Learned Categorical Perception for Natural Faces

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
Department of Psychology, University of Adelaide, SA 5005, Australia

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide, SA 5005, Australia

Hannah Nikkerud (hannah.nikkerud@adelaide.edu.au)
Department of Psychology, University of Adelaide, SA 5005, Australia

Abstract

We present an experiment involving learned categorical perception for natural faces, using four different category learning tasks. In the four tasks, participants learned to classify faces divided on the basis of gender, hair color, a subjectively determined level of 'trust', or at random. After category learning, participants rated the similarity of each pair of faces, and their judgments were compared to a previously collected 'base rate' set of similarities for the same stimuli. Evidence for learned categorical perception was then sought, in the form of increased differences between intra-category and inter-category similarities. Over the four conditions, we observed no learned categorical perception when categories were based on obvious properties of the faces (gender and hair color), nor when the category structure was intentionally random. However, when the loosely defined category structure of trust was employed, a learned categorical perception effect emerged.

If every stimulus were perceived as a unique event, we would be rapidly inundated with pointless information. So we organize our perceptions into categories, allowing us to describe the world in a simpler manner, and to generalize better to novel situations. As a consequence of this restructuring, items belonging to the same categories tend to be perceived as more similar to one another, while items belonging to different categories appear less similar. These *categorical perception* phenomena have been observed in a range of domains such as phonemic categories (e.g., Liberman, Harris, Hoffman & Griffith, 1957), color categories (e.g., Bornstein & Korda, 1984; Ozgen & Davies, 2002), musical pitch categories (Burns & Ward, 1978), and facial expressions (e.g., Bimler & Kirkland, 2001; Etcoff & Magee, 1992). Categorical perception effects rely on an interaction between the perceptual representation of the environment and the conceptual representations used to interpret it, and so are a central issue in cognitive modeling.

In this paper we consider the acquisition of categorical perception for naturalistic faces. We present an experiment in which a prior category learning task is shown to influence subsequent similarity judgements, but only for some category structures.

Learned Categorical Perception

While categorical perception effects may be partially innate (e.g., Barrera & Maurer, 1981), much of human conceptual structure is likely to be learned, leading to the notion of *learned categorical perception*. Such studies typically measure the effect of a prior learning task on a subsequent discrimination task. For instance, Goldstone (1994) demonstrated that prior learning of categories based on stimulus size or brightness improved performance on subsequent same-different judgements based on the same dimension. The goal in studies of this kind is to see how the act of learning the category structure affects the subsequent judgements, and specifically to see if the recently learned category structure is reflected in these judgements.

Two main theoretical explanations have been proposed to account for learned categorical perception (Goldstone, Lippa & Shiffrin, 2001). According to the *altered object description* account, simply learning that a set of stimuli share a label is sufficient to change the way that they are perceived. In its strongest form, the altered object description account can be viewed as a variant of the linguistic relativity hypothesis (Whorf, 1964), in which linguistic labels are taken to form the underlying basis of perception itself. In a weaker form, it asserts that the prior learning experience allows people to observe or attend (perhaps temporarily) to different properties of the stimuli, adapting their representations to suit the context.

The alternative account is the *strategic judgement bias* account, which argues that categorical perception effects can be viewed as a response to task demands, in which people adopt a new decision strategy, but do not alter the representational description of stimuli. As with the altered object account, the strategic judgement bias account has strong versions and weak versions. A strong version of this theory states that the judgements are a deliberate response to the experimental task, and do not reflect any change in the way stimuli are perceived. For instance, if participants are told that two stimuli belong to the same class, then they will assume that they are supposed to rate them as more similar. Accordingly, they alter their similarity ratings, but do not alter the stimulus representation. In a weaker form, it asserts that prior category learning informs people that some properties are more likely to be relevant to the task than others, so people will focus their decision processes on these properties.

The weak form of the altered object description account seems similar to the weak form of the strategic judgement bias account. In both cases, the role of category learning is to allow people to attend to different

properties of the stimulus enviroment. The only difference is that the altered object description account treats this shift as a representational process, while the strategic judgement bias account views it as a decision process. Arguably, this is more a semantic disagreement than anything else. Accordingly, we refer to this as the *attentional reweighting* account.

The most notable feature of attentional reweighting is that it implies that not all category structures will produce categorical perception. If learned categorical perception results from a dimensional attention process (or feature weighting, in the discrete case) applied to the stimulus representations, then it will only appear in situations when the learned category structure allows this kind of attentional shift to take place. That is, the mechanism underlying learned categorical perception is constrained to a limited class of representational changes. Categories with no structure will not induce learned categorical perception, since no representational learning would be expected to take place. This is in contrast to the strong versions of both the altered object and strategic judgement bias accounts, in which the effect should not depend on the amount of structure inherent in the category. In what follows, we present an experiment into the learned categorical perception of faces, in which we systematically vary the amount of structure in the category, and observe a corresponding variation in the extent of categorical perception.

Categorical Face Perception

The stimulus domain we used was naturalistic faces. Faces are one of the most studied stimulus domains in categorical perception, since they are important visual stimuli in human environments (Ellis & Young, 1989). Moreover, faces are complex multidimensional stimuli, allowing a great variety in the kinds of properties that they possess. Faces can often be successfully classified remarkably quickly, even under changes in lighting, viewing angle, size and expression (Bruce, 1994), though often with a decrement in accuracy (Laughery & Wogalter, 1989). Furthermore, there is evidence that categorical perception for faces can be learned. For instance, Beale and Keil (1995) constructed a set of faces that interpolated between photographs of Bill Clinton and John Kennedy, and found evidence for strong boundary effects in terms of the resemblances to the two original photographs, whereas there were no such effects for faces that interpolated between photographs of two unfamiliar people. In a later study, Levin and Beale (2000) found categorical effects for unfamiliar faces, including other-race and inverted faces. Further evidence in favor of learned categorical perception for faces is provided by Stevenage (1998), who demonstrated that participants were better able to discriminate between twins after a category learning task.

In an elegant study designed to compare the altered object description account to the strategic judgement bias account, Goldstone et al. (2001) used four faces (A, B, C and D) that were morphed to produce a total of 16 faces with different degrees of similarity to the originals. Participants completed a pre-categorization similarity task, a category learning task (e.g., in which A and B formed one category, and C and D formed the other), followed by another similarity task. There was a decrease in the similarity ratings for between-category pairs, but no corresponding increase in within-category similarities. In order to discriminate between the altered object description account and the strategic judgement bias account, they included a neutral face E in the similarity rating tasks. Under a strategic decision view, A, B, C and D should all have the same relationship to E after the category learning task as they did beforehand. However, under an altered description view, the similarity between A and E should grow more like the similarity between B and E. Since the latter was observed empirically, they concluded that some representational change had occurred.

Experiment

The aim in the current experiment is to examine the effect of the category structure itself on learned categorical perception. To that end, we are interested in three qualitatively different types of category. First, we considered highly structured categories, to which people are presumably already attuned. In these cases, the category structure is likely to have been learned prior to the experiment, and we would predict that categorical perception effects would be observed irrespective of whether participants were "trained" on this category. Secondly, we considered thoroughly arbitrary categories with no particular structure. In these cases, we would predict that no categorical perception would exist *a priori*, nor would any be learned. Finally, we considered category structures with limited structure, which people would not perceive *a priori*, but could learn to do so.

Method

Participants. Forty participants (10 males, 30 females) aged 18 to 58 years took part in the experiment, recruited through the University of Adelaide.

Materials. The stimuli were the 25 frontal photographs of faces, originally obtained from the Psychological Image Collection at Stirling (PICS) repository[1], for which pairwise similarity ratings had previously been collected by O'Doherty and Lee (2002). The faces were presented in color against a light background and occupying an area of approximately 5cm×5cm on a computer screen. These faces are shown in Figure 1.

Procedure. Participants completed a category learning task followed immediately by a similarity-rating task. We used four conditions for the category learning task in a between-subjects design. The similarity judgements collected by O'Doherty and Lee (2005) for these stimuli were used as a control condition. In the first experimental condition, the category to be learned separated the female faces (A-L) from the male faces (M-Y), intended to represent a highly structured category. In the second condition, also intended to be structured, the faces were

[1] http://pics.psych.stir.ac.uk/

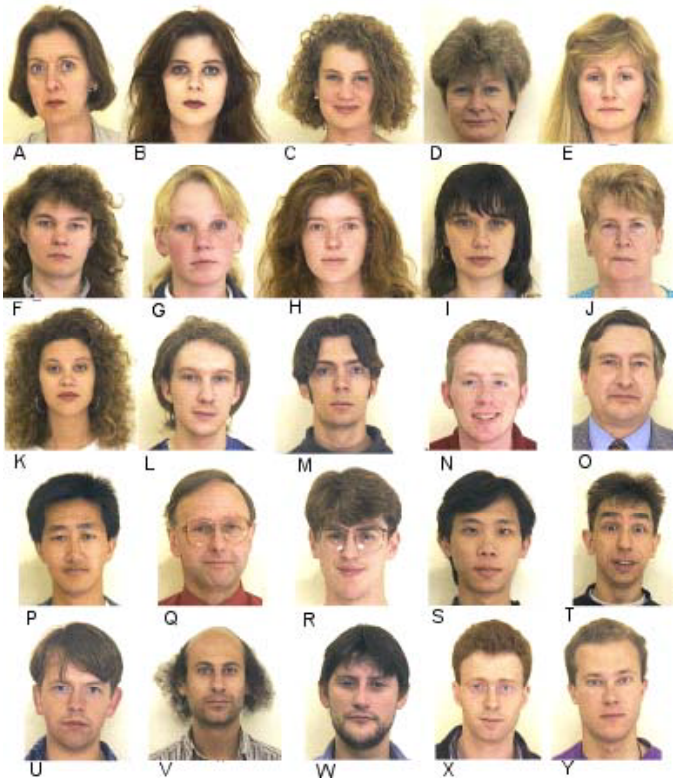

Figure 1: The 25 photographs used in the experiment, labeled A to Y.

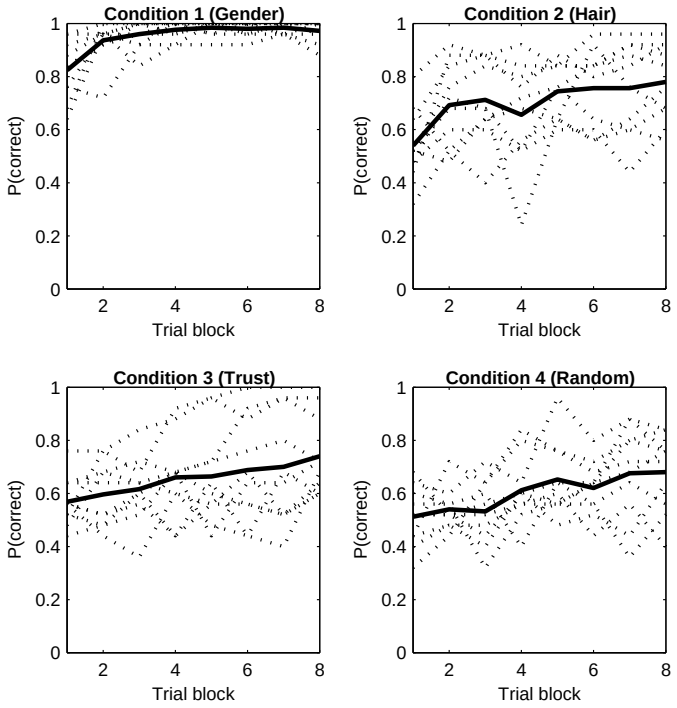

Figure 2: Learning curves for faces across the four conditions. The dotted lines show the learning curves for every participant in the experiment, while the solid lines show the averaged curves for each condition.

distinguished by hair color, with the "lighter" category consisting of faces C, D, E, G , J, L, N, Q and Y.

In the third condition, the categories were based on a subjective measure of "trustworthiness", with the intent to capture a loosely-defined, partially structured category. Broadly speaking, "trustworthy" faces were based on features such as large eyes, large forehead, small chin, softer face shape, high eyebrows, and smiling expression (see Berry, 1990; McArthur & Apatow, 1983/34; Zebrowitz & Montepare, 1992; Berry & McArthur, 1986). Faces C, D, E, H, K, L, N, O, P, T, U and X were treated as trustworthy.

In the final condition, faces were divided into "random" categories, intended to lack any noticeable structure. The categories contained an equal number of males, an equal number of lighter-hair faces and an equal number of the trustworthy faces. Additionally, we attempted to ensure that no other obvious feature distinguished between the two categories. Faces A, C, E, F, J , K, M, N, P, R, U and V formed one of the categories.

During the category learning task, stimuli were presented for 8 blocks in which each face was presented exactly once in a random order. On each trial, participants were asked which category the face belonged to, and received feedback after their response. In the subsequent similarity-rating task, participants rated the similarity of all $\binom{25}{2} = 300$ pairs of stimuli on a five-point scale.

Results

Category Learning. Figure 2 shows the learning curves for all participants and all conditions. Each of the dotted lines represents the learning curve for an individual participant, while the solid lines represent the averaged performance for each condition. Not surprisingly, condition 1 was the easiest category structure to learn, with the average proportion of correct responses rising from 82% in the first block to 97% in the final block. As shown in Figure 3, the only face that caused difficulty in the first 75 trials was face L.

As expected, the second easiest category to learn was the one based on hair color (condition 2), in which performance rose from 56% correct in the first block to 82% correct in the final block. When broken down by item, as in Figure 3, the interesting face is R, for which performance was at chance at the beginning of the experiment, but was almost universally misclassified over the last 75 trials.

For condition 3, the category structure based on "trust", the proportion of correct responses rose from 57% in the first block to 74% in the final block. The comparison of the individual item performance for the first 75 trials with the last 75 trials shown in Figure 3 shows no obvious structure. Performance seems to have improved in a reasonably similar way for most items. In condition 4, performance rose from 50% to 65%, and

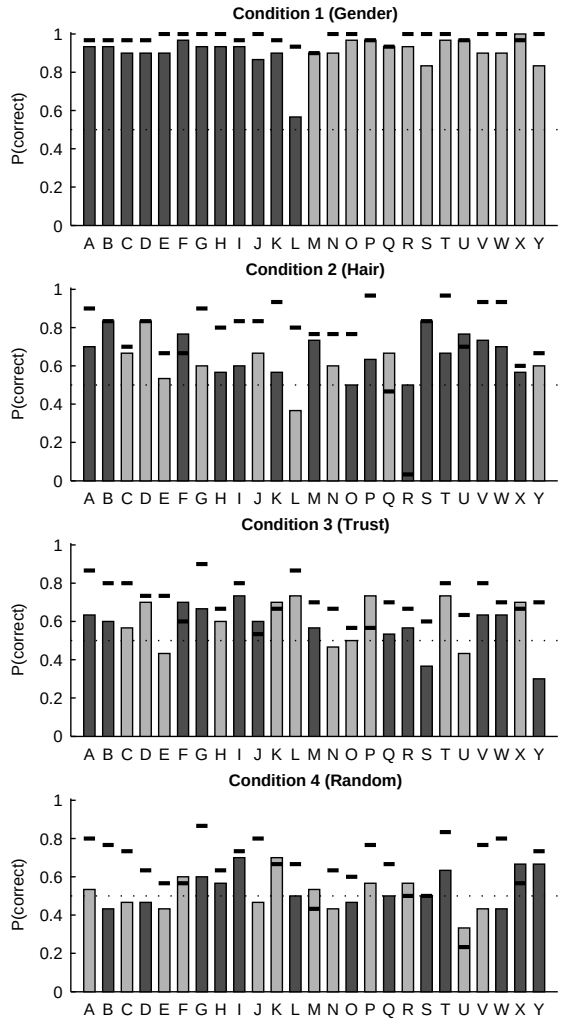

Figure 3: Proportion correct for each face in each condition. The solid bars represent performance over the first 75 trials, while the floating bars represent performance over the last 75 trials.

once again there is no obvious pattern to the participants' performance.

Similarity Data. Figure 4 plots the reported similarities between faces in all four conditions. On the right hand side of each plot are the pairwise similarities between stimuli that belong to different categories, while the left hand side shows similarities between stimuli that belong to the same category. Ratings by participants in the base-rate or control condition are shown in white, while those by participants who had experienced the prior category learning are shown in black. The line plots shown with circular markers correspond to the mean similarities, and the squares plotted on either side show the full distribution of reported similarities.

The most salient aspect of the plots is the overall decrease in similarity from the control condition to the learning condition. The notable exception, however, is in the "hair color" condition, where the mean performance of the two groups is indistinguishable (although there is more variability in the learning group). The second salient aspect of Figure 4 are the effects of categorical structure itself. Conditions 1 and 2 show greater similarity for within-category items in both the control group and the learning group. Condition 4 shows less similarity for within-category items in both the control group and the learning group, though there is some evidence that this effect is slightly stronger in the learning group. Most importantly, however, in condition 3 there is no difference in within-category and between-category similarities for the control group, while the learning group rated between-category items substantially less similar than within-category items.

Discussion

Exposure and Dissimilarity. It is a general result that increased familiarity with a set of stimuli facilitates greater discriminability. In a sense, one can imagine a representational space being stretched, causing all stimuli to be a little further apart. Similarly, greater exposure may allow more appropriate features to be observed, fine-tuning the representation. This differentiation between stimuli should be expected to decrease the overall similarity between all stimuli, when compared with a group of participants that had not undertaken a prior category learning task. This decrease in similarity from the control condition to the learning condition is the most salient property in Figure 4.

The notable exception, however, is in the "hair color" condition, where the two groups are indistinguishable. A possible reason for this involves the nature of the feature that distinguishes between the two classes. If, during learning, people attend only to hair color and ignore other features, there would be little reason to expect any substantial differentiation between stimuli (a kind of "representational irrelevance"). In contrast, gender is not a "primitive" feature, and requires that attention be paid to a number of aspects of the face. Thus, even though gender is likely to be highly salient a priori, other features are still likely to be observed. Similarly, the "trust" condition and the "random" condition requires that a whole host of facial features and properties receive attention.

Variation in Learned Categorical Perception. In the gender and hair color conditions, the within-category similarity is higher than the between-group similarity for both the control and the learning groups, and there is no evidence of any kind of interaction effect. In other words, besides the effects discussed earlier (which are unlikely to be a categorical perception effect), the learning appears to have induced no learned categorical perception.

The reason for this seems to be simple. Gender and hair color are salient properties that influence the perception of faces in the real world (and form categories that are used in real life, e.g., "dark-haired male"), and to the extent that any categorical perception occurs, participants were sensitive to it prior to beginning the experiment. In other words, the learning in these con-

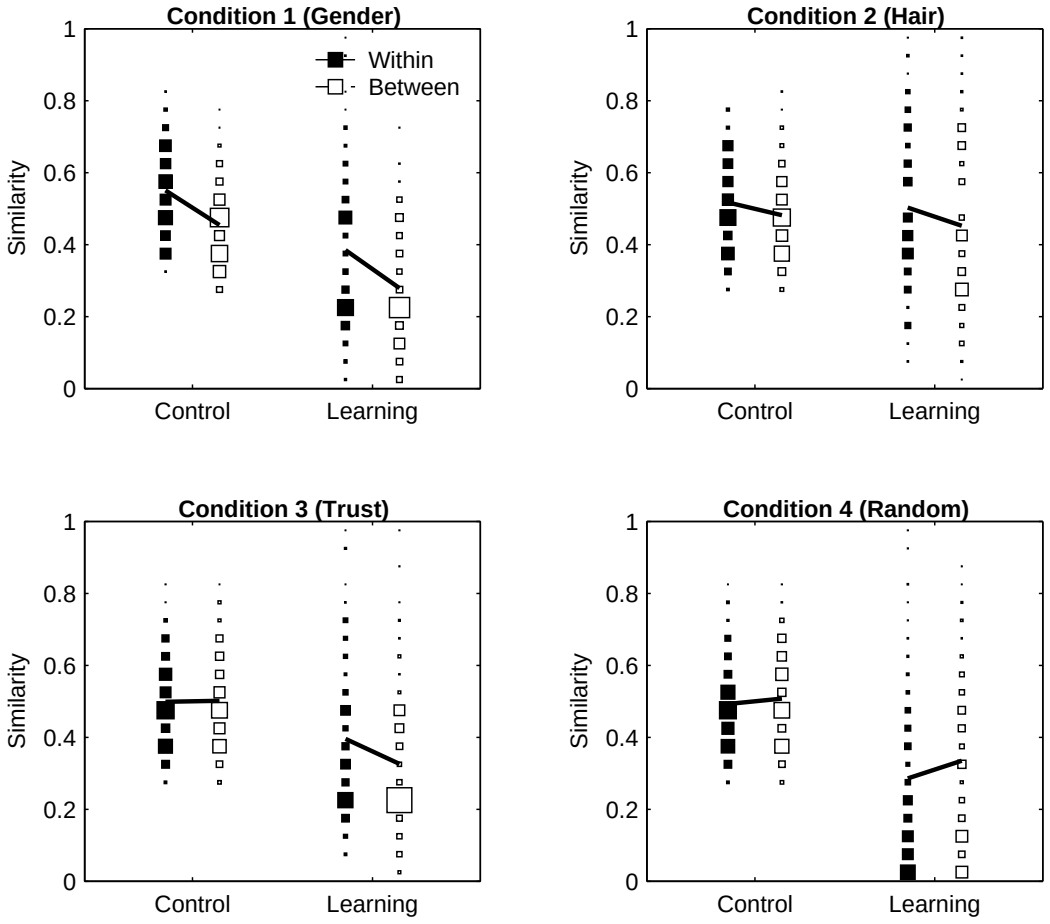

Figure 4: The distributions of the (normalized) pairwise similarity ratings for the 25 faces. Each of the 16 vertical plots is a histogram, in which the area of each rectangle is proportional to the probability associated with a particular similarity. Within each panel, the left side shows similarity distributions for the control group, while the right hand side shows the distribution of similarities for the learning group. In both cases, the similarity between stimuli that belonged to the same category (for the learning group) is plotted in black, and the similarity between stimuli that belonged to different categories is plotted in white. The line plots show the differences between the means. A potential categorical perception effect is evident in condition 3, in which the learning group rated within-category similarities higher than the between-category similarities, but the control group did not.

ditions was that certain properties that were already highly salient happened to be the appropriate basis for the classification decisions. Nothing new needed to be noticed, and no representational changes were required.

The trust and random conditions are rather different. In both cases, the control group shows no difference between the within-category and the between-category stimuli, suggesting that there is little structure in the categories that was salient to participants a priori. In condition 4, the learning group is similar to the control group, suggesting that no categorical perception was induced by learning. However, in condition 3, there is a marked difference: the between-category items are assumed to be much less similar than the within-category items.

General Discussion

When viewed from a rational perspective (e.g., Anderson, 1990) it is natural to expect a certain amount of variation in learned categorical perception. The most enduring regularity in our world is change. Some aspects of the environment fluctuate rapidly, such as weather and lighting conditions. If we entirely reshaped our assumptions about the structure of the world every time the lights were turned off, we would learn very little about the world. Instead, we "decide" to dilate our pupils, and behave more cautiously since we have less sensory information arriving. We do not assume that the walls disappear just because we cannot see them. On the other hand, some aspects of change are lasting, and our interpretations of the world shift with them. With the advent of the internet, the conceptualisations of commu-

nication (via email, for instance) and scholarship (via online libraries) have undergone substantial shifts, not easily reversible. In this context, we assume that the nature of the world has permanently changed, and reshape our world view to suit.

Given this, we suggest that focusing on the distinction between "representational shift" and "strategic decisions", though useful, can be somewhat misleading. Instead, when one learns a new category label, it is important to consider in which contexts the label is likely to be useful. Possibly, what we perceive as structure in a category are those regularities that are expected to be stable and important, enduring over a range of contexts and retaining some diagnosticity across them. If a new category label reflects a newly observed but possibly enduring regularity in the environment (e.g., finding tumours in x-rays), we should expect to observe relatively enduring shifts in representation and behavior. However, if the new label is arbitrary and of no general significance (e.g., most category learning experiments), any changes to behavior should be highly limited. If so, learned categorical perception effects are only likely to be observed when the new category is sufficiently important that people might reasonably expect it to be relevant in the future.

Acknowledgements

This work was supported by Australian Research Council grant DP-0451793. We thank Lama Chandrasena for his programming assistance.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Barrera, M. E. & Maurer, D. (1981). The perception of faces by the three-month old. *Child Development, 52*, 203-206.

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

Berry, D. (1990). Taking people at face value: Evidence for the kernel of truth hypothesis. *Social Cognition, 8*, 343-361.

Berry, D. & McArthur, L. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin, 100*, 3-18.

Bimler, D. & Kirkland, J. (2001). Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition and Emotion, 15*, 633-658.

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.

Bruce, V. (1994). Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology, 47*, 5-28.

Burns, E. M. & Ward W. D. (1978). Categorical perception – phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic music intervals. *Journal of the Acoustical Society of America, 63*, 456-468.

Ellis, H. D. & Young, A. W. (1989). Are faces special? In A. W. Young and H. D. Ellis (Eds) *Handbook of Research on Face Processing* (pp. 1-26). Amsterdam: Elsevier.

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

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

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

Laughery, K. R. & Wogalter M. S. (1989). Forensic applications of facial memory research. In A. W. Young and H. D. Ellis (Eds) *Handbook of Research on Face Processing* (pp. 519-555). Amsterdam: Elsevier.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*, 358-368.

McArthur, L.& Apatow, K. (1983/84). Impressions of baby-faced adults. *Social Cognition, 2*, 315-342.

O'Doherty, K. C. & Lee, M. D. (2002). The featural representation of animals based on similarity. *Australian Journal of Psychology, 54*, 60.

Ozgen, E. & Davies, R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*, 477-493.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology, 89*, 39-58.

Whorf, B. L. (1964). *Language, Thought and Reality*. New York: Macmillan.

Zebrowitz, L. & Montepare, J. M. (1992). Impressions of babyfaced individuals across the lifespan. *Developmental Psychology, 28*, 1143-1152.