# Replicating Color Term Universals through Human Iterated Learning

**Jing Xu (jing.xu@berkeley.edu)**
**Thomas L. Griffiths (tom_griffiths@berkeley.edu)**
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

**Mike Dowman (Mike@ImageScope.net)**
ImageScope

## Abstract

In 1969, Berlin and Kay proposed that there exist cross-cultural universals in the form of basic color terms. To test this hypothesis, the World Color Survey (WCS) collected color naming data from 110 non-industrial societies, identifying regularities in the structure of languages with different numbers of terms. This leaves us with the question of where these universals come from. We use a simple model of cultural evolution known as "iterated learning" to explore the hypothesis that universals emerge from human perceptual and learning biases. We conducted an experiment simulating the process of cultural transmission in the laboratory, and compared the results to the systems of color terms that appear in the WCS data. Our results show that cultural evolution results in convergence of systems of color terms towards a form consistent with the WCS, supporting the hypothesis that universals are the result of perceptual and learning biases.

**Keywords:** basic color terms; iterated learning model; color term universals; cultural evolution; Bayesian inference.

## Introduction

Linguistic universals – properties that seem to hold across all human languages – have the potential to provide unique insight into the nature of human cognition. Universals in systems of color terms are among the best documented of these properties. Berlin and Kay (1969) proposed that color naming systems across different cultures are based on one or more of eleven focal colors corresponding to the English color terms *black*, *white*, *red*, *green*, *yellow*, *blue*, *brown*, *purple*, *pink*, *orange*, and *gray*. Kay and McDaniel (1978) and Kay and Maffi (1999) later refined this model to emphasize the six Hering primary colors (*black*, *white*, *red*, *green*, *yellow*, and *blue*) (Hering, 1964), and to characterize the process by which societies might transition from one system of color terms to another as new terms are introduced.

The World Color Survey (WCS) was initiated in the late 1970's to provide a more comprehensive empirical test of the univerality hypothesis (Kay, Berlin, & Merrifield, 1991; Kay, Berlin, Maffi, & Merrifield, 1997). In the WCS, a total of 330 color chips, comprised of 40 equally spaced Munsell hues at 8 levels of lightness and achromatic chips at 10 levels of lightness (see Figure 1), were presented to speakers of 110 different languages in non-industrial societies. Those speakers were asked to name each color chip, and also to point out the most representative chip for each color term. Later analysis of the WCS data showed that the universality hypothesis was by and large confirmed (Kay et al., 1997). Recently, several statistical analyses of the WCS data have also been conducted in an attempt to resolve the debate over the universality of color naming. For example, Kay and Regier (2003; Regier, Kay, & Cook, 2005) showed that the focal colors in the WCS data largely fall in similar regions to those seen in English; in another study they defined a statistical measure of "well-formedness", and used this measure to show that observed systems of color terms correspond to a near-optimal partition of color space (Regier, Kay, & Khetarpal, 2007).

The consistent cross-linguistic structure highlighted by the WCS raises a new question: Where do these universals come from? They may be a result of cultural universals that may arise from the homogeneity of biological traits and evolutionary paths across cultures that constrain people to consider only a limited range of color categories when learning language, thus forcing color term systems to conform to a limited range of universal types (Hawkins, 1988). However, if we view language as a system culturally transmitted from generation to generation, a simpler hypothesis is that these universals may arise directly from biases that cause learners to prefer some color categorizations over others, but that do not place absolute constraints on the types of color categories that are learnable. One way to explore this hypothesis is using the *iterated learning* model, a simple model of cultural transmission in which a sequence of agents each learns from the behavior of the previous agent in the sequence (Kirby, 2001). In an iterated learning model of the transmission of systems of color terms, each agent learns a system of color terms from examples provided by another agent, and then generates examples which are provided to the next agent in the sequence. Mathematical analyses of iterated learning show that as this process continues, the information being transmitted gradually changes to become consistent with the learning biases of the agents involved (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007). If systems of color terms similar to those seen in the WCS emerge from a process of cultural

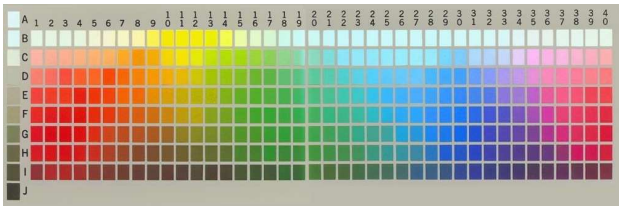

Figure 1: The World Color Survey stimulus array.

transmission by iterated learning, then the biases of individual learners may be sufficient to explain the regularities seen across human societies.

Previous work has used computer simulations to demonstrate that iterated learning with simulated agents can produce systems of color terms similar to those seen in the WCS (Dowman, 2007, 2009). In this paper, we test the hypothesis that color naming universals may be a result of the perceptual and learning biases of human learners by conducting a large-scale laboratory experiment based on iterated learning. In our experiment, human learners acquire and transmit novel systems of color terms, providing a human simulation of the process of cultural evolution. We examine how these systems of color terms change over time, comparing the results to the WCS. We show that, consistent with the hypothesis that perceptual and learning biases are the source of color-naming universals, the systems of color terms generated by our iterated learning chains converged over time to become more consistent with the WCS.

The plan of the paper is as follows. The next section provides further details of the iterated learning model and its predictions about the influence of learning biases on the outcome of cultural transmission. We then present our experiment, which used human learners to simulate the cultural transmission of systems of color terms. Analyzing the results of this experiment raises some technical challenges, which we address by introducing a novel method for quantifying the correspondence between the languages produced by our participants with those in the WCS. We conclude the paper by discussing the implications of our results, and consider some of the potential limitations of our analysis.

## Iterated Learning

Much of human knowledge is not learned from the world directly, but from other people. When we learn languages, we learn them from the utterances of existing speakers, and our utterances inform the next generation of speakers. A simple way to model this process of cultural transmission is in terms of "iterated learning", as illustrated in Figure 2. We imagine a sequence of learners, each of whom observes data, forms a hypothesis about the process that produced those data, and then generates data for the next learner based on that hypothesis.

We can analyze the process of iterated learning by assuming that our learners are rational Bayesian agents. In this framework, learners come up with the *posterior* probability $P(h|d)$ of a hypothesis $h$ given the observed data $d$ by applying Bayes' rule,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'}(d|h')P(h')} \quad (1)$$

where $P(d|h)$ is the *likelihood*, indicating the probability of observing $d$ if $h$ were true, and $P(h)$ is the *prior* probability, indicating the extent to which the learner was willing to accept $h$ prior to observing $d$. The prior encodes the learner's inductive biases, being a factor that combines with the observed data to yield a conclusion.

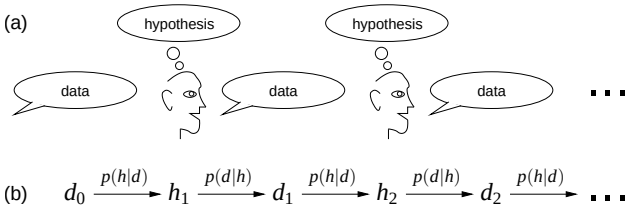


Figure 2: Iterated learning. (a) Each learner sees data produced by the previous generation, forms a hypothesis about the process by which those data were produced, and uses this hypothesis to produce the data that will be supplied to the next generation. (b) In iterated learning with Bayesian agents, each learner sees data $d$, and uses Bayes' rule to compute the posterior probability of each hypothesis $h$, $p(h|d)$. The learner samples a hypothesis from this distribution, and then generates data from the distribution $p(d|h)$.

In the cultural transmission process, data are passed along a chain of learners. Assume that the same Bayesian inference process happens repeatedly at each generation, with each learner sampling a hypothesis from their posterior distribution and generating data by sampling from the likelihood function associated with that hypothesis. This process can be analyzed as a Markov process: The probability each learner selects a particular hypothesis depends only on the data produced by the previous generation. Griffiths and Kalish (2005, 2007) showed that when learners share a common prior distribution, the probability a learner selects a hypothesis converges to the prior probability of that hypothesis as the process of iterated learning continues. Likewise, the probability of generating data $d$ converges to the prior predictive distribution, being the average of the likelihood over the prior, $p(d) = \sum_h p(d|h)p(h)$.

The convergence of iterated learning to the prior potentially provides an explanation of linguistic universals, including universals in color naming. Languages are constantly being passed from speaker to speaker via a process of cultural transmission similar to iterated learning. If this process provides a way for perceptual and learning biases of the kind captured by a prior distribution to have an effect on the structure of languages, we should expect languages to demonstrate properties that are consistent with human biases. If this hypothesis is correct, we should expect to see systems of color terms transmitted via a process of iterated learning to change over time to resemble those that appear in the WCS. To test this idea, we ran a series of iterated-learning chains among an English-speaking population in our laboratory, comparing the systems of color terms produced by those chains to those seen in the WCS.

## Color Term Transmission in the Lab

### Methods

**Participants** Participants were 390 members of the community at the University of California, Berkeley, receiving either course credit or approximately $10/hr for taking part in the experiment.

**Stimuli** Each participant learned a system of color terms by being provided with examples of colors and the terms that were associated with them, and then generalized those terms to new colors. The color stimuli were presented on an Apple iMac computer by a Java program, and the monitor was calibrated using a ColorVision Spyder2 colorimeter/color calibrator on regular basis. A total of 330 colors were used as stimuli, corresponding to the computer screen analogues of the 330 Munsell color chips used in the WCS. Each term was a randomly-allocated pseudo word (from Rastle, Harrington, & Coltheart, 2002), and varied across participants.

**Procedure** We simulated a total of 30 iterated learning chains, each with 13 "generations" of learners. Each chain varied in the number of terms that were allowed in the "language" being transmitted, with two, three, four, five or six terms per language. The first learner in each chain received data generated from one of three types of initial partition of the WCS color space: hue, lightness, and random. The "hue" and "lightness" partitions were approximately equal vertical and horizontal partitions of the color space into the relevant number of categories; the "random" partitions were a truly random partition of the color space, generated uniquely for each chain. These three kinds of initial partitions were used as a means of checking the convergence of iterated learning: by starting the chains with very different systems of color terms, we could easily establish when the influence of the inital partition had disappeared. The following generations of learners all received data generated from the responses of the previous generation, as detailed below. We ran a total of 20 random chains, four for each number of terms, and five hue and five lightness chains, one for each number of terms.

Each participant was trained on the system of color terms by being shown a set of chips together with the corresponding terms. The total number of observed chips was six times the number of terms in the language. These chips were chosen at random from the 330 chips making up the full array, and then provided labels according to either the initial partition (for the first learner) or the responses of the previous learner (for subsequent learners). These training examples remained on the screen while the participant went on to label all 330 color chips from the WCS array. On every trial, they were presented a color chip and asked to select one of the terms to label the color chip. No feedback was given during this phase of the experiment. The responses of each participant thus produced a partition of the set of 330 chips, and this partition was used to generate the labels given to chips for the next learner in the chain.

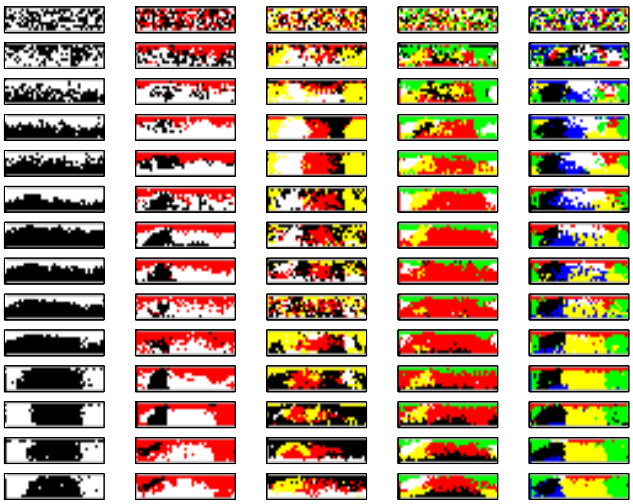

Figure 3: Representative examples of the data produced by simulating iterated learning of systems of color terms in the laboratory. Each panel shows one system of color terms, with arbitrary colors indicating the term assigned to each chip in the World Color Survey array. Each column is one chain with a particular number of terms, each row shows a different generation. The first row shows the random partitions used to initialize each chain.

### Results

Figure 3 shows one set of chains initialized with random partitions, with the number of terms varying from two to six. Through this simple visualization of the data, we can see that each chain started from an unnatural color term system (a random partition), and that transmission along the chains resulted in a very rapid restructuring towards a more regular form. However, it is not clear how well these laboratory-generated data fit the WCS data. In the next section, we use a measure of the difference between each system of color terms and a randomly selected set of responses from the WCS data to test the convergence of the chains and to compare them to the kinds of systems seen across human languages.

## Using Variation of Information to Analyze Color Term Systems

Analyzing the results of our experiment presents a challenge: how can we evaluate whether two systems of color terms are similar? Various methods have been proposed for solving this problem. For example, Kay and Regier (2003) converted the color chips from Munsell space to CIE L*a*b* space so they could compute the centroid for each color term. Centroid distances could then be used to compare clusterings. However, just using centroid measurements may discard important information about the variance of a cluster, and about the locations of boundaries. This method is also dependent on the psychological validity of the CIE L*a*b* representation of colors, which is disputable (Dowman, 2007).

Since our participants' responses consisted of partitions of

the same set of colors as those used in the WCS, we used a technique that compares the Munsell arrays directly, without referring to another color space. This technique uses an information-theoretic measure known as *Variation of Information* (VI) to compare clusterings of a set of items (Meila, 2003). Given two clusterings $C$ and $C'$, the VI is

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \quad (2)$$

where $H(C)$ is the *entropy* of $C$,

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k) \quad (3)$$

where $k$ ranges over the cluster labels and $P(k)$ is the probability of an item being assigned to each cluster, and $I(C,C')$ is the *mutual information* between the two clusterings

$$I(C,C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')} \quad (4)$$

where $P(k,k')$ is the probability an item belongs to cluster $k$ in clustering $C$ and to $k'$ in clustering $C'$.

While VI was originally developed for comparing clusterings, a clustering is simply a partition of a set of items, just as our systems of color terms partition colors according to the terms applied to them. The VI value for two systems of color terms is thus calculated by comparing the distribution of the terms in the two systems, as well as the extent to which they agree with one another. A high VI value reflects a larger difference between two clusterings, whereas a small VI value indicates that the two clusterings are more similar. Our primary analytic tool was comparing the partitions produced by our participants with those observed in the WCS data. We did this by randomly selecting one speaker from each of the 110 languages in the WCS data set, and then calculating the VI of the partition produced by our participants with the 110 partitions from the WCS. Finally, we average across all of the languages from the WCS, to give us a single measure of consistency.[1]

## Testing Convergence

The theoretical analyses of iterated learning outlined above predict that, no matter what language begins a chain, it will eventually converge to a distribution over languages reflecting the prior. We could evaluate this prediction by comparing the chains generated with different initial partitions. A necessary characteristic for convergence is that the VI to the WCS data should not differ between chains, since they should all have reached the same distribution over languages. Figure 4 shows the VI values for the three types of initialization in each language system, showing individual chains with two to six terms for the hue and lightness initialization and the average over all chains for the random initialization.

To test for a difference in VI values across chains, we ran a two-way ANOVA at each generation with initialization and number of terms as the two factors. The main effect of number of terms are significant for all generations ($p < 0.05$); while only the initial systems ($F(2,23) = 196.78, p < 0.0001$) and the first generation ($F(2,23) = 11.19, p < 0.001$) showed a statistically significant effect of initial partition. These results are consistent with a relatively rapid convergence towards a common distribution. Rapid convergence is to be expected in this experiment, since only a very small proportion of the color chips were labeled in each generation, providing a good opportunity for other factors (such as the learning and perceptual biases of the participants) to influence the resulting systems of color terms. This can be seen in Figure 3, where the initial partitions quickly give way to more systematic responses.

## Comparison to the WCS

As described above, to compare our experimental results with the WCS data, VI values were calculated between the responses of each participant and 110 randomly selected WCS systems. Figure 5 shows the VI values for all 20 random chains. A paired t-test on the VI values for the initial and final systems in those random chains showed a statistically significant difference ($t(19) = 11.44, p < 0.0001$), indicating a significant reduction of VI along iterated learning chains, resulting a better fit to the WCS data.

The remaining question is how close our data are to the WCS data: What counts as a low VI score? To address this question, we randomly selected another set of systems from the WCS data, one from each of the 110 languages. Using the same method as used above, we computed the VI between the two sets of WCS data. The average pairwise VI is shown in Figure 5. This average lies close to the mean VI seen in our random chains once they converge. We tested the difference between the VI scores produced by the final participants in each of our random chains and the VI scores for speakers sampled from the WCS using a two-samples t-test. The result was not significant ($t(128) = -0.29$, $p = 0.78$). These results suggest that the systems of color terms generated from our lab are indeed consistent with the data collected from the WCS.

## Rotation Analysis

One potential objection to the conclusion that our chains are moving closer to the WCS could be that the reduction in VI may merely be a result of increasing regularity in the responses. As the systems of color terms in the random chains move towards more regular forms, the VI scores will go down naturally, regardless of whether the actual partition of terms reflects the structure of the WCS or not. To further test the consistency between our experiment results and the WCS

[1] While it would be desirable to also average over speakers, this was too computationally intensive to be practical in our current analyses. We observed little variation in average VI across sampled sets of speakers. We chose to use a single speaker rather than a composite formed by aggregating across speakers within a language (a "mode map") on the grounds that this might not produce a system typical of the language of any individual, especially as different speakers of the same language sometimes use different numbers of color words (Kay & Maffi, 1999).

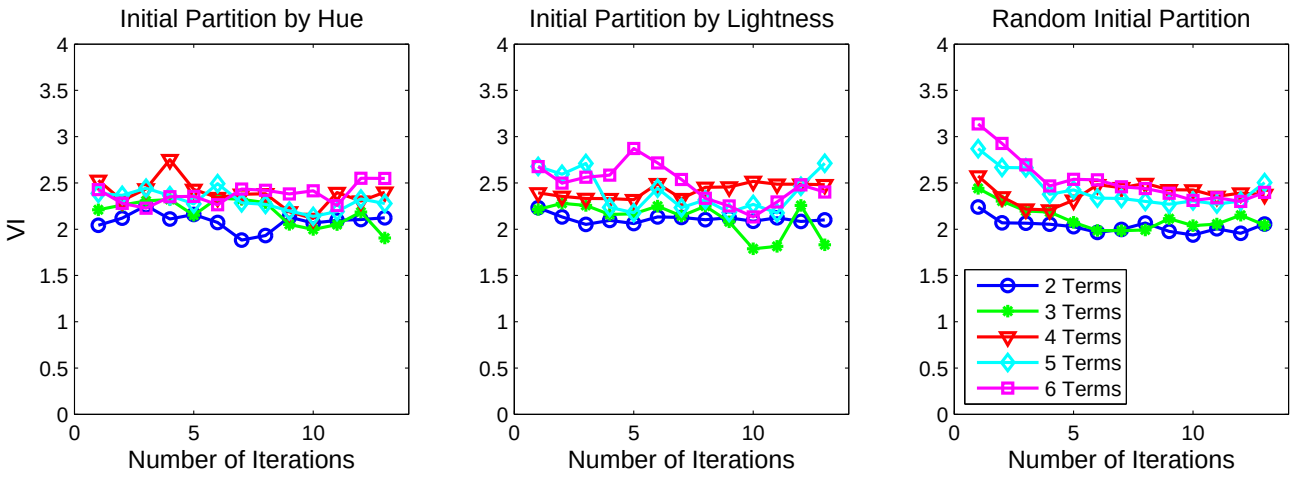


Figure 4: Variation of Information (VI) fit to WCS data for iterated-learning chains with three types of initial partitions and two to six color terms. Results for random initial partitions are averaged over four chains each.

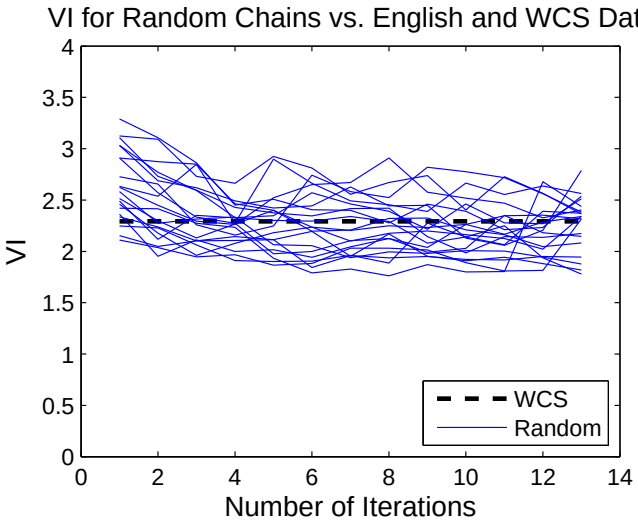


Figure 5: Variation of Information (VI) fit to WCS data for random chains. The dashed line shows the VI for comparing the WCS to itself.

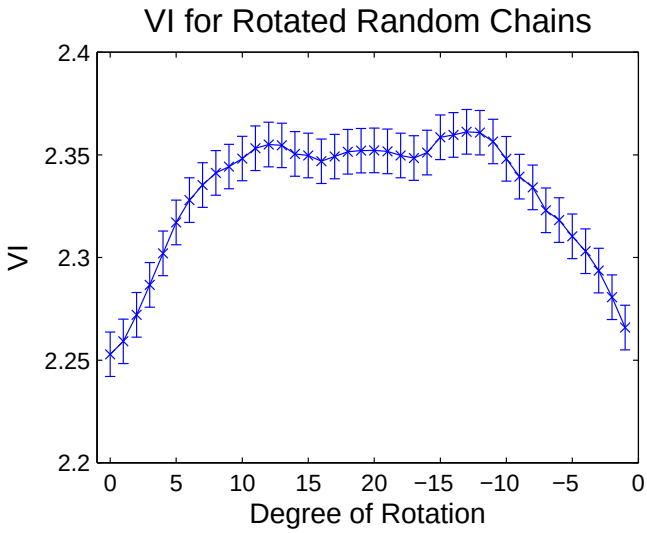


Figure 6: Variation of Information (VI) fit to WCS data for rotations of the final partitions produced by chains initialized with random partitions.

data, we compared the degree of match of each system to the WCS data when it was rotated in the hue dimension by varying amounts. We would expect that the more a partition was rotated out of position, the lower the resulting degree of match would be. This procedure was used by Regier et al. (2007) in connection with their measure of how "optimal" a set of color terms was as a division of the color space into maximally perceptually distinct regions.

Figure 6 shows the mean VI values of the partitions generated by the final participants in our random chains, when rotated from 0 to 20 steps in the hue dimension. Paired t-tests on VI values for no-rotation vs. maximum-rotation ($t(19) = -6.12, p < 0.01$), no-rotation vs. quarter-rotation ($t(19) = -3.66, p < 0.001$), and no-rotation vs. three-quarter-rotation ($t(19) = -4.61, p < 0.001$) all showed statistically significant differences, indicating that the data from the experiment fits the WCS data significantly better than the rotated systems. This analysis thus confirmed that the iterated learning chains did converge to forms closer to the WCS.

## Discussion

We tested the idea that human color-naming universals may be a result of shared learning and perceptual biases, demonstrating that systems of color terms similar to those seen in a variety of non-industrial societies emerge purely as a result of cultural transmission. Using Variation of Information as a measure of the difference between systems of color

terms generated in our experiment and the WCS data, we showed that the VI for systems generated by iterated learning rapidly decreases as the systems moves from unnatural random partitions to more regular forms. Our rotation analysis also showed that this reduction of VI can not be explained as simply a result of the emergence of more regularity, but reflects the adoption of a form consistent with the WCS data.

One objection that could be made with respect to our study is that our English-speaking subjects could have been imposing a system of colour naming reflecting that of English on the languages in our experiments, rather than using pre-linguistic universal biases. As English has 11 basic color terms, many more than the 2 to 6 terms in our experiments, none of the emergent languages could reflect English very closely, which we could expect would minimize the potential for our participants knowledge of language to shape the colour categories formed in the experiment. We take the finding that systems of color terms similar to those seen in the WCS can be produced by cultural transmission by English speakers as supporting our argument that human learning and perceptual biases may be sufficient to explain universals, under the assumption that the English-speaking participants in our experiments share the same learning and perceptual biases as the members of non-industrial societies surveyed by the WCS. This result is less surprising when we take into account previous findings relating the color term categories produced by English speakers with cross-linguistic trends. For example, Boster (1986) found that when English speakers were asked to recursively split a set of color chips into subsets, the partitions they produced corresponded to those seen in other languages with a corresponding number of terms.

Our experiment and subsequent analyses not only demonstrate that iterated learning may provide a valuable experimental method for investigating human inductive biases, but also show that languages formed in the laboratory by English speaking participants seem to converge toward a form consistent with the WCS. These results suggest that the color-naming universals may come from the learning and perceptual biases of human learners, brought out through the process of cultural transmission. In particular, our results supplement previous computational modeling results demonstrating that such properties could be produced by iterated learning with simulated agents. We anticipate that similar pairings of laboratory experiments and computer simulations will be effective in further elucidating how languages and concepts change through cultural transmission.

**Acknowledgments.** This work was supported by grant number BCS-0704034 from the National Science Foundation. We thank Tony Lai, Jason Martin, Linsey Smith, and Joe Vuong for their assistance in collecting and analyzing the data.

## References

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

Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, *25*, 61-74.

Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, *31*(1), 99-132.

Dowman, M. (2009). Evolution of basic color terms. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language evolution and the brain* (p. 109-139). Hong Kong: City University of Hong Kong Press.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 827-832). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441-480.

Hawkins, J. (Ed.). (1988). *Explaining language universals*. Oxford: Blackwell.

Hering, E. (1964). *Outlines of a theory of the light sense.* Cambridge, MA: Harvard University Press.

Kay, P., Berlin, B., Maffi, L., & Merrifield, W. R. (1997). Color naming across languages. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language.* Cambridge, UK: Cambridge University Press.

Kay, P., Berlin, B., & Merrifield, W. R. (1991). Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology*, *1*, 12-25.

Kay, P., & Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicon. *American Anthropologist*, *101*, 743-760.

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

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

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241-5245.

Meila, M. (2003). Comparing clusterings by the variation of information. In *Learning theory and kernel machines* (p. 173-187).

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339-1362.

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

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