In press, *Psychological Science*

Mutual information and categorical perception

Jacob Feldman
Dept. of Psychology, Center for Cognitive Science, Rutgers University, New Brunswick, NJ USA

Categorical perception (CP) refers to the enhancement of perceptual sensitivity near category boundaries, generally along dimensions that are informative about category membership. But it remains unclear exactly which dimensions are treated as "informative" and why. This paper reports a series of experiments in which subjects were asked to learn statistically-defined categories in a novel, unfamiliar two-dimensional perceptual space of shapes. Perceptual discrimination was tested before and after category learning at various features in the space, each defined by its position and orientation relative to the maximally informative dimension. The results support a remarkably simple generalization: the magnitude of improvement in perceptual discrimination at each feature is proportional to the *mutual information* between the feature and the category variable. This finding suggests a "rational" basis for categorical perception, in which the precision of perceptual discrimination is tuned to the statistical structure of the environment.

Statement of relevance: Categorical perception (CP) is a famous phenomenon of learning in which subjects learning categories become measurably more sensitive to the underlying perceptual features that distinguish the categories. CP has been extensively studied, but key aspects of it are still poorly understood: in particular, we do not know exactly *which* features subjects become more sensitive to, and why. This paper demonstrates a remarkably simple generalization: subjects learn to be more sensitive to features in proportion to how much information they convey about the categories to be learned. Mutual information (MI) is a basic information-theoretic measure that expresses the relationships among variables. This finding suggests a "rational" allocation of neural resources to perceptual features, shedding light on the way learning builds on and modifies the underlying neural representation of perceptual features. It should be of interest to all researchers interested in the underlying mechanisms of learning.

Introduction

Categorical perception (CP) refers to the enhancement of perceptual sensitivity near category boundaries (Harnad, 1987). After learning to classify stimuli into discrete classes, subjects' ability to make fine discriminations along perceptual dimensions that are "informative" about the categories can measurably improve. CP and associated changes in perceptual discrimination were first observed in phonological perception (Liberman, Harris, Hoffman, & Griffith, 1957) but have since been observed along a number of visual features, including orientation (Rosielle & Cooper, 2001), facial features (Rotshtein, Henson, Treves, Driver, & Dolan, 2005; Viviani, Binda, & Borsato, 2014) and shape (Gauthier, James, Curby, & Tarr, 2003; Folstein, Palmeri, & Gauthier, 2014). For example, in one influential study (Goldstone, 1994), subjects trained to classify objects into two categories of objects distinguishable by their size became more sensitive to size differences (improved discrimination, d'), but not (as much) to brightness differences. Such findings are remarkable because they reflect a profound interaction between cognitive and perceptual mechanisms: a change to basic perceptual processes attributable to the acquisition of a new concept by an adult organism (Schyns, Goldstone, & Thibaut, 1998).

In modern terminology, the term CP is sometimes reserved for changes to categorization performance (referring to the tendency for subjects' category judgments to change abruptly near the category boundary), while concomitant changes to discrimination performance are referred to as *acquired distinctiveness* (AD, for improvements in discrimination between categories) or *acquired equivalence* (AE, for

Supported by NIH (NEI) R01 021494. I am grateful to Brooke Ambert, Isaac Bazian, Scott Hoberman, Yoganandasai Koka, Alexandra Lerman, Lijie Lin, and Dip Patel for discussions and assistance in data collection, to Eileen Kowler for helpful discussions.

degradation of discrimination within categories, which is observed in some though not all studies) (Goldstone, 1994; Livingston, Andrews, & Harnad, 1998; Goldstone, Lippa, & Shiffrin, 2001; Notman, Sowden, & Ozgen, 2005; Folstein, Gauthier, & Palmeri, 2010; Folstein, Palmeri, & Gauthier, 2013; Folstein et al., 2014). The neural basis of AD and AE are thought to involve modification of receptive field structure (Sigala & Logothetis, 2002; Kang, Shapley, & Sompolinsky, 2004; Li, Ostwald, Giese, & Kourtzi, 2007; Folstein, Palmeri, Van Gulick, & Gauthier, 2015). However, notwithstanding several neural-network models of CP (Casey & Sowden, 2012; Damper & Harnad, 2000), the computational mechanisms underlying these effects are still poorly understood.

Some studies (e.g. Goldstone & Steyvers, 2001) have concluded that improvement in perceptual discrimination (AD) tends to occur in proportion to each feature's *informativeness* (or *relevance* or *diagnosticity*) about the categories to be learned. But it is not clear which perceptual features the system treats as "informative" and why. In many studies, categories are distinguished by a clear, deterministic boundary separating one class from another—often a linear boundary separating a 2D perceptual space into two clear-cut halves. In such a space, a perceptual feature that crosses the category boundary is perfectly predictive of category membership, making it informative by any reasonable metric, while any feature that does *not* cross the boundary is completely uninformative. But hard classification boundaries are not characteristic of natural categories, which have been understood for decades to have typicality gradients and correspondingly "soft" classification boundaries (Posner & Keele, 1968; Rosch, 1973). When categories are defined more naturalistically via statistical distributions (Huttenlocher, Hedges, & Vevea, 2000), no dimension is perfectly predictive, and a variety of definitions of informativeness are possible. For example, some studies have defined categories as bivariate Gaussian (normal) distributions in a 2D space (e.g. Maye, Werker, & Gerken, 2002; Lake, Vallabha, & McClelland, 2009). In this case, Lake et al. (2009) found that a particular measure of informativeness, the L^2 norm between the posterior distributions, predicted improvements in perceptual discrimination, but this measure was not compared with alternatives.

However classical information theory provides a more natural and well-motivated measure of informativenesss: the *mutual information* (MI), which quantifies how much of the variation in one variable is predicted by another (Cover & Thomas, 1991). MI is widely used in neuroscience (Piasini & Panzeri, 2019), animal learning (Balsam, Fairhurst, & Gallistel, 2006) and machine learning (Battiti, 1994) to quantify informational relationships among variables. The MI between a category variable C and a feature f is defined as

$$\mathrm{MI}(C, f) = H(C) - H(C|f),$$

where $H(C)$ is the prior Shannon uncertainty about the category, and $H(C|f)$ is the conditional uncertainty about the category once the feature is known, both measured in bits if logs are taken in base 2. The MI represents the degree to which learning the value of the feature reduces the observer's uncertainty about which category the stimulus belongs to, and thus constitutes a natural measure of the "informativeness" of the feature. Indeed recently Bates, Lerch, Sims, and Jacobs (2019) showed that features that provide mutual information about a category variable undergo more improvement in discrimination than those that do not, and Bates and Jacobs (in press) provided a comprehensive theoretical argument that the quantity of conveyed information is capped at the MI.

However, the manner in which the system quantifies informativeness cannot be determined using a hard category boundary, as has been used in virtually all studies (including those of Bates et al., 2019). With such a boundary, all of the information (however defined) is concentrated at the boundary, and features that do *not* cross the category boundary convey no information whatsoever about the category variable. This makes it impossible to test intermediate values of informativeness (again, however defined), and moreover completely confounds all reasonable measures of informativeness, because all of them are maximal at the boundary and minimal everywhere else in the feature space. The experiment below solves some of these problems by using probabilistically-defined categories separated by a soft boundary in a novel two-dimensional feature space. The allows for the evaluation of "diagonal" features through the space, in addition to the category-relevant and category-irrelevant axes to which previous studies have been restricted. The resulting experiment includes a whole range of levels of informativeness (rather than just relevant and irrelevant), and also deconfounds various potential measures of informativeness.

Moreover, most studies of CP use features such as color or facial features with which the visual system has enormous prior experience, and which also may have some degree of innate categorical structure (Folstein et al., 2015), making it difficult both to induce changes to perceptual sensitivity and to attribute them directly to training. To more carefully isolate the effect of learning, it is desirable to use a feature space that is as unbiased and unfamiliar to subjects as possible.

The experiments below use a space of randomized, subjectively novel perceptual features with which subjects can be assumed to have little or no prior experience (Fig. 1). Stimuli are drawn from a high-dimensional space of "blob" shapes, created by modulating radial fourier components (Op de Beeck, Wagemans, & Vogels, 2003; Dickinson, Bell, & Badcock, 2013) defining shape contours (Fig. 1a). Shape is very high-dimensional space in which most dimensions involve subtle combinations of contour geometry that are novel and difficult to verbalize (Destler, Singh, & Feldman, 2019). From the initial high-dimensional space, a 2D feature space is randomly selected by choosing 3 random points in the space, which define a random plane, and then randomly choosing an origin and two orthogonal basis vectors in this plane (via the Gram-Schmidt process), which together define a coordinate frame (Fig. 1b). This results in a 2D manifold of shapes from which stimuli are chosen, any subspace of which defines a potential feature (Fig. 1c). Note that unlike many previous studies (e.g. Folstein, Gauthier, & Palmeri, 2012; Folstein et al., 2013; Viviani et al., 2014; Wallraven, Bülthoff, Waterkamp, van Dam, & Gaissert, 2014; Dieciuc, Roque, & Folstein, 2017) this feature space is not a "morph space" constructed by weighted combinations of fixed stimuli at the poles. Rather it is a completely novel space newly novelized (randomized) for each subject. Unlike a morph space there is no familiar or consistent stimulus shape at the poles, and indeed there are no poles, thus reducing the possibility of a preexisting categorical bias present in most previous experiments.

Within the feature space, two categories are defined by circular bivariate Gaussian distributions (Fig. 1d), which define a "soft" linear optimal classification boundary (shown as a dotted line in Fig. 1e). The overlap between the two Gaussians can be modulated by changing their (common) standard deviation σ, which determines the maximum possible proportion correct (ideal performance level or IPL, equal to one minus the Bayes error). IPL was set to 95% in Exps. 1-3 (σ = .150), 90% in Exp. 4 (σ = .188), and 99% in Exp. 5 (σ = .105).

Critically, this procedure defines a feature space that is fully "rotatable" (with a Euclidean L^2 norm) meaning that any direction through this space defines a potential shape feature—including some that might be somewhat verbalizable, but many others that are not (Op de Beeck et al., 2003; Hockema, Blair, & Goldstone, 2005). In contrast feature spaces in most studies (if 2D at all) consist of two separable features (e.g. size and brightness) with only the two cardinal axes as potential features, implying an L1 (citiblock) norm. It is well-established that diagonal dimensions that combine cardinal axes are more difficult to learn than axis-aligned features (Ashby & Maddox, 2011). But in the spaces used here no direction is any more cardinal than any other (and thus no space any more "diagonal" than any other), and moreover the orientation of the subspace is randomized for each subject. Hence in the experiments below exactly which features are informative depends *only* on the category structure chosen. This procedure makes it possible to cleanly assess the informativeness of shape dimensions purely as a function of the category learned, without confounding from the subjects' prior experience.

Experiments

Participants. Subjects were adult members of the undergraduate community (N = 20, 22, 21, 21, and 21 in Exps. 1–5 respectively), recruited from introductory psychology classes and naive to the goals of the experiment.

Discrimination task. Perceptual discrimination was assessed at selected features of interest (FOIs) before and after the categorization task. Each FOI is defined as a point $\mathbf{x} = (x, y)$ and direction $\mathbf{v} = (u, v)$ in the shape space; 5 or 6 such features were evaluated in each experiment (details in Supplementary Table **??** and Fig. 2). To measure discrimination, pairs of shapes (size about 4dva) were presented (white on a dark background, one at a time each for .25sec separated by .5sec ISI, and spatially offset by about 10dva), and the subject was asked to indicate if they were the same or different. Each pair of shapes was located at the desired location in feature space plus or minus a variable discrepancy in the given vector direction, $\mathbf{x} \pm \lambda\mathbf{v}/2$. The featural difference λ was then adaptively reduced on successive trials by the Psi method (Kingdom & Prins, 2010) until the shapes could no longer be distinguished, resulting in an estimate of the threshold of distinguishability at each FOI. Staircases were randomly interleaved. Threshold estimates stabilized in about 15 minutes (about 50-100 trials per feature). Subjects performed the discrimination task before and after the categorization task, providing pre- and post-training estimates of discrimination threshold at each FOI. The main dependent measure is the difference in thresholds pre-training minus post-training at each FOI, Δthreshold.

Categorization task. Stimulus shapes (size about 4dva) were drawn randomly with equal probability from either the A category (a circular bivariate Gaussian centered at $x = .25, y = .5$ of the unit square, see Fig. 2) or the B category (also a circular bivariate Gaussian, centered at $x = .75, y = .5$). The two-Gaussian category structure defines a maximally informative dimension, depicted as horizontal in figures. Stimulus shapes moved downwards from the top of the screen at about 8dva/sec over a starry field, and were visible for maximum of 2.5 seconds or until response. (The motion was intended to draw subjects' attention to the stimulus (Franconeri & Simons, 2003).) The instructions framed the task as a space-based video game (see sample screen in Fig. 1f) in which subjects had to use keyboard buttons to fire at "hostile" ships (category A) or welcome "friendly" ships (B), and received feedback after each response in the form of a happy face (correct classification) or frown (incorrect). Each subject ran 300 trials, taking about 20 minutes, a number that piloting suggested was sufficient to induce measurable CP effects.

Design. Exps. 1-3 all used the same two-Gaussian category structure (95% IPL), differing only in the choice of FOIs (Fig. 2). FOIs were chosen so as to broadly survey the space including a broad range of MI levels (see below), and also

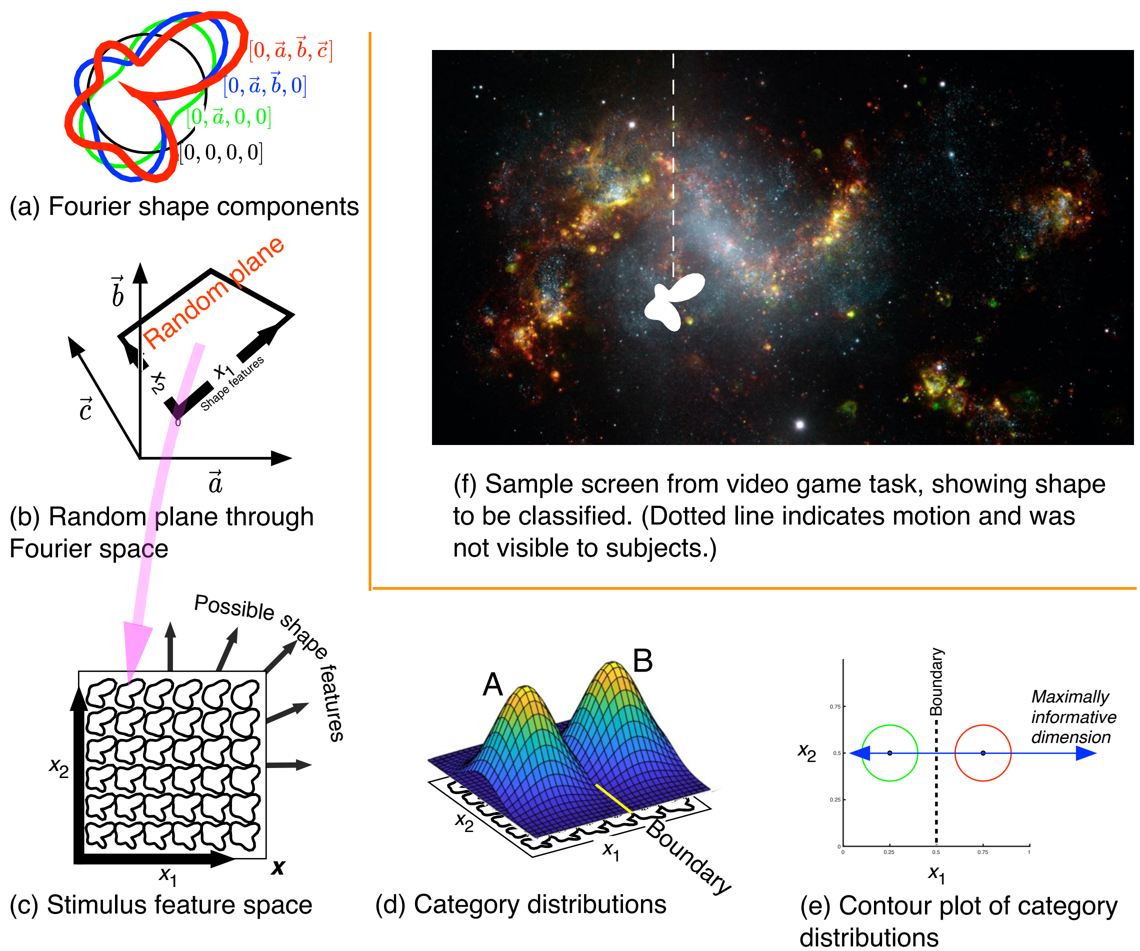

Figure 1. Procedure for creating a novel, randomized perceptual space in which categorization occurs. (a) Shapes are defined by four radial Fourier components (only three depicted). (b) A random plane through this 4D space is chosen. (c) In this plane, a random origin and two orthogonal basis vectors are chosen, resulting in a 2D shape space through which any direction is a potential shape feature. (d) In this space, A and B categories are each circular bivariate Gaussians, respectively centered at (.25,.5) and (.75,.5) of the unit square. (e) Viewed as a contour plot, this category structure defines a linear optimal classification boundary (vertical) with a maximally informative dimension (horizontal). (f) Sample screen from categorization task, showing a shape to be classified.

to target several specific comparisons. Exp. 1 (Fig. 2a) tested six FOIs, including three at the inter-category midpoint (with $\alpha = 0°, 45°$ and $90°$ relative to the maximally informative dimension) and three at the center of category A (at the same three orientations). This comparison is potentially interesting because some studies (e.g. (Folstein et al., 2010)) have suggested that mere exposure to stimuli rather than category training per se is sufficient to induce CP; stimuli near a category center are more frequent, but less diagnostic, than those between categories. Exp. 2 (Fig. 2b) also used 6 FOIs, 3 at the midpoint ($\alpha = 0°, 45°, 90°$) and 3 at a point elsewhere on the optimal classification bound ($\alpha = 0°, 45°, 90°$). Comparing features on and off the main axis is interesting because most studies use a 1D space so all comparisons are necessarily "on axis." Exps. 3–5 investigated the effect of α more finely, using 5 features at the inter-category midpoint ranging from maximally to minimally informative in equal angular steps ($\alpha = 0°, 22.5°, 45°, 67.5°, 90°$). Exp. 3 used the same category structure as Exps. 1 and 2 (95% IPL), while Exp. 4 used a "softer" category boundary (IPL = 90%) and Exp. 5 a "sharper" one (IPL = 99%). The manipulation of IPL does not change the optimal classification boundary, but (as discussed below) it does change the quantity of information available at the FOIs, allowing a more fine-grained evaluation of CP as informativeness is varied. Supplementary Table **??** give a complete list of the FOIs used in all five experiments.

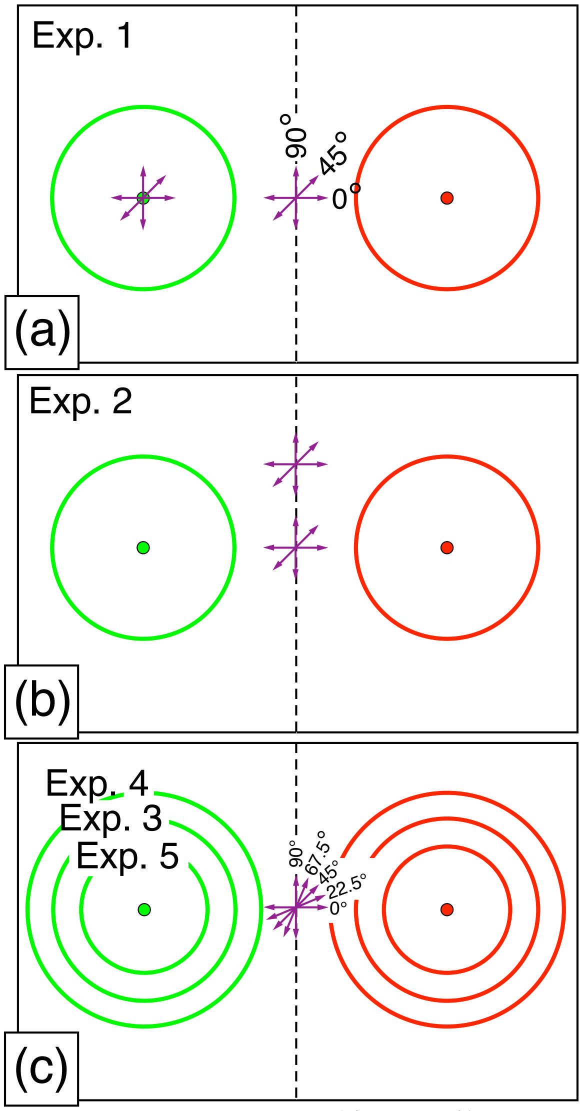

Figure 2. Category structure and features of interest (FOIs) at which discrimination was evaluated in (a) Exp. 1 (b) Exp. 2 and (c) Exps. 3–5. Green and red circles indicate A and B categories respectively (each is a bivariate Gaussian indicated by a circle of radius σ). Each FOI is a point and direction in the 2D perceptual space.

Results

Categorization task. Performance on the categorization task was very variable, with average performance far below the theoretical limit (IPL), presumably reflecting the extremely unfamiliar and nonverbalizable shape features over which categories were defined. Mean performance (s.d.) in Exps. 1-5 were respectively 85% (6%), 83% (6%), 80% (5%), 83% (7%), and 86% (7%). Results reported below include only subjects with overall performance over 70%, which includes 80/105 subjects (76%). Setting the criterion to 50% includes 97/105 subjects (92%), which adds noise to the results but does not affect the main conclusions.

Notwithstanding the subjects' uneven performance, their responses showed clear evidence of the "sharpening" of the category boundary over the course of learning associated with CP. To quantify this, subjects' responses in the categorization task were fitted to a one-dimensional Gaussian classifier of the form $p(A|x) = N(x; \mu_A, \sigma^2)/[N(x; \mu_A, \sigma^2) + N(x; \mu_B, \sigma^2)]$ (i.e., an ideal observer classifier) with the single free parameter σ fitted by least-squares to the data in each block of 50 trials for each subject. In this model, the parameter σ modulates the sharpness of the classification boundary, with high values of σ indicating broad category distributions and a more gradual transition between categories, and low values indicating narrower category distributions and a more abrupt transition. Fig. 3 shows plots of the progression of subjects' mean estimated sigmas over the course of training in each experiment. In all of the plots sigmas start high and progressively decrease (sharpen), gradually approaching their respective target values (i.e. those from which the stimuli were actually generated). The change from broader to narrower sigmas from the first block of the experiment to the last was statistically substantial (BF > 3) in all 5 experiments. In these plots, the classification curve is approximately linear in the first block—meaning that the classification probability changes in approximately equal increments with each step through the feature space, that is, completely non-categorically. By the last block, the fitted values of sigma are such that the classification is a relatively sharp step near the boundary, in the classical pattern associated with CP. Note though that since sigmas seem to be asymptoting near their "true" values (that is, the values used to generate the stimuli), this increasingly categorical performance simply seems to reflect approximately optimal category learning.

Discrimination task. Discrimination improved substantially ($BF_{10} > 3$ pre- vs. post) in all 20 of the FOIs at which α was less than 90° (and thus the feature was diagnostic at all), but not in the other seven FOIs ($BF_{10} < 3$). That is, subjects became more sensitive to those features—and only those features—that were predictive of category membership. The subjects demonstrated acquired distinctiveness after only about 20 minutes of category training, in contrast to thousands of trials of training in many studies. This unusually rapid induction of AD presumably reflects the novel feature space, whose unusual initial difficulty allowed subjects to improve rapidly with training. The finding of AD with integral shape dimensions contrasts with the study of Op de Beeck et al. (2003) (though see Hockema et al., 2005). Overall thresholds decreased from a mean (s.d.) of .35 (0.007) to 0.26 (0.009) after training (recall that the category means were separated by .5 of the unit square). The magnitude of discrimination improvement (Δthreshold) was not correlated with performance on the categorization task ($R^2 = 0.00084$,

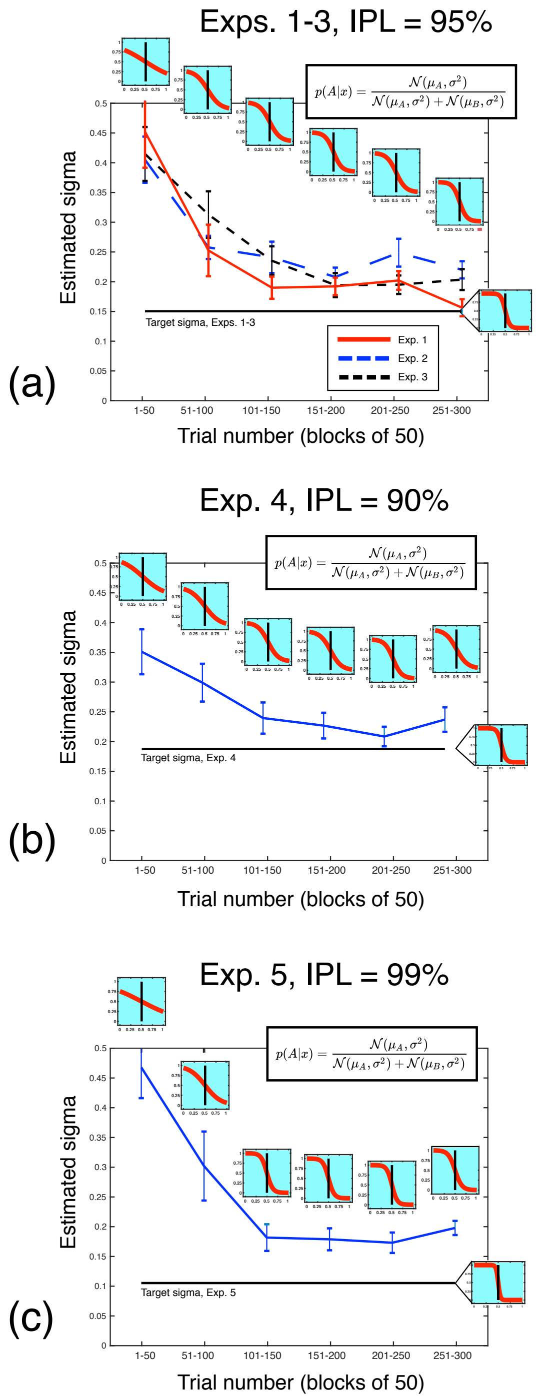

Figure 3. Plots of fitted values of σ in (a) Exps. 1-3 (95% IPL) (b) Exp. 4 (90% IPL) and (c) Exp. 5 (99% IPL). Each subject's responses in the categorization task were fitted to the ideal classification curve $p(A|x) = N(x; \mu_A, \sigma^2)/[N(x; \mu_A, \sigma^2) + N(x; \mu_B, \sigma^2)]$, with μ_A and μ_B set to their true values of .25 and .75 respectively and σ fitted by least squares to the subject's responses. The fitted value of σ modulates how broad or narrow the subjects' induced category is, with larger values entailing a "softer" decision boundary and smaller values entailing a "sharper" one (visualized in insets). As can be seen in the plots, the subjects generally progressed over the course of training from softer boundaries to sharper ones.

$\mathrm{BF}_{10} = 0.09$). The manipulated position factors had relatively small effects in individual experiments (Fig. 4). However a clear pattern emerges when the results of all five experiments are combined, as follows.

The main analysis is the magnitude of improvement in discrimination (Δthreshold) as a function of the mutual information $\mathrm{MI}(C, f) = H(C) - H(C|f)$ between the category variable C (= A or B) and a given FOI f. $H(C) = -p(A)\log_2 p(A) - p(B)\log_2 p(B)$ is the prior uncertainty about the category, which in the experiments is always 1 bit because the two categories are equally likely. $H(C|f) = -p(f)\log_2(A|f) - p(f)\log_2(B|f) - p(\neg f)\log_2(A|\neg f) - p(\neg f)\log_2(B|\neg f)$ is the conditional uncertainty about the category once the feature is known. Intuitively, each feature f can be thought of as a binary division of the perceptual space into two halves (Fig. 5a); $\mathrm{MI}(C, f)$ measures how much information an observer gains about C (which category a given stimulus belongs to) from learning which "half" of f it falls in.

In this sense MI quantifies the diagnosticity of a given stimulus property with respect to the shape's category membership.

MI is maximal for "horizontal" ($\alpha = 0°$) features lying on the classification boundary, but its value there is affected by the sharpness of the boundary, modulated in the experiments by the IPL. For example the MI for such features in Exps. 1–3 (95% IPL) is .71 bits, in Exp. 4 (90% IPL) .53 bits, and in Exp. 5 (99% IPL) is .92 bits. Features with $\alpha = 90°$ (perpendicular to the classification boundary) have MI = 0; they are completely "uninformative." Between these extremes MI depends in a more complex way on both the feature's position and orientation.

Fig. 6 show Δthreshold as a function of MI for all 27 FOIs aggregating across Exps. 1–5 (with different colors/symbols for each experiment). The plot shows a clear linear relationship: the magnitude of AD (Δthreshold) rises with MI ($R^2 = 0.5663, \mathrm{BF}_{10} = 3,697$). The linear relationship between MI and Δthreshold in individual experiments was respectively $R^2 = 0.48, 0.71, 0.89, 0.006$, and 0.94, suggesting that the effect is robust and replicable. As discussed above, previous papers have found that AD is larger in category-relevant features than in category-irrelevant ones. The current results show that, as Bates and Jacobs (in press) suggested, the degree of AD at each feature is proportional to its informativeness, as measured by the magnitude of mutual information it shares with the category variable.

Changes to discrimination performance observed were all positive, indicating "acquired distinctiveness" rather than "acquired equivalence." The regression intercept of .0550 gives the magnitude of AD at MI = 0, i.e. an overall practice effect. On the basis of information-theoretic constraints, Bates et al. (Bates et al., 2019; Bates & Jacobs, in press) have argued that if channel capacity is fixed then improvements to representational precision in some dimensions need to be

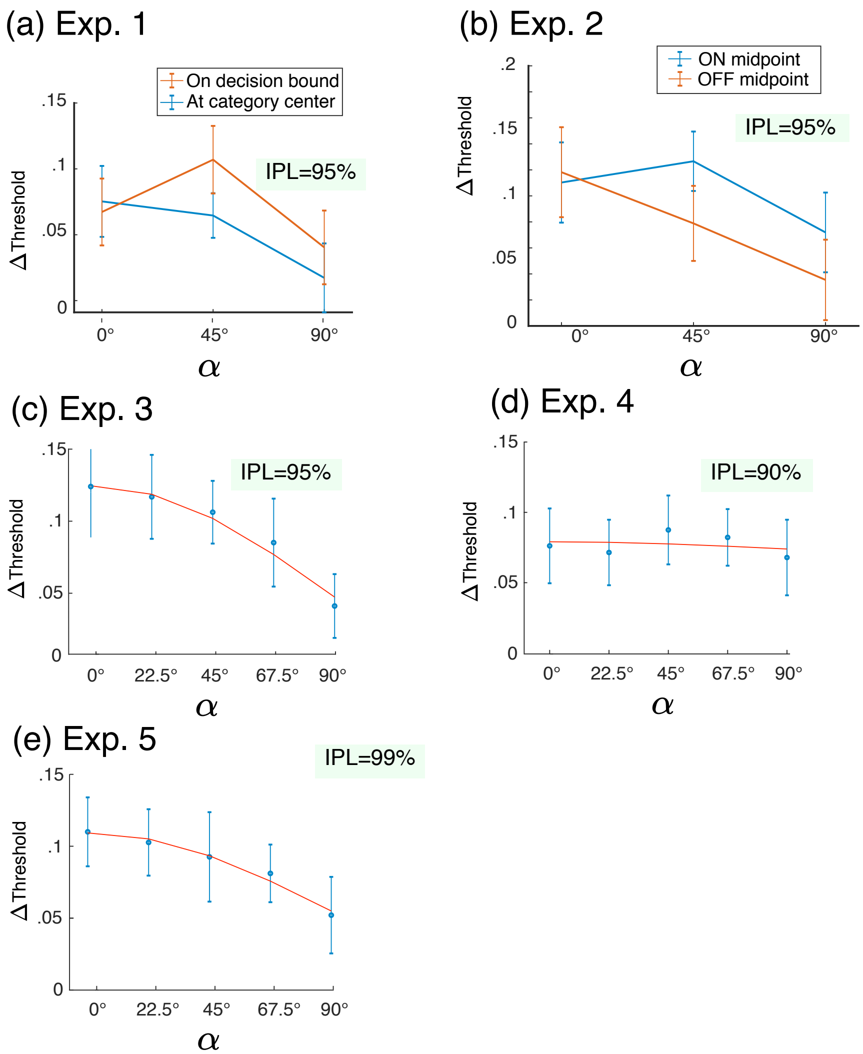

Figure 4. Results for Exps. 1–5, showing Δthreshold as a function of α (angular deviation from most informative dimension) and position in the feature space (Exps. 1 and 2). Panels c-e show best fitting cosine function.

offset by degradations in others. No such effect is apparent in the current data, as all discrimination changes were in the same direction. It it is possible that such tradeoffs may have been swamped by an overall practice effect, meaning that the total channel capacity allocated to featural representation may have increased over the course of training. Unfortunately the current data do not allow this issue to be addressed more decisively.

Given the particular categories and features used in these experiments, most of the variation in MI (about 85%) is due to α, while the rest is due to feature position and IPL. Hence it is fair to wonder whether the effect of MI on AD might in fact be entirely due to α rather than MI per se. However a regression of Δthreshold onto α alone is less predictive than MI ($R^2 = 0.4257$ compared to $R^2 = 0.5663$ for MI; the difference in fits is statistically substantial, $BF_{10} = 44.4$). Moreover, a Bayesian ANOVA on the entire dataset favors (maximum posterior) the additive model that includes all three factors (α, feature position, and IPL) over any subset model ($BF_{10} = 11.820$). Thus the effect of MI appears to depend on all three component factors, and in particular is not attributable to α alone, although the contribution of feature position and IPL is relatively subtle and should be more comprehensively explored in future experiments.

As mentioned, several other definition of informativeness have been proposed, including the L^2 norm between the posterior distributions (Lake et al., 2009), and the (squared) derivative of the posterior, which reflects the sharpness of the category boundary (Clayards, Tanenhaus, Aslin, & Jacobs, 2008). The squared posterior derivative is related to the MI and Fisher information, and plays an important role in the theoretical literature on neural population coding (Pouget &

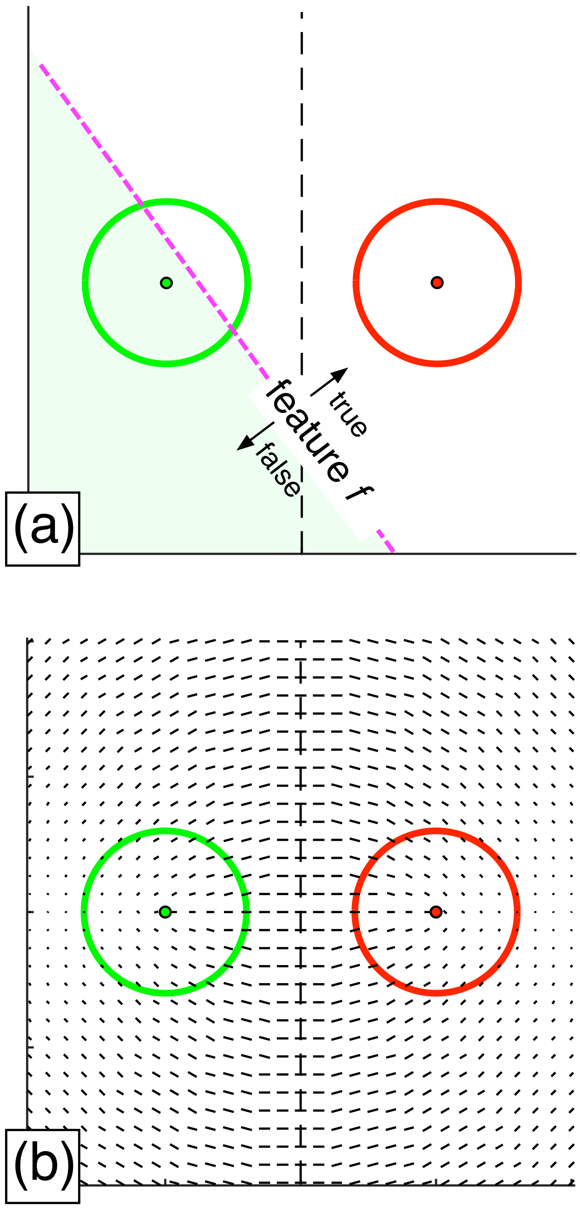

Figure 5. (a) Calculation of MI at an arbitrary feature f. The MI is the shared information between the feature f (which side of the boundary defined by f does the stimulus fall on?) and the category C (which category does it belong to?) (b) A map of MI as induced by the category structure used in Exps.1-3, showing the magnitude and direction of maximum MI at each point in the space.

Zemel, 2007; Bonnasse-Gahot & Nadal, 2008). However in the current data the the L^2 norm predicts Δthreshold less well than does MI ($R^2 = 0.45$, worse than the fit for MI by $BF_{10} = 24.3$) as does the squared posterior derivative ($R^2 = .25$, worse than the fit for MI by BF = 1,626). Hence in addition to MI's more natural axiomatic derivation as a measure of the information conveyed by one variable about another, MI gives a better fit to the human data.

Discussion

Several studies have found that feature discrimination tends to improve more for features that are informative about learned categories than for those that are not (e.g. Goldstone & Steyvers, 2001; Folstein et al., 2013, 2014; Bates et al., 2019). The results of Exps. 1–5 show that "informativeness" can be quantified by mutual information: the more information a feature conveys about the category, in a classical Shannon sense, the more subjects (on average) tend to gain in sensitivity at that feature. This improvement in discrimination (AD) is directly attributable to category training, and is associated with the progressive development of sharper category boundaries over the course of training (CP). The effect is better predicted by MI than it is by other measures of informativeness, such as the posterior slope, the posterior L^2 norm, or the orientation of the feature with respect to the maximally informative dimension. Overall, these results corroborate the role of information theory in quantifying how the brain allocates representational resources (Balsam et al., 2006; Nelson, McKenzie, Cottrell, & Sejnowski, 2010; Sims, 2018), and suggest that such allocation is rationally tuned to the category structure of the world (Lake et al., 2009; Maye et al., 2002; Feldman, Griffiths, & Morgan, 2009; Soto & Ashby, 2015; Bates & Jacobs, in press).

One notable consequence of these results is to deemphasize the division between features that cross category boundaries and those that do not, which is often highlighted in definitions of CP. In the MI account features that cross the category boundary are the *most informative*, but are not qualitatively different from other features in the space that convey information about the category albeit to lesser degrees. This observation helps explain a variety of previous results, e.g. those of Goldstone (1994), who found that discrimination improvement was not limited to the category boundary, but was distributed throughout the space in a somewhat complex pattern. Note that this pattern is not consistent with traditional attention-weighting models (e.g. Kruschke, 1992), which elevate or attenuate entire perceptual dimensions rather than specific feature values. As mentioned, this pattern cannot be clearly established using a hard category boundary, where boundary-crossing features are the *only* informative ones; nor in a 1D perceptual space, where all features lie along the (sole) informative dimension. The clear relationship between MI and AD only becomes apparent with statistically-defined category structure over at least two dimensions.

Dieciuc et al. (2017) have suggested that some feature learning can be explained by relatively short-term reallocation of attention. The current experiments cannot address the time-course of the observed changes to perceptual dis-

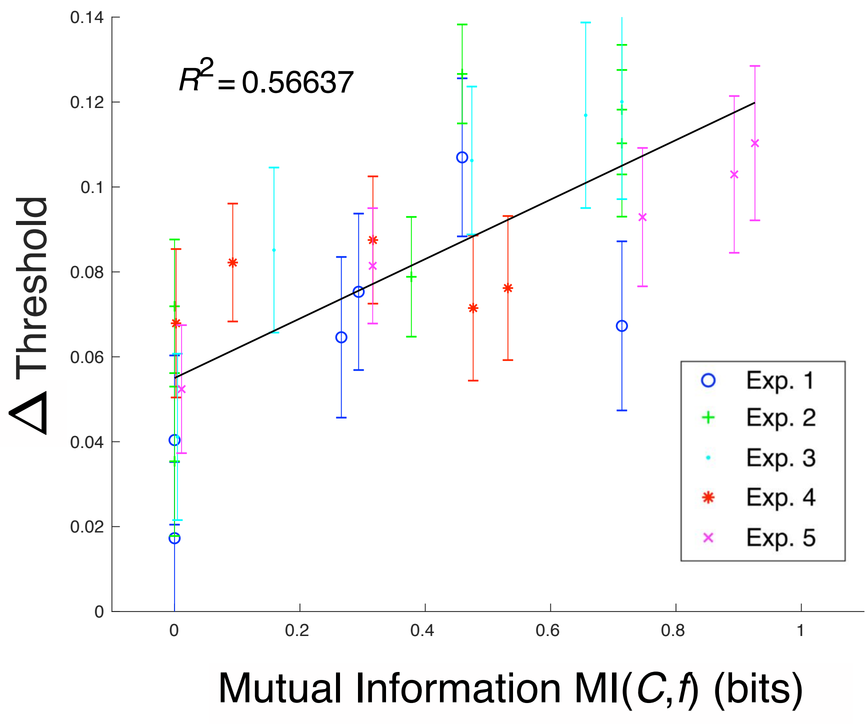

Figure 6. Improvement in discrimination (Δthreshold) as a function of MI across FOIs tested, showing linear regression (Δthreshold = .0701 × MI + .0550). Each plotted point averages over the subjects included in that condition (N = 20 or 21), with different colors/symbols for different experiments (see legend). Units on the ordinate are proportions of the feature space (in which categories were separated by .5). Error bars are ±1 s.e.

crimination, since discrimination was only tested in the immediate aftermath of category training. Note however that allocation of *spatial* attention cannot explain these results, since the shape features tested were all global aspects of each stimulus shape, and could not be localized to any one location within it. The results might however reflect the reallocation of *feature-based* attention (Maunsell & Treue, 2006). But note that these shape features represented novel, complex combinations of shape contour features, and thus could not be evaluated simply by reallocating resources within an existing feature space. Hence while these data do not directly address the role of attention, it seems difficult to explain the observed improvements in discrimination by reallocation of feature-based attention alone. Future experiments evaluating the durability of these discrimination changes would be very valuable.

Conclusion

The results reported here suggest that feature learning is rationally tuned to the statistical structure of the environment (Lake et al., 2009; Maye et al., 2002; Feldman et al., 2009; Soto & Ashby, 2015; Bates & Jacobs, in press), and support a principled information-theoretic quantification of the way representational resources are allocated. More specifically, the new finding supports previous arguments (Harnad, 1993; Schyns et al., 1998) that CP reflects the process by which the brain constructs a "vocabulary" of features suitable for representing the world.

Important questions for future studies include how to apply the MI measure to unsupervised categorization, in which the category variable C is not directly available to the subject. In unsupervised learning, which is ubiquitous in everyday cognition, MI might be computed between features and an estimated latent category variable (Lake, Salakhutdinov, & Tenenbaum, 2015). Another important question is whether the relationship between AD and MI extends to more complex conceptual structures such as multimodal categories (Briscoe & Feldman, 2011), in which the MI "map" can become much more complex.

References

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Science*, *1224*, 147–161.

Balsam, P. D., Fairhurst, S., & Gallistel, C. R. (2006). Pavlovian contingencies and temporal information. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(3), 284–294.

Bates, C. J., & Jacobs, R. A. (in press). Efficient data compression in perception and perceptual memory. *Psychological Review*.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

Bonnasse-Gahot, L., & Nadal, J. P. (2008). Neural coding of categories: information efficiency and optimal population codes. *Journal of Computional Neuroscince*, *25*(1), 169–187.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*, 2–16.

Casey, M. C., & Sowden, P. T. (2012). Modeling learned categorical perception in human vision. *Psychological Bulletin*, *33*, 114–126.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley.

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics*, *62*(4), 843–867.

Destler, N., Singh, M., & Feldman, J. (2019). Shape discrimination along morph-spaces. *Vision Research*, *158*, 189–199.

Dickinson, J. E., Bell, J., & Badcock, D. R. (2013). Near their thresholds for detection, shapes are discriminated by the angular separation of their corners. *PLoS ONE*, *8*(5), e66015.

Dieciuc, M., Roque, N. A., & Folstein, J. R. (2017). Changing similarity: Stable and flexible modulations of psychological dimensions. *Brain Research*, *1670*, 208–219.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology*, *1*, 1–6.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: not all morphspaces stretch alike. *J Exp Psychol Learn Mem Cogn*, *38*(4), 807–820.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: the case of shape and motion. *Frontiers in Psychology*, *5*(1394), 1–7.

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current Directions in Psychological Science*, *24*(1), 17–23.

Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & Psychophysics*, *65*(7), 999–1010.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, *20*(3), 507–523.

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

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology*, *130*(1), 116–139.

Harnad, S. (1987). *Categorical perception: the groundwork of cognition*. Cambridge: Cambridge University Press.

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think*, *2*, 57–62.

Hockema, S. A., Blair, M. R., & Goldstone, R. L. (2005). Differentiation for novel dimensions. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 953–958). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Kang, K., Shapley, R. M., & Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *Journal of Neuroscience*, *24*(15), 3726–3735.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction*. London: Academic Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Vallabha, G. K., & McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 35–43.

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *Journal of Neuroscience*, *27*(45), 12321–12330.

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*(5), 358–368.

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*(3), 732–753.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neuroscience*, *29*(6), 317–322.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101–111.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.

Notman, L. A., Sowden, P. T., & Ozgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition*, *95*(2), 1–14.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, *132*(4), 491–511.

Piasini, E., & Panzeri, S. (2019). Information theory in neuroscience. *Entropy*, *21*, 62.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pouget, A., & Zemel, R. S. (2007). Population codes. In K. Doya, K. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding.* Cambridge, MA: MIT Press.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosielle, L. J., & Cooper, E. E. (2001). Categorical perception of relative orientation in visual object recognition. *Memory & Cognition*, *29*(1), 68–82.

Rotshtein, P., Henson, R. N., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*(1), 107–113.

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

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Soto, F. A., & Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition*, *139*, 105–129.

Viviani, P., Binda, P., & Borsato, T. (2014). Categorical perception of newly learned faces. *Visual Cognition*, *15*(4), 420–467.

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaissert, N. (2014). The eyes grasp, the hands see: metric category knowledge transfers between vision and touch. *Psychonomic Bulletin & Review*, *21*(4), 976–985.