. *Science 265,* 679–682.

Kay, P., & Kempton, W. (1984). What is the Sapir–Whorf hypothesis? *American Anthropologist, 86,* 65–79.

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language, 54,* 610–646.

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.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review, 79,* 410–419.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler, S. Yantis, D. Medin, R. Gallistel, & J. Wixted (Series Eds.) & H. Pashler & J. Wixted (Vol. Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339–390). New York: Wiley.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review, 1,* 476–490.

Lucy, J. A., & Schweder, R. A. (1979). Whorf and his critics: Linguistic and non-linguistic influences on color memory. *American Anthropologist, 81,* 581–615.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal Experimental Psychology: General, 115,* 39–57.

Özgen, E. (2000). *Language, learning, and colour categorisation.* Unpublished doctoral dissertation, University of Surrey, Surrey, England.

Özgen, E., & Davies, I. R. L. (1998). Turkish color terms: Tests of Berlin and Kay's theory of color universals and linguistic relativity. *Linguistics, 36,* 919–956.

Pastore, J. (1987). Psychophysical aspects of categorical perception. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 29–52). New York: Cambridge University Press.

Pilling, M., Wiggett, A., & Davies, I. R. L. (2002). *Is color categorical perception a direct-language effect?* Manuscript submitted for publication.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341–365.

Ratner, C. (1989). A sociohistorical critique of naturalistic theories of color perception. *Journal of Mind and Behavior 10,* 361–373.

Reber, P. J., Stark, C. E., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences, USA, 95,* 747–750.

Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience, 13,* 87–103.

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

Roberson, D. M. J., Davidoff, J., & Braisby, N. (1999). Similarity and categorization: Neuropsychological evidence for a dissociation in explicit categorization tasks. *Cognition, 71,* 1–42.

Roberson, D. M. J., Davies, I. R. L., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence in favor of linguistic relativity. *Journal of Experimental Psychology: General, 129,* 369–398.

Rosch, E. (1975). The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 303–322.

Saunders, B. A. C., & van Brakel, J. (1997). Are there non-trivial constraints on color categorization. *Brain and Behavioural Sciences,* 20, 167–178.

Schoups, A. A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature, 412,* 549–553.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–54.

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition, 69,* 243–265.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 681–696.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

Shiu, L., & Pashler, H. (1992). Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Perception & Psychophysics, 52,* 582–588.

Sowden P., Davies, I., Rose, D., & Kaye, M. (1996). Perceptual learning of stereo acuity. *Perception, 25,* 1043–1052.

Sowden, P. T., Özgen, E., & Schyns, P. G. (2001). Expectancy effects are spatial frequency tuned. *Perception, 30*(Suppl.), 91.

Sowden, P. T., Rose, D., & Davies, I. R. L. (2002). Perceptual learning of luminance contrast detection: Specific to spatial frequency and retinal location but not orientation. *Vision Research, 42,* 1249–1258.

Valberg, A. (2001). Unique hues: An old problem for a new generation. *Vision Research, 41,* 1645–1657.

Weiss, Y., Edelman, S., & Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation, 5,* 695–718.

Whorf, B. L. (1956). *Language, thought and reality.* Cambridge, MA: MIT Press. (Original work published 1940)

Zanker, J. M. (1999). Perceptual learning in primary and secondary motion vision. *Vision Research, 39,* 1293–1304.

Received March 13, 2002
Revision received July 29, 2002
Accepted July 31, 2002 ■