Exploring the Cost Function in Color Perception and Memory: An Information-Theoretic Model of Categorical Effects in Color Matching

Chris R. Sims (chris.sims@drexel.edu)
Applied Cognitive & Brain Sciences, Department of Psychology, Drexel University

Zheng Ma (zma4@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Sarah R. Allred (srallred@scarletmail.rutgers.edu)
Department of Psychology, Rutgers University–Camden

Rachel A. Lerch (rachel.a.lerch@drexel.edu)
Applied Cognitive & Brain Sciences, Department of Psychology, Drexel University

Jonathan I. Flombaum (flombaum@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Abstract

Recent evidence indicates that color categories can exert a strong influence over color matching in both perception and memory. We explore this phenomenon by analyzing the cost function for perceptual error. Our analysis is developed within the mathematical framework of rate–distortion theory. According to our approach, the goal of perception is to minimize the expected cost of error while subject to a strong constraint on the capacity of perceptual processing. We propose that the cost function in color perception is defined by the sum of two components: a metric cost associated with the magnitude of error in color space, and a cost associated with perceptual errors that cross color category boundaries. A computational model embodying this assumption is shown to produce an excellent fit to empirical data. The results generally suggest that what appear as 'errors' in working memory performance may reflect reasonable and systematic behaviors in the context of costs.

Keywords: color perception; visual working memory; information theory; rate–distortion theory

Visual working memory is central to daily life. Even extremely simple tasks, such as visually comparing the size or color of two objects, requires the storage and manipulation of perceptual information in working memory. Given the central role of visual working memory in natural tasks, it is quite surprising that this system is also quite limited. Previous studies have demonstrated that the capacity of visual working memory for simple unitary features such as color or orientation is on the order of 2–4 bits (Sims, Jacobs, & Knill, 2012; Sims, 2015). With such a strong constraint on information processing, it seems especially paramount that the brain use its available working memory capacity in an efficient manner. But what defines an 'efficient' perceptual system?

A natural and intuitive answer is that visual working memory is used efficiently when it minimizes task-relevant costs and errors. According to this perspective, the key to understanding perceptual processing is identifying the particular *cost function* that it seeks to minimize. Abstractly, if a sensory signal x is misperceived or misremembered as a different signal y, then there exists some subjective cost (or disutility) associated with this error, and this can be quantified by some function $\mathcal{L}(x,y)$. An efficient perceptual system is one that minimizes the expected cost according to a particular cost function, while subject to a constraint on the capacity of the perceptual channel. Recent work (Sims et al., 2012; Sims, 2015) has demonstrated that this problem statement corresponds quite naturally to a branch of information theory known as rate–distortion theory (Berger, 1971). Rate–distortion theory concerns the optimal solution to the problem of minimizing the costs of communication error, subject to constraints on available capacity.

The goal of the present paper is to apply rate–distortion theory in order to identify the cost function that drives color matching in perception and memory. For example, if a particular shade of red is misperceived or misremembered as a slightly different shade of red, how costly is that error to the brain? Although this may seem like a trivial question, color perception–even in simple laboratory contexts–exhibits many subtle properties that are not completely understood (Allred & Flombaum, 2014). Of particular relevance is the finding that categories can strongly influence perception and memory (Huttenlocher, Hedges, & Vevea, 2000; Bae, Olkkonen, Allred, & Flombaum, 2015). The question considered in this paper is how color categories influence both color perception and color memory, as formalized within the mathematical framework of rate–distortion theory.

Bae and colleagues (2015) reported a series of four experiments examining how categories influence color perception and memory, and also developed a computational model to account for their results. According to their model (referred to as CATMET), colors are encoded in two separate channels in perceptual processing: one channel encodes a category-based representation, while the other encodes a metric-based representation of color on a continuous scale. Color perception results from the heuristic integration of these two channels to form an estimate of the afferent sensory signal.

The current paper seeks to explain these same experimental data, but using an alternative modeling approach based on rate–distortion theory. The goal of this effort is not to supplant the CATMET model. Rather, rate–distortion theory can offer an explanation at Marr's computational level of anal-

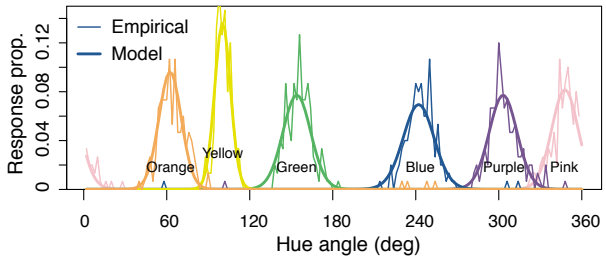

Figure 1: Empirical data and model fit for the color identification experiment conducted by Bae et al. (2015).

ysis (Marr, 1982). Whereas the CATMET model provides a process-level explanation for *how* categories influence perception, the goal is to propose an explanation as to *why* categories influence perception. The model developed in this paper represents a preliminary step in that direction.

To us, it is intuitive and plausible that there is a cost to committing categorical errors in perception. Consider the task of selecting fruit based on ripeness. Both ripe and unripe fruit may encompass a wide range of hues, but for the purpose of finding the best one to eat, perceptual errors that do not cross this category boundary have little practical implication. Much more nutrition is provided by yellow bananas and red berries than green bananas and berries; hence it seems plausible that perception should be sensitive to category boundaries.

In a nutshell, we propose that color perception is the result of the rational minimization of a particular cost function, subject to a constraint on capacity. In the model we will describe, the cost function driving perception is the sum of a metric error term (the distance between a stimulus and its perceived value in color space) as well as an additional cost when perceptual errors cross color category boundaries.

Before describing the mathematical details of the model, we first introduce the experimental results to be explained.

Experimental results

Bae et al. (2015) conducted four experiments examining the relationship between color categories, and color matching in perception and memory. In the current paper we restrict our attention to three of these datasets: color identification, undelayed estimation, and delayed estimation.

In the *color identification* experiment, subjects were presented with a color wheel containing 180 equiluminant colors varying only in hue, along with 6 color category labels (pink, orange, yellow, green, blue, purple). Subjects were asked to simply click on a point along the color wheel to indicate the best example for each of the six color categories. The results from this experiment are shown in Figure 1. Subjects were highly consistent in their identification of color categories.

In the *undelayed estimation* experiment, subjects were presented with a color patch as well as a color wheel from which the color was sampled. The task for the subject was to choose a point along the color wheel to indicate the best color match for the given probe. This procedure was repeated using 180 different color targets, collecting a large number of trials per subject. For complete experimental methods the reader is directed to the source publication (Bae et al., 2015). Note that since the color patch and color wheel remained visible throughout the duration of each trial, it would seemingly be an easy task for subjects to click on the exact matching color on each trial. The *delayed estimation* experiment was methodologically similar, except that in this experiment the color patch disappeared during the response portion of the trial and subjects had to rely on a memory representation of the probe color.

The important results from the undelayed and delayed estimation experiments are illustrated in Figure 2. Figure 2a shows the overall histogram of responses grouped into 180 equal-width bins. Although probe stimuli were sampled uniformly along the color wheel, responses are clearly nonuniformly distributed. The magnitude of this effect is greater in the delayed experiment. Figure 2b shows the mean bias observed for each of the 180 probe stimuli. Positive values indicate responses that were clockwise, on average, relative to the probe stimulus. The bottom panel shows the circular standard deviation of the response distribution. The figure shows that response variability also varied systematically across hues.

Figure 3a gives an overhead view of the both datasets in their entirety, showing the conditional response distribution for each of the 180 stimuli used in the estimation experiments. An unbiased perceptual system would exhibit a straight diagonal line; in contrast, human performance shows a consistent pattern of distortion (bias) and variability.

The empirical results summarized in Figures 2 and 3 demonstrate unquestionably that color perception and color memory both show strong stimulus-specific properties (see also Allred & Flombaum, 2014). Bae and colleagues (2015) previously developed a model that could produce these effects; our goal is to model the effects specifically as the consequence of a cost function, framed by rate-distortion theory.

An information-theoretic model of color perception and color memory

Rate–distortion theory concerns the optimal solution to minimizing costs according to a particular cost function, subject to a constraint on channel capacity. We assume that color perception can be modeled as a communication channel where some input signal x is perceived or remembered as a possibly different signal y. In the experiments under consideration, stimuli are color hues drawn from a circular color wheel; hence x and y can be considered as the angle of a given stimulus, and the response angle around this color wheel. Our model assumes a cost function consisting of two terms: a metric cost related to the angular difference between x and y, and a categorical cost that is based on the probability that x and y would be assigned different color category labels. Hence,

$$\mathcal{L}(x,y) = f(y-x) + P(C_x \neq C_y). \quad (1)$$

The first term represents the metric cost of error. Potential

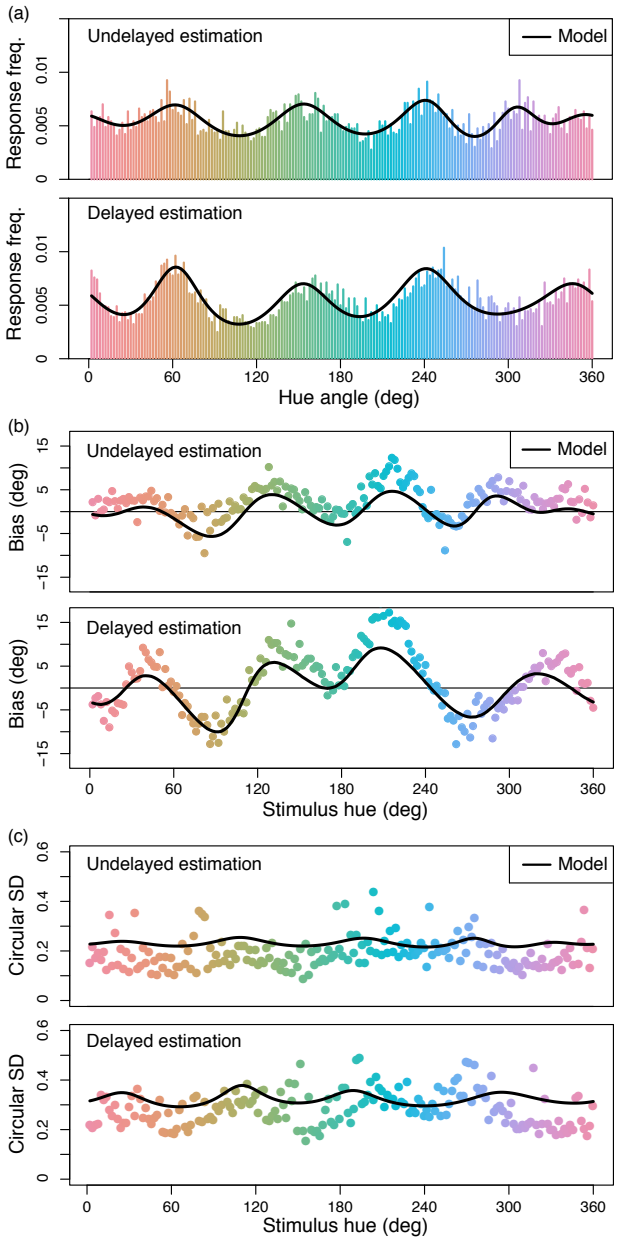

Figure 2: Results from the estimation experiments conducted by (Bae et al., 2015). (a) Response frequency, demonstrating systematic biases in color perception and memory. (b) The magnitude of bias for each stimulus hue. (c) The circular standard deviation in the response distribution for each stimulus hue. In all panels, empirical data is given by colored markers/lines while the black curve shows the information-theoretic model fit to the data.

candidates for this function include the squared error, $(y-x)^2$ or absolute error $|\,y-x\,|$. However, since x and y are points in a circular space we first assume a metric cost function based on the cosine of the difference between x and y:

$$f(y-x) = \frac{1}{2}\left[1-\cos(y-x)\right]. \tag{2}$$

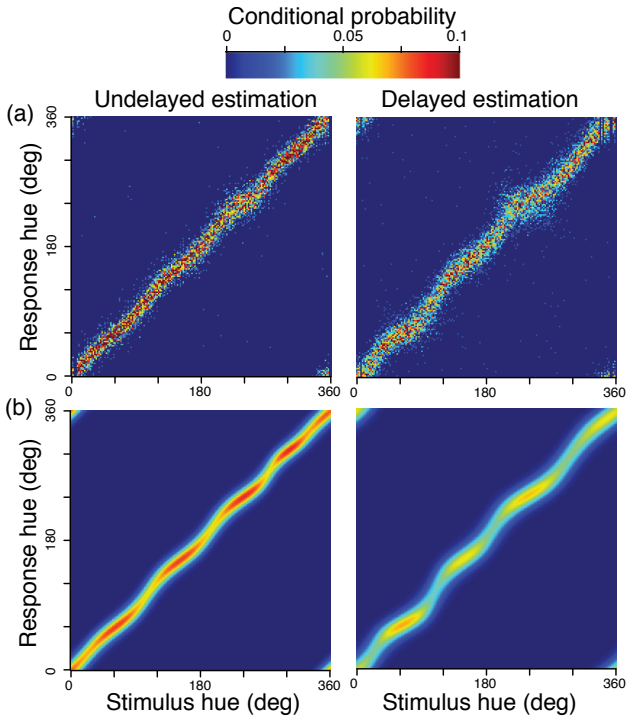

Figure 3: (a) Conditional probability distributions showing the response distribution for each of the 180 stimuli used in the undelayed (left panel) and delayed (right panel) estimation experiments. (b) Model fit to the experimental data.

This function is zero when $x = y$, and reaches a maximum of 1 when x and y differ by the maximum of 180 degrees. The second term in the cost function given by Eq 1 indicates the probability that x and y would be assigned different category labels, indicated by C_x and C_y. Specifying this requires a model of how color hues are mapped on to color categories. Our model assumes that different hues in color space are better or poorer examples of each color category. The strength or 'goodness' of a given hue for a particular color category k is modeled as a scaled Von Mises distribution:

$$\phi_k(x) = \frac{\alpha_k}{e^{\tau_k}} \exp\left(\tau_k \cos(x-\mu_k)\right). \tag{3}$$

Hence, each color category is described by three parameters: its central location (μ), the category precision (τ, the inverse of width), and the maximum strength of the color category (α). If there are K categories, then a given hue angle corresponds to a vector of category strengths, $\langle \phi_1(x), \phi_2(x), \ldots, \phi_K(x) \rangle$. The probability that the hue x is assigned to category k is modeled using the softmax equation:

$$P(C_x = k) = \frac{\exp(\phi_k(x))}{\sum_{j=1}^{K} \exp(\phi_j(x))}. \tag{4}$$

This process allows for noise in color category assignment: when all $\alpha_k = 0$, category assignment is performed at chance. With this generative model of color category assignment, the probability that two hues x and y are assigned different labels

is equal to one minus the probability that they are given the same label:

$$P(C_x \neq C_y) = 1 - \sum_{k=1}^{K} P(C_x = k)P(C_y = k). \quad (5)$$

Equations 1–5 fully describe the cost function $\mathcal{L}(x,y)$ that we assume the perceptual channel seeks to minimize. Specifically, the goal is to minimize the expected cost, subject to a constraint on available channel capacity,

$$\begin{aligned} &\text{Minimize } \mathrm{E}[\mathcal{L}(x,y)] \text{ w.r.t. } p(y \mid x) \\ &\qquad = \sum_x \sum_y \mathcal{L}(x,y)p(y \mid x)p(x) \\ &\text{Subject to } \mathrm{I}(x,y) \leq \mathcal{C}, \end{aligned} \quad (6)$$

where the minimization is performed over the space of conditional probability distribution $p(y \mid x)$ (this conditional probability distribution specifies the probability that the channel produces an output y for a given input x). The second line of this equation specifies that the mutual information $I(x,y)$ must be less than a specified channel capacity $\mathcal{C}$. Equation 6 represents a standard poblem statement in rate–distortion theory (Berger, 1971). For readers unfamiliar with information theory, in the present case it only matters that the solution to this constrained optimization problem represents an information channel that minimizes expected cost according to a given cost function, subject to a specified constraint on channel capacity. Our model obtains a solution to this equation using an efficient numerical algorithm due to Blahut (1972).

With a cost function specified, it is possible to examine the predictions of the model by comparing the distribution $p(y \mid x)$ (the distribution of channel outputs for a given stimulus input) against empirical data. However, the model just described requires the specification of three parameters for each color category (μ_k, τ_k, α_k), along with a constraint on available memory capacity ($\mathcal{C}$).

Our approach is to estimate the mean (μ_k) and precision (τ_k) of each color category using data from the color identification experiment (Figure 1). To do so requires one additional modeling assumption. If a color category is defined by its central location μ and precision τ, it is necessary to describe how a single color is selected from this category as the best example of the category. Our current model assumes that color identification is also based on the softmax equation. Now however, rather than selecting between categories, the goal is to select a hue that best represents a particular category k. Mathematically, this is stated as

$$P(\theta \mid k) = \frac{\exp(\beta \cdot \phi_k(\theta))}{\sum_\Theta \exp(\beta \cdot \phi_k(\Theta))}. \quad (7)$$

This introduces one additional parameter, β, which controls the noise in color identification (as $\beta \to \infty$, the model deterministically selects the peak of the color category, μ_k).

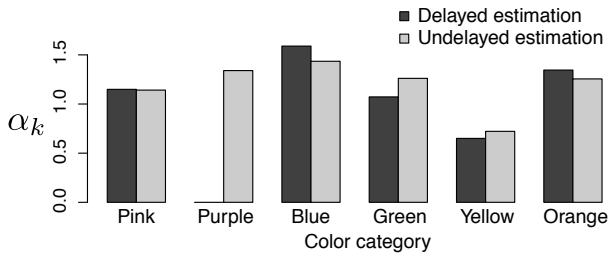

Figure 4: Color category strengths, α_k, fit to the undelayed and delayed estimation experiments.

In the model a single parameter β is used for all color categories. In summary, the parameters μ_k, τ_k and β were estimated from the color identification experiment. The best fitting parameters were determined by maximum likelihood estimation using numerical optimization. As shown in Figure 1, this model produces a close fit to the data.

With these parameters fixed, the category strengths (α_k) and capacity $\mathcal{C}$ were fit to the undelayed and delayed estimation experiments via maximum likelihood estimation. Model predictions are based directly on the optimal channel distribution $p(y \mid x)$ obtained from Equation 6. Corresponding model fits are shown in Figures 2 and 3. In terms of parameter estimates, the primary difference between the undelayed and delayed conditions is the estimated memory capacity. For the undelayed condition, channel capacity was estimated as 3.02 bits; for the delayed condition estimated capacity was 2.60 bits. These estimates are well in line with previous analyses of visual working memory capacity (Sims et al., 2012; Sims, 2015). What is notable is that in the undelayed experiment, capacity is still strongly limited even while stimuli remain continuously visible throughout each trial. This underscores the fact that perception has limited channel capacity in an information sense. In other words, there is always uncertainty in the interpretation of sensory signals.

The color category strengths, α_k are illustrated in Figure 4. Recall that these values determine the maximum strength of each color category, which in turn influences the probability of category assignment via Equation 4. The parameter estimates from the two experiments are highly similar, with one notable exception: in the delayed estimation condition of the experiment, the purple color category exhibits no influence over perception. This can also be seen in the histograms in Figure 2a. The reason for this difference between conditions is not clear; perhaps the influence of categories on perception shows strong individual differences, or perhaps the difference is due to the memory retention interval imposed by the delayed estimation experiment. Exploring this question remains a topic for future research.

Figure 5 shows the estimated cost functions for the undelayed and delayed experiments. The function $\mathcal{L}(x,y)$ is visualized as a two-dimensional heat map, where colors correspond to the cost of a particular perceptual error. Inter-

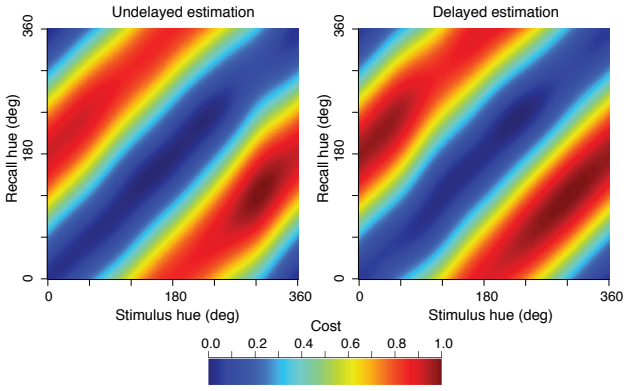

Figure 5: Cost functions estimated from the undelayed and delayed estimation experiments. The plots are re-scaled so that costs lie in the range (0, 1).

Comparison	Model	ΔBIC Delayed	ΔBIC Undelayed
(a)	Category + Metric	**0**	**0**
	Metric only	1159.71	294.92
	Category only	10682.77	7697.35
(b)	Cosine cost	151.94	246.93
	Parameterized cost	**0**	**0**

Table 1: Model comparisons based on Bayesian Information Criterion (BIC) scores, for the delayed and undelayed estimation experiments. Models with the lowest relative score (ΔBIC) are favored, indicated with bold type. Comparisons: (a) Examining whether category and metric costs of error are both necessary to account for performance. (b) Comparing the choice of alternative metric cost functions.

estingly, the cost function is highly similar between the two experiments; the change in performance is almost entirely due to the decrease in channel capacity in the delayed estimation experiment. According to the model, as capacity decreases, the relative influence of category errors on perception increases (as illustrated in Figure 2).

Model extensions and additional analyses

The model we have described so far assumes that the perceptual cost function is the sum of two terms, a metric and a categorical cost. Are both of these components necessary to explain performance in the information-theoretic model?

To answer this question, we fit two additional information-theoretic models: one that uses only a metric cost, and one that uses only a category cost. Relative performance of the three models was assessed using the Bayesian Information Criterion (BIC score). The results indicate that the model combining both categorical and metric costs is strongly supported over the two alternatives (BIC scores are reported in Table 1a; $\Delta\text{BIC} > 10$ is typically interpreted as strong support). This result is particularly important, as nearly all existing models of visual working memory ignore or overlook the influence of categories on perception (for further discussion, see Allred & Flombaum, 2014; Bae et al., 2015).

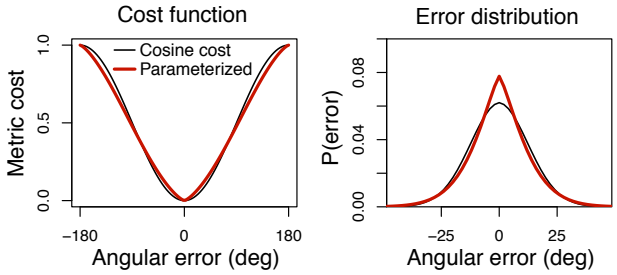

Figure 6: Left: Comparison of two different metric cost functions: a cosine cost function and a parameterized function fit to the data from the delayed estimation experiment. Right: Predicted error distributions for each cost function. In both cases a channel capacity of 2.75 bits is assumed.

One additional assumption of the model that we examine is the form of the metric cost function. The model described so far has assumed a cosine cost function (Equation 2). We developed an alternate model that includes category error costs, as well as a parameterized metric cost function of the form

$$f(y-x) = \frac{\delta z^{\gamma}}{\delta z^{\gamma} + (1-z)^{\gamma}}, \quad z = \frac{|\, y-x \,|}{\pi}, \tag{8}$$

where $|\, y-x \,|$ represents the absolute angular difference. This cost function introduces two additional parameters into the model, δ and γ, controlling the shape of the function. The category cost was left unchanged from the original model. Relative BIC scores comparing the cosine and parameterized metric cost functions are reported in Table 1b. The results indicate that the parameterized cost function offers a superior account for both the delayed and undelayed experiments.

The differing form and predictions of the two cost functions are illustrated in Figure 6. Although the two cost functions appear similar, the parameterized function predicts an error distribution with a sharper peak and heavier tails than a cosine cost function. This result is also in line with previous findings in the visual working memory literature (Sims, 2015; Van den Berg, Shin, Chou, George, & Ma, 2012).

Finally, we note that in the current paper model parameters were fit to the aggregated data from all participants. An important direction for future research is to examine individual differences in the influence of color categories as they relate to differences in memory capacity.

Conclusions

This paper argues that a useful approach for understanding perception is to understand the cost function that it seeks to minimize. We proposed a specific cost function for the perception and memory of color, based on a combination of a metric cost, and a cost associated with memory errors that cross category boundaries. This hypothesis was explored using a branch of information theory known as rate–distortion theory, which concerns optimal communication or information transmisison subject to strong limits on channel capacity.

We demonstrated that the application of this framework was able to provide a close quantitative fit to experimental data previously collected (Bae et al., 2015).

There is substantial evidence to suggest that categories can influence perception (Goldstone & Hendrickson, 2010). For example, Huttenlocher et al. (2000) found that observers' memory for the size of simple shapes is influenced by learned categories. Feldman, Griffiths, & Morgan (2009) conducted a rational analysis of perception of speech sounds in noise, and demonstrated that an optimal solution results in the reduced discriminability near prototypical vowel sounds. At the same time, there is conflicting evidence about the role of categories in color perception specifically. Witzel and Gegenfurtner (2013) argue that category effects are not inherent to perception, rather that these effects can be explained by attention to categorical distinctions which stem from linguistic category boundaries. We believe the model developed in the current paper may offer a productive theoretical tool for elucidating the influence of categories on perception.

The current model was partly inspired by an existing model of how categories influence color working memory (CATMET; Bae et al., 2015). Whereas CATMET assumes that color matching results from the integration of evidence from two independent information sources (category and metric representations), the current model assumes a single perceptual channel that optimizes a cost function with two terms. At present there is no consensus regarding the nature of high-level color representation in the brain to inform the selection between these two approaches (for related work see Stoughton & Conway, 2008; Wade, Augath, Logothetis, & Wandell, 2008; Bird, Berens, Horner, & Franklin, 2014). In the absence of clear physiological evidence for a dual channel representation, a one channel model is more parsimonious. A more meaningful distinction between the CATMET model and the current approach, however, lies in their intended roles as process-level and computational explanations for behavior, respectively (Marr, 1982). Unlike CATMET, the current approach explains biases in perception from the rational perspective of minimizing the costs of error.

A closely related approach is a Bayesian model of color perception developed by Persaud & Hemmer (2014). Their model assumes that color perception is the result of Bayesian inference, combining noisy sensory evidence with prior knowledge of color categories. Both Bayesian inference and rate–distortion theory offer normative theoretical frameworks. Bayesian models of perception typically make assumptions regarding the nature of noise that limits the fidelity of perceptual processing. Information theory, by contrast, assumes a limit on channel capacity; the nature of noise in the channel is optimized with respect to a given cost function. Hence, rate–distortion theory offers a more direct approach to studying perception as an adaptive system. We believe that the computational approach we are developing represents a promising framework for understanding color perception as a boundedly optimal system.

Acknowledgments

This work was supported by NSF BCS 0954749 (SA).

References

Allred, S. R., & Flombaum, J. I. (2014). Relating color working memory and color perception. *Trends in cognitive sciences*, *18*(11), 562–565.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall.

Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, *111*(12), 4590–4595.

Blahut, R. E. (1972). Computation of channel capacity and rate–distortion functions. *IEEE Transactions on Information Theory*, *IT-18*(4), 460–473.

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.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78.

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

Marr, D. (1982). *Vision*. New York, NY, USA.

Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1162–1167).

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of vision*, *15*(3), 1–27.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Stoughton, C. M., & Conway, B. R. (2008). Neural basis for unique hues. *Current Biology*, *18*(16), R698–R699.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.

Wade, A., Augath, M., Logothetis, N., & Wandell, B. (2008). fMRI measurements of color in macaque and human. *Journal of Vision*, *8*(10), 6.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of vision*, *13*(7), 1.