Predicting focal colors with a rational model of representativeness

Joshua T. Abbott (joshua.abbott@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Best examples of categories lie at the heart of two major debates in cognitive science, one concerning universal focal colors across languages, and the other concerning the role of representativeness in inference. Here we link these two debates. We show that best examples of named color categories across 110 languages are well-predicted by a rational model of representativeness, and that this model outperforms several natural competitors. We conclude that categorization in the contested semantic domain of color may be governed by general principles that apply more broadly in cognition, and that these principles clarify the interplay of universal and language-specific forces in color naming.

Keywords: Language and perception; semantic universals; color naming; representativeness; Bayesian inference.

Introduction

Do the world's languages reflect a universal repertoire of cognitive and perceptual categories? Or do different languages partition the experienced world in fundamentally different ways? These questions have been pursued in depth in the domain of color naming and cognition (e.g. Berlin & Kay, 1969; Kay & McDaniel, 1978; Lindsey & Brown, 2006; Roberson, Davidoff, Davies, & Shapiro, 2005; Roberson, Davies, & Davidoff, 2000), and current findings suggest an interestingly mixed picture. There are clear universal tendencies of color naming across languages, but there is also substantial cross-language variation (e.g. Regier, Kay, & Khetarpal, 2007), more than is suggested by traditional universalist accounts. At the center of this debate is the disputed role of *focal colors*, or the best examples of named color categories.

It has long been claimed that color naming across languages is constrained by six universal privileged points, or foci, in color space, corresponding to the best examples of what would be described in English as *white*, *black*, *red*, *yellow*, *green*, and *blue*. This view has received empirical support: the best examples of color terms across languages tend to cluster near these six points (Berlin & Kay, 1969; Regier, Kay, & Cook, 2005), and these colors have also been found to be cognitively privileged (Heider, 1972; but see Roberson et al., 2000). A natural and influential proposal (Kay & McDaniel, 1978) is that these privileged colors constitute a universal foundation for color naming, such that languages differ in their color naming systems primarily by grouping these universal foci together into categories in different ways.

Roberson et al. (2000) advanced a diametrically opposed view of color naming, and of the role of best examples in it. They argued that color categories are not defined around universal foci, but are instead defined at their boundaries by local linguistic convention, which varies across languages. They proposed: "Once a category has been delineated at the boundaries, exposure to exemplars may lead to the abstraction of a central tendency so that observers behave as if their categories have prototypes" (p. 395). On this view, best examples do not reflect a universal cognitive or perceptual substrate, but are merely an after-effect of category construction by language: best examples are derived from language-specific boundaries, rather than boundaries from universal best examples.

A proposal by Jameson and D'Andrade (1997) has the potential to reconcile these two opposed stances. This proposal holds that there are genuine universals of color naming, but they do not stem from a small set of focal colors. Instead, universals of color naming may stem from irregularities in the overall shape of perceptual color space, which is partitioned into categories by language in a near-optimally informative way. This proposal has been shown to explain universal tendencies in the *boundaries* of color categories (Regier et al., 2007). However it has not yet provided an account of *best examples* of these categories, which lie at the heart of the debate.

Here, we address this open issue, completing the reconciliation of the two standardly opposed views. We suggest that best examples are largely universal (in line with the universal-foci view), but nonetheless derived from category boundaries (in line with the relativist view). Specifically, given the independent explanation of category boundaries in terms of the shape of color space, we propose that universal tendencies of best examples are derived from those of boundaries, rather than the other way around as has been traditionally assumed. Moreover, we propose that best examples are derived from category boundaries in an optimal manner, echoing the optimal or near-optimal partition of color space into categories. To pursue this idea, we draw on previous work on a rational model of representativeness, and ask whether the best examples of color categories can be well-predicted as those colors that are most representative of a given category.

The remainder of the paper proceeds as follows. In the next section, we discuss the previous work on representativeness on which we draw, and contrast it with other approaches to that problem. We then describe the color naming data we consider, and a set of competing models that predict the foci

of color categories from the extensions of those categories. We first test these models broadly against data from 110 languages, and then test them in a targeted fashion against the data of a language with an unusual color naming system. In both cases, we find that the rational model of representativeness provides a good fit to the empirical data, and outperforms competing models. We close by discussing the implications of our findings.

Representativeness

Why do people believe that the sequence of coin flips HHTHT (where H=heads, T=tails) is more likely than the sequence HHHHH to be produced by a fair coin? Using simple probability theory, it is easy to show that the two sequences are in fact equally likely. Cognitive psychologists have proposed that people use a heuristic of "representativeness" instead of performing probabilistic computations in such scenarios (Kahneman & Tversky, 1972). We might then explain why people believe HHTHT is more likely than HHHHH to be produced by a fair coin by arguing that the former is more representative of the output produced by a fair coin than the latter. If this heuristic is a correct account of such inferences, how do we define it? Numerous proposals have been made, connecting representativeness to existing quantities such as similarity (Kahneman & Tversky, 1972), and likelihood (Gigerenzer, 1996). Tenenbaum and Griffiths (2001) took a different approach to this question, providing a *rational analysis* (Anderson, 1990) of representativeness by trying to identify the problem that such a quantity solves. They noted that one sense of representativeness is being a good example of a concept, and showed how this could be quantified in the context of Bayesian inference.

Formally, given some observed data d and a set of of hypothetical sources, $\mathcal{H}$, we assume that a learner uses Bayesian inference to infer which $h \in \mathcal{H}$ generated d. Tenenbaum and Griffiths (2001) defined the representativeness of d for hypothesis h to be the evidence that d provides in favor of a specific h relative to its alternatives:

$$R(d,h) = \log \frac{p(d|h)}{\sum_{h' \neq h} p(d|h')p(h')} \quad (1)$$

where $p(h')$ in the denominator is the prior distribution on hypotheses, re-normalized over $h' \neq h$. This measure was shown to outperform similarity and likelihood in predicting human representativeness judgments for a number of simple stimuli. We propose this measure can also be used to determine focal colors from the set of colors named with a particular color term - that is, the extension of that named color category.

Representativeness and color foci

Evaluating formal models of representativeness as an account of color foci requires a good source of color naming data. The data we considered were those of the World Color Survey (WCS), which collected color naming data from native speakers of 110 unwritten languages worldwide (Cook, Kay, & Regier, 2005). Participants in the WCS were shown each of the 330 color chips from the stimulus array in Figure 1, and were asked to name each chip with a color term from their native language; we refer to the resulting data as "naming data". Afterwards, participants were asked to pick out those cells in the stimulus array that were the best examples (foci) of each color term they used; we refer to these as "focus data". The WCS dataset is available at http://www.icsi.berkeley.edu/wcs/data.html.

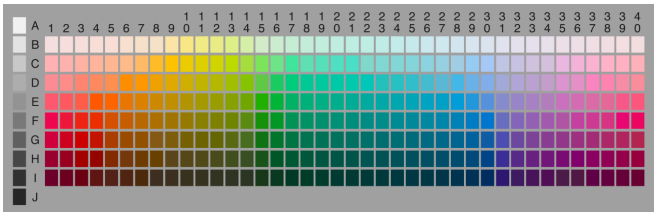

Figure 1: The WCS stimulus array. The rows correspond to 10 levels of Munsell value (lightness), and the columns correspond to 40 equally spaced Munsell hues. The color in each cell corresponds approximately to the maximum available Munsell chroma for that hue-value combination.

We applied Tenenbaum and Griffiths' (2001) representativeness model, and a set of natural competitor models, to the problem of predicting best examples of color categories from the extension of those categories. Thus, the models we consider are different formalizations of our central proposal that best examples may be derived from category boundaries. Following Kay and Regier (2003), we represented each color in the stimulus array as a point in 3-dimensional CIELAB color space. For short distances at least, Euclidean distance between two colors in CIELAB is roughly proportional to the perceptual dissimilarity of those colors. For each named color category used by each speaker in each language of the WCS, we modeled that category as a 3-dimensional Gaussian distribution in CIELAB space, and estimated the parameters of that distribution using a normal-inverse-Wishart prior, a standard estimation method for multivariate Gaussian distributions of unknown mean and unknown variance (Gelman, Carlin, Stern, & Rubin, 2004). Specifically, given a set of M chips $\mathbf{x}_i$ in color category t we obtain the estimates:

$$\mu_t = \frac{1}{M}\sum_i^M \mathbf{x}_i$$

$$\Sigma_t = \frac{SS_t + \lambda_0}{n_t + \nu_0}$$

where SS_t is the sum of squares for category t: $\sum_i^M (\mathbf{x}_i - \mu_t)(\mathbf{x}_i - \mu_t)^\intercal$, n_t is the number of chips in category t for the current speaker, and λ_0 and ν_0 are the parameters of the prior. λ_0 was set by taking an empirical estimate of the variance in CIELAB coordinates over all chips in the stimulus array, and ν_0 was set to 1. We chose this Bayesian formulation of parameter estimation over standard Maximum Likelihood Es-

timation (MLE) since MLE will result in singular covariance matrices for color categories containing few color chips.

With an estimate of the distribution characterizing the category named by color term t, we can now adopt the representativeness measure given in Equation 1 to determine how good an example each color chip x is of a color term t. Substituting x in for the observed data d and t for hypothesis h we obtain the expression:

$$\mathrm{R}(x,t) = \frac{p(x|t)}{\sum_{t' \neq t} p(x|t')p(t')} \tag{2}$$

where $p(x|t)$ is computed from the density function of the estimated Gaussian described above and the priors $p(t')$ are proportional to $n_{t'}$, the number of chips in named color category t'. We test this Bayesian measure against the alternative proposals of representativeness mentioned above (Gigerenzer, 1996; Kahneman & Tversky, 1972): a likelihood model and two similarity models (a prototype model and an exemplar model). In addition, we explore a model that selects as the focus for category t that chip in the extension of t that has the highest chroma. Chroma, or saturation, corresponds loosely to how colorful or "un-gray" a given color is, and in exploring this model we follow the suggestion (Jameson & D'Andrade, 1997; Regier et al., 2007) that focal colors tend to be those with high chroma. We note that each of these models captures some variant of the category *central tendency* idea promoted by Roberson et al. (2000), as described above. We present the details of the competing models below. As with the representativeness model, for a given color x and color term t, each model assigns a score indicating how good x is as an example of t.

Likelihood model. In this model, the goodness score of color x as an example of color category t is given by the density function of the Gaussian distribution that was fit to the naming data for t. Thus,

$$\mathrm{L}(x,t) = p(x|t) \tag{3}$$

Note that this model is similar to the representativeness model, but without the denominator which captures competition among categories in that model.

Prototype model. In this model we define the focus, or prototype, of color category t to be the mean μ_t of the distribution characterizing t. The score for this measure then becomes the similarity of x to that prototype:

$$\mathrm{P}(x,t) = \exp\{-\mathrm{dist}(x,\mu_t)\} \tag{4}$$

where $\mathrm{dist}(\cdot,\cdot)$ is the Euclidean distance between two colors in CIELAB color space.

Exemplar model. We define the exemplar model using a scoring metric similar to that in the prototype model, except rather than computing the similarity of color x to a single prototype, we compute its similarity to each color chip that falls in the extension of category t, and sum the results. This similarity measure is thus computed as

$$\mathrm{E}(x,t) = \sum_{x_j \in \mathbb{X}_t} \exp\{-\lambda\, \mathrm{dist}(x,x_j)\} \tag{5}$$

where $\mathbb{X}_t$ is the set of color chips that fall in the extension of category t, $\mathrm{dist}(\cdot,\cdot)$ is is the Euclidean distance between two colors in CIELAB space, and λ is a free parameter. For the results presented below, λ was set to the value that yielded the best performance overall, which was 0.25.

Chroma model. The score for this model is computed similarly to that for the prototype model, but rather than computing the similarity of color x to the mean of a distribution characterizing category t, we compute its similarity to that color chip c_t which has the highest chroma (saturation) value within the extension of category t. The chroma values for each chip in the stimulus array are provided with the WCS data. Thus we compute

$$\mathrm{C}(x,t) = \exp\{-\mathrm{dist}(x,c_t)\} \tag{6}$$

where $\mathrm{dist}(\cdot,\cdot)$ is the Euclidean distance between two colors in CIELAB space, and c_t is the chip within the extension of t that has the highest chroma value. In the case of ties for c_t - that is, several chips with the same maximum value for chroma - we randomly select a chip from the set of ties.

Predicting foci from category extensions

We assessed these models as follows. For each speaker of each language in the WCS, we first considered that speaker's naming data, and modeled the categories in those data as a set of Gaussians in the manner described above. Then, for each such category, we determined how representative of that category each of the 330 chips in the stimulus array is, according to each model. This yielded, for each model, a ranking of chips in the array by predicted representativeness, and we then compared this model prediction with empirical focus data from the WCS. In the following sections we present both qualitative and quantitative evaluations of the models.

Distribution of foci

A simple means of assessing the models is to generate predicted focal choices from each model's ranking of chips, and to then compare those predicted focal choices with the actual focus data of the WCS. Some speakers in the WCS provided more than one focus (best example) for some categories; if a speaker provided n foci for a given category, we selected the n top-ranked chips as a given model's predicted focal choices for that category and speaker. In this manner we obtained, for each model, one predicted focal choice for each empirical focal choice in the data. We then counted the number of times each of the 330 color chips in the stimulus array was selected as a focal choice, yielding a distribution of focal choices over the stimulus array. We then compared the

empirical distribution of foci across the array with the distribution predicted by each of the models. Following Regier et al.'s (2005) empirical analysis of WCS focus data, we plotted these distributions over the chromatic portion of the array, where the 2-dimensional layout makes contours easily interpretable. Accordingly, we did not plot the focal choices for the terms a speaker used to name A0 and J0, corresponding to English focal *white* and *black*. The resulting contour plots, of the empirical WCS focus distribution and the five models' predicted focus distributions, are shown in Figure 2.

The empirical distribution is shown in panel (a), and replicates the findings of Regier et al. (2005). The distribution predicted by the Bayesian representativeness model (panel b) matches this empirical distribution qualitatively fairly well. Moreover, at least on informal inspection, the Bayesian model appears to approximate the empirical distribution more closely than do the competing models. The chroma model (panel f) at first appears to also approximate the empirical distribution fairly well, but closer inspection reveals that several of the peaks of the model distribution do not align correctly with those of the empirical distribution.

This qualitative assessment is reinforced by a quantitative one. The Jensen-Shannon divergence (JSD) is a measure of the dissimilarity between two probability distributions, P and Q, defined as

$$\mathrm{JSD}(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M) \qquad (7)$$

where $M = \frac{1}{2}(P+Q)$, and $KL(\cdot)$ is the more commonly-known Kullback-Leibler divergence. JSD is closely related to Kullback-Leibler divergence, with the important difference that JSD is always a finite value, ranging from a value of 0 when the two distributions are identical, to a value of 1 when they are maximally different.

We computed the JSD between the WCS empirical focus distribution (normalized so that it may be considered a probability distribution, taken to be P in Equation 7), and each of the model distributions (similarly normalized, taken to be Q in Equation 7). The results are shown below in Table 1. The Bayesian model outperforms the other models, diverging less from the empirical distribution than its competitors.

Rank position of foci

Each model produces as output a ranking of the stimulus chips, where rank is assigned in descending order. Thus, another natural way to assess the models is to note the position of the true empirical focal choice in this ranked list. For example, if a model correctly ranked the true focal chip as the single most representative example of a given color category, it would receive a score of 1/330. As noted previously, sometimes a speaker provided multiple foci for a given color term. To accommodate this we averaged the positional ranking of each focus empirically provided and took the resulting quantity as the model performance for a given color term. In turn, we averaged this performance over the number of color terms a speaker used, then averaged over the number of speakers in a language, and finally computed an average overall model performance for all 110 WCS languages. The average rank position of empirical WCS foci for each model is presented in Table 2.

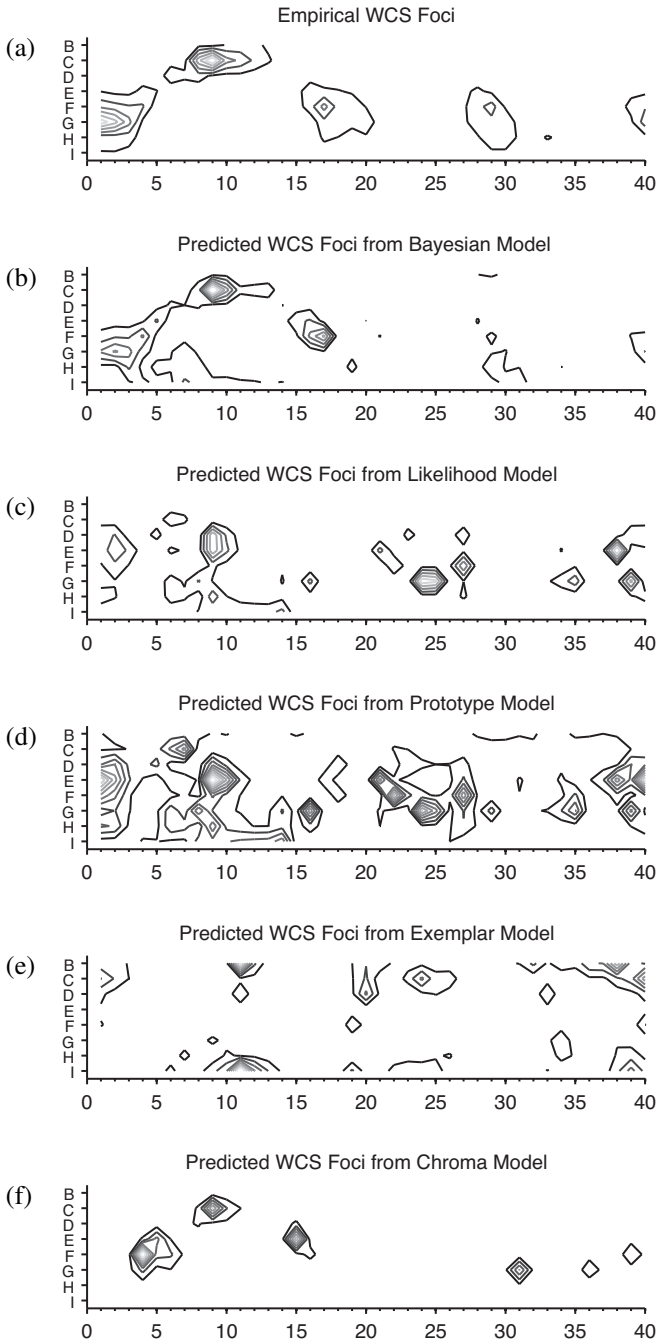

Figure 2: Contour plots of the focus distributions in (a) the WCS, and as predicted by (b) the representativeness model, (c) the likelihood model, (d) the prototype model, (e) the exemplar model, and (f) the chroma model. Each contour line corresponds to 100 focal choices.

As before, we find that the Bayesian measure of representativeness outperforms the other models, ranking the true

Table 1: Divergence between empirical WCS focus distribution and model prediction

Model	Jensen-Shannon Divergence
Bayesian	0.0368
Likelihood	0.1977
Prototype	0.1750
Exemplar	0.1760
Chroma	0.1698

Table 2: Average rank position of empirical WCS foci for each model

Model	Average Rank Position
Bayesian	0.1026
Likelihood	0.1381
Prototype	0.1559
Exemplar	0.1457
Chroma	0.2306

foci within the top 11% of chips on average. In comparison, the likelihood model, which has the second highest average, ranks the true foci in the top 14% of chips on average. It is noteworthy that the chroma model, which captures the natural idea that best examples correspond to chroma maxima, performs most poorly, ranking the true foci only within the top 24% of chips.

A final test: Karajá

So far, we have suggested that color foci may be derived from category boundaries as representative members of a category - and we have shown that this idea accounts well for universal tendencies in focal colors. Thus, foci may inherit their universal tendencies from category boundaries, rather than projecting their universal tendencies to those boundaries. Note, however, that the demonstrations we have seen so far do not discriminate between these two hypotheses. For languages with common color-naming systems, the two hypotheses make the same prediction: foci should tend to fall in the canonical positions shown in Figure 2(a). This is predicted on the traditional universal-foci account, because these are the proposed locations of the universal foci. Roughly the same outcome is predicted by our account, as seen in Figure 2(b).

In a final investigation, then, we attempt to discriminate between these two hypotheses. The hypotheses diverge in their predictions for languages with color categories that have unusual extensions. If foci are a universal groundwork for color naming, then in such unusual cases, foci will fall in the universal (canonical) positions, despite the non-canonicality of the category boundaries. In contrast, our account predicts that in such cases, foci should follow the category boundaries, and fall in non-canonical positions. We test these predictions against a language that is known (Regier, Kay, & Khetarpal, 2009) to have color categories with unusual extensions: Karajá, a language of Brazil.

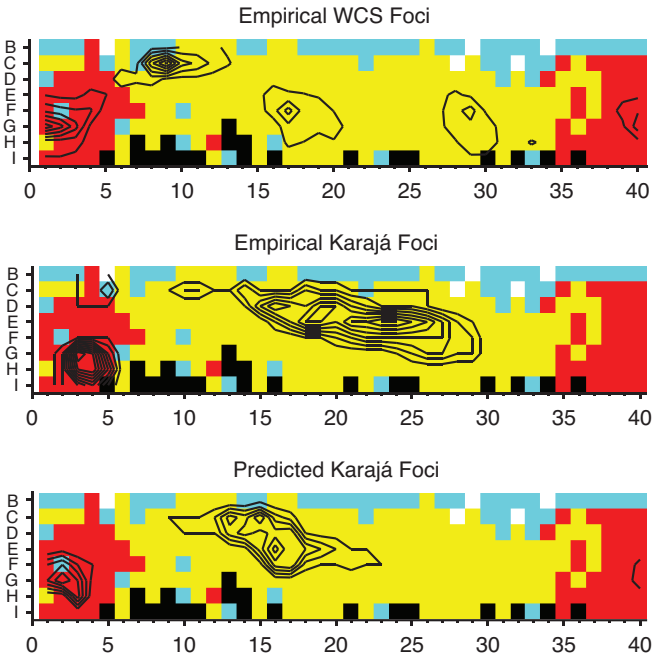

Figure 3: Naming data for the Karajá language, overlaid with contour plots of three different focus distributions: the empirical focus distribution for all languages in the WCS (upper panel), the empirical focus distribution for Karajá itself (middle panel), and the focus distribution predicted by the Bayesian model of representativeness (lower panel).

Figure 3 presents WCS color naming data for Karajá. Here, chips displayed in the same color were named with the same color term by a plurality of participants. These modal naming maps are overlaid with three different focus distributions: the full empirical focus distribution of the WCS (upper panel), the empirical focus distribution from Karajá only (middle panel), and the focus distribution for Karajá predicted by the Bayesian representativeness model (lower panel). The difference between the focus distributions in the top two panels is clearly seen, demonstrating that the foci of Karajá follow the language's color boundaries and are not in line with the universal foci found across the WCS. Additionally, the focus predictions from the Bayesian model of representativeness follow the empirical Karajá focus distribution relatively closely. As before, these qualitative results are confirmed by a quantitative analysis that measures the Jensen-Shannon divergence between the empirical Karajá focus distribution and the distribution predicted by each of the models. As can be seen in Table 3 below, the Bayesian model outperforms the other models on Karajá considered by itself, not just on the entire WCS dataset. We also examined the rank position of the empirical Karajá foci in the ranking produced by each model, and by this measure as well, the Bayesian model fits the data more closely than the competitors, as shown in Table 4 below.

In sum, when boundaries fall in non-canonical positions,

Table 3: Divergence between empirical Karajá focus distribution and model prediction

Model	Jensen-Shannon Divergence
Bayesian	0.3272
Likelihood	0.4430
Prototype	0.5524
Exemplar	0.5137
Chroma	0.5848

Table 4: Average rank position of empirical Karajá foci for each model

Model	Average Rank Position
Bayesian	0.2064
Likelihood	0.2298
Prototype	0.2877
Exemplar	0.3023
Chroma	0.3199

foci do as well - suggesting that foci may in fact be derived from boundaries. This conclusion is reinforced by the observation that the Bayesian representativeness model predicts foci from boundaries fairly well in this non-canonical case, as well as more generally across the WCS.

Conclusion

Focal colors, or best examples of color terms, lie at the center of the debate over color naming. These foci have traditionally been viewed either as the underlying source of color naming universals, or as derived from category boundaries that vary with local linguistic convention. In contrast, we have argued for a novel account of this disputed construct, in which focal colors show strong universal tendencies, but are nonetheless derived from category boundaries, as the most representative members of categories. In support of this proposal, we have shown that an existing Bayesian model of representativeness can predict the distribution of focal colors in the world's languages, from category extensions. This account synthesizes traditionally opposed views of color naming (Kay & McDaniel, 1978; Roberson et al., 2000), and accounts for data that challenge the traditional views.

Our proposal also coheres naturally with a recent theoretical account that explains universal tendencies in color naming in terms of optimally informative partitions of an irregularly shaped perceptual color space (Jameson & D'Andrade, 1997; Regier et al., 2007). Significantly, that view explains universal tendencies in color category boundaries without reference to a small set of focal colors, and it leaves the nature of focal colors unexplained. Our proposal fills that gap. Taken together, the two proposals suggest a single overall account of color naming: foci are optimally representative members of categories that are defined at their boundaries - and the boundaries themselves result from near-optimally informative partitions of color space.

Acknowledgments. We thank Paul Kay for his helpful comments. This work was supported by grants IIS-0845410 and IIS-1018733 from the National Science Foundation.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

Cook, R., Kay, P., & Regier, T. (2005). The world color survey database. *Handbook of categorization in cognitive science*.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. Chapman & Hall/CRC press.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, *103*, 592–596.

Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10–20.

Jameson, K., & D'Andrade, R. (1997). *Color categories in thought and language*. Cambridge University Press, Cambridge, UK.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

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

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*(15), 9085–9089.

Lindsey, D., & Brown, A. (2006). Universality of color names. *Proceedings of the National Academy of Sciences*, *103*(44), 16608–16613.

Regier, T., Kay, P., & Cook, R. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(23), 8386–8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(4), 1436–1441.

Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, *85*(4), 884–892.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*, 378–411.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369–398.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1036–1041).