

Cultural, Gender, and Individual Differences in Perceptual and Semantic Structures of Basic Colors in Chinese and English*

CARMELLA C. MOORE**, A. KIMBALL ROMNEY*** and
TI-LIEN HSIA***

ABSTRACT

In this paper we examine the judged similarity among the eight basic *focal* colors, and their names, among female and male Chinese (n = 68) and English (n = 52) speaking respondents. The major findings are: (1) all respondents share approximately sixty percent of their knowledge of the judged similarity structures of both semantic and perceptual tasks, (2) there are genuine individual differences among respondents that account for about fourteen percent of their knowledge on average, (3) there are small but statistically significant gender differences that come to about three percent on average, (4) there are small but statistically significant differences between Chinese and English respondents of about one-and-a-half percent, (5) there are differences in the semantic structure of the names of colors as compared to the judgments of the color samples that amounts to about five percent, and (6) there is about a three percent difference in the paired comparison task and the triads task. The results place strong constraints on theories relating to individual differences, linguistic relativity, and the relation of perceptual and semantic structures for colors.

KEYWORDS

Universals, language, cognition.

*The research was supported in part by National Science Foundation Grant SBR-9730831 to C.C.M. and in part by National Science Foundation Grant SBR-9631213 to A.K.R. and W.H. Batchelder. Correspondence concerning this article should be addressed to Carmella C. Moore, Department of Cognitive Sciences, University of California, Irvine, California 92697-5100. E-mail: ccmoores@uci.edu.

**Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100.

***School of Social Sciences, University of California, Irvine, CA 92697-5100.

Introduction

The domain of colors and color terms constitutes a strategic field of investigation because it is central to a number of current theoretical issues across several of the social sciences. Since the landmark study on the evolution of color terminology by Berlin and Kay (1969) anthropologists, psychologists, and linguists have debated the existence of universal basic colors and the extent to which the evolution of color terminologies follows universal regularities in all languages. Their work stimulated an impressive amount of research and has been cited over 700 times. We cannot review that work here although much of it is reviewed in one way or another in the collection edited by Hardin and Maffi (1997), and Maffi (1999) published an updated 116 item bibliography in the reissue of the original Berlin and Kay (1999) monograph.

One central question, as judged by the number of controversial articles, seems to revolve around the so-called Sapir Whorf hypothesis of linguistic relativity versus linguistic universals with respect to color (see, e.g., Kay & Kempton, 1984). Fifty years ago there was little empirical evidence (although see Berlin & Kay, 1999, Appendix II on 19th century history of color work) as to whether different languages classified colors in basically similar ways or, to phrase it in anthropomorphised fashion, each language was free to classify colors in any way it desired. There are ethnographic studies of several individual societies (e.g., Conklin, 1955; Heider, 1972; Heider & Olivier, 1972; Levinson, 2000) that show rather clearly that languages need not have elaborate color terminologies. The critical question is whether, when a language does develop a color lexicon, there is a universal tendency for languages to lexically label similar areas of the color space in similar ways.

The major thrust of the Berlin and Kay study is that all languages tend to acquire terms for various colors in a similar order and that the color terms in different languages refer to the same specific areas of the color space, i.e., that there are universal constraints (or regularities) on color terminologies (see Kay & Maffi, 1999 for a recent summary). In their scheme, for example, in those languages that have the full set of eleven “basic” color terms, the terms would label similar areas of the color space. That is, “red” in one language would correspond to the same area of the color space as would “red” in another language, and so on for all

the “basic” color terms. One might infer from this that judged similarity structures among colors ought to be similar across cultures, an hypothesis we investigate further below. Of course languages may also develop a large repertoire of color terms and modifiers beyond the basic set. The difference between the basic terms and the additional terms may be ambiguous on occasion but the distinction appears to be very useful in practice.

Another topic raised by the Berlin and Kay study is the question of the amount of within culture variability compared to between culture variability. They say, for example, that “controlling for the number of terms, two informants speaking the same language are, on the average, no more similar than two informants speaking different languages” (Berlin & Kay, 1999, p. 12). That is *intracultural* variability is at least as great as *intercultural* variability. This leads naturally to the question of the amount of true individual differences between speakers. In a previous study we found considerable individual differences in the judgments of similarity among color stimuli and color terms (Moore, Romney & Hsia, 2000).

We are also interested in other aspects of color terms and color stimuli beyond those raised by the Berlin and Kay work. A few examples that have influenced the design of this study may be noted. Anthropologists have reported that judged similarity structure among color stimuli are different for males and females (Furbee et al., 1997). In a study of unique hues, “slight discrepancies in the means between females and males and somewhat larger discrepancies in the ranges” was found by Kuehni (2001, p. 26).

Psychologists have found that in many domains, including color, that “judgments of similarity among objects are essentially the same whether the objects are presented or only named” (Shepard, 1975, p. 96; Shepard & Cooper, 1992). Philosophers have long speculated on the question of qualia, for example, whether the green seen by one person is the same as the green seen by another person (Block & Fodor, 1972; Searle, 1992). While our work does not directly answer the qualia question, it does provide an answer to a closely related question, namely, are the relationships among a set of colors the same for one person as for another person.

In this paper we compare judgments of similarity among eight basic chromatic colors (excluding black, white, and gray from the eleven

basic color terms of Berlin and Kay) using names of colors as well as color samples as stimuli. The data were collected from male and female Mandarin Chinese speaking participants in Taiwan, and from male and female monolingual English speaking participants in California. Two measuring tasks were used: a paired comparison rating task of the 28 pairs of eight colors, and a triadic comparison of the 56 triads formed by eight colors. This study is designed to investigate the following questions about the perceptual and semantic structures of color: (1) To what extent do the two cultures with radically different languages share the same structures? (2) To what extent do females and males share the same structures? (3) To what extent do individuals have different structures? (4) How different are the results obtained from paired comparisons from those obtained from the triads test? (5) To what extent are the semantic structures (of names of colors) the same as the perceptual structures (of the color samples)? An important feature of our approach resides in the fact that we quantify the size of effects in answering these questions.

The Participants

We conducted our study of Taiwanese participants in Taiwan using Mandarin speaking Chinese at the National Chengchi University in Taipei. Responses were obtained from 35 females and 33 males, and of the 68 participants 59 were Chengchi University students, six were librarians working for the University and three were alumni. All had normal color vision based on the Ishihara color plates (Ishihara, 1997). All were fluent in Mandarin and most of them also knew Taiwanese, a dialect of Chinese. They had all been exposed to English beginning in junior high school at about 12 to 14 years of age. None learned any English before the age of eight and none would be classified as bilingual, although a couple of dozen say they can speak English. The average age of the participants was 22.5 years old. Participants were paid the equivalent of about ten U.S. dollars for their participation.

The English speaking participants were students at the University of California, Irvine who were monolingual English speakers. Responses were obtained from 25 females and 27 males. All were students at the University of California, Irvine and had normal color vision based on the Ishihara color plates. They all received course credit for their participation.

Materials and Study Design

We use Boynton and Olson's (1987) definitions of eight chromatic *focal* colors (excluding gray, white, and black, and retaining blue, brown, green, orange, pink, purple, red, and yellow) as identified in the Optical Society of America (OSA) system (Nickerson 1981) as color sample stimuli. In the Boynton and Olson study nine subjects named each of the 424 different OSA color chips under carefully controlled conditions. Reaction times for this naming task were recorded. Consensus colors were defined as color chips given the same name by all subjects. The focal colors were defined "as those samples, named with consensus, that exhibit the shortest response times within their categories" (Boynton & Olson, 1987, p. 99).

In this study each participant completed four tasks: (1) a triads task using color terms as stimuli, (2) a paired comparison rating scale task using color terms as stimuli, (3) a triads task using the colors as stimuli, and (4) a paired comparison rating scale task using the colors as stimuli. In the triads task all 56 possible triadic combinations of terms or colors were presented with the instruction to pick the color (term or sample) most different from the other two. In the paired comparison rating scale task, the 28 possible pairs of colors (terms or samples) were presented with the instruction to rate the similarity between them on a rating scale from 1 (most different) to 7 (most similar). In the color sample tasks, printed stimuli were prepared that matched the OSA color chips as nearly as possible by visual inspection iterated over many trials by the researchers and were printed with an inkjet printer. Lighting effects were not controlled. In the tasks using terms, English color words were used in the United States and Chinese characters were used in Taiwan. The order of items in the tasks was individually randomized for each participant. The order of the tasks was also randomized across subjects. The names of the colors, the OSA coordinates of the focal colors, and the Chinese characters used in the study are shown in Table 1. Figure 1 shows examples of the two kinds of stimuli used for the color paired comparison and the triad color tasks. Because of printing limitations the colors used in Figure 1 are only approximate.

The names of the colors in English are the eight basic colors (excluding black, white, and gray) listed by Berlin and Kay (1999, p. 94). Berlin and Kay list only four of these terms (again excluding black and white) for Mandarin (*area* China, North China), namely terms for blue, green,

Table 1

The names of the colors, the OSA coordinates of the focal color samples, and the Chinese characters used in the study.

English word	Focal coordinates			Chinese character
	<i>L</i>	<i>j</i>	<i>g</i>	
blue	-6	-4	2	藍色
brown	-6	2	-2	咖啡色
green	-1	5	5	綠色
orange	0	6	-6	橘色
pink	3	-1	-5	粉紅色
purple	-4	-4	-2	紫色
red	-4	2	-8	紅色
yellow	4	12	0	黃色

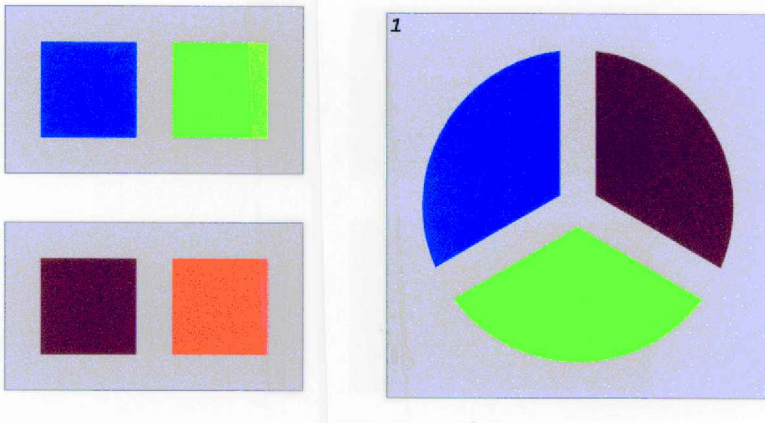


Figure 1. Examples of the stimuli used for respondents to make comparisons among colors with the paired comparison samples on the left and a sample triad on the right. The colors are not exactly as they appear in the experiment due to reproduction imperfections.

red, and yellow (1999, pp. 84-85). We would definitely add the term for purple to the basic list on the basis of extensive interviewing in Taiwan. The term for pink that we used has a prefix that may be glossed as

“powdery” combined with the term for red; the characters used could be loosely translated as “light red.” The term for orange is used very much as in English. Finally the term “coffee” is used in a manner similar to orange except that it refers to brown. We also note that in writing Chinese characters that all terms used as a color referent have a color classifier character added as a kind of suffix denoting color as shown in Table 1. The use of these eight terms facilitates the comparison of terms and colors for both Mandarin and English in the same Euclidean space. We might note that any errors we make in the interpretation of these terms adds to the error variance of our study.

Results

Recent advances in scaling and psychometric measurement methods make possible precise comparisons among semantic structures and perceptual structures of colors. We define a semantic structure as a cognitive representation in which the meaning of the terms, names of colors in this case, relative to each other is represented in Euclidean space. The meaning of an item is defined by its location relative to all the other items. The perceptual structure is defined as a cognitive representation in which the similarity among perceptual objects, color samples in this case, relative to each other is represented in Euclidean space. The perceptual similarities among colors is defined by their location relative to other colors. Our methods facilitate the comparison among semantic and perceptual structures by scaling them into a common Euclidean space. The estimates of the structures are obtained by scaling judged similarity data to obtain a representation in which names of colors and color samples that are judged more similar are closer to each other in the representation than items that are judged less similar. The methods make it possible to scale the semantic structures of any number of individuals, including repeated measures of the same individual, into a common Euclidean spatial representation. Since the methods have been described and illustrated in detail in a series of recent articles (Moore, Romney & Hsia, 2000; Moore et al., 1999; Romney, Moore & Brazill, 1998; Romney, Moore & Rusch, 1997; Romney et al., 1996; Romney et al., 2000) we will not repeat them here.

Figure 2 illustrates the representation that results from scaling all four tasks for each of the 120 participants into a common Euclidean

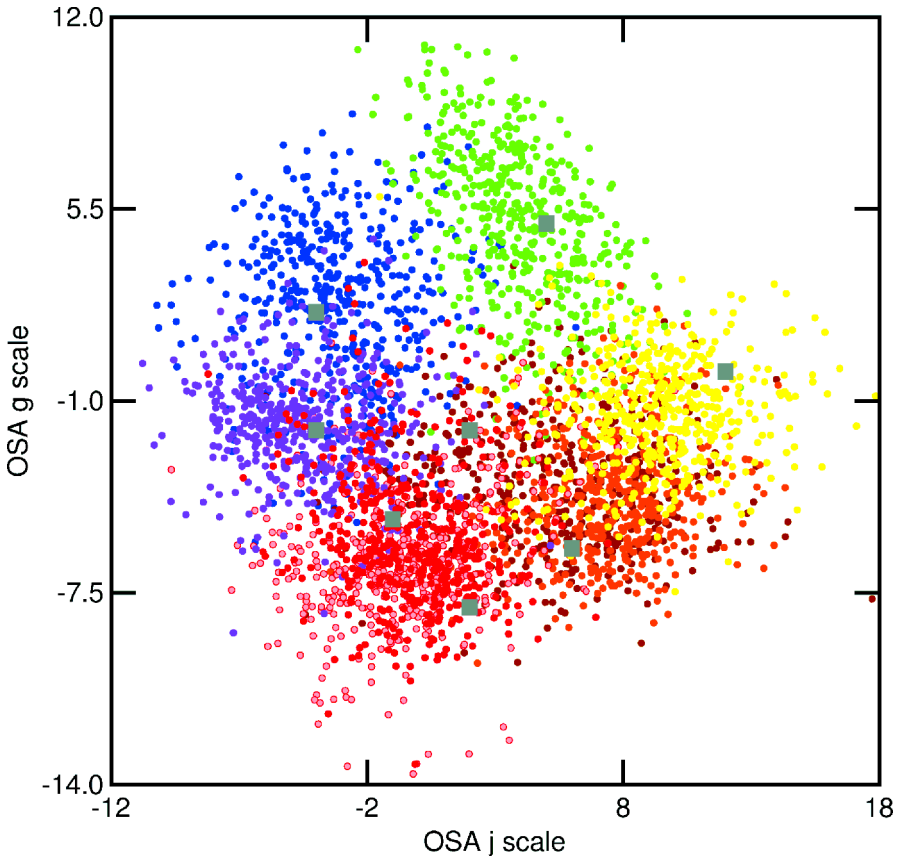


Figure 2. The placements of all participants for all tasks after scaling, rotation, and sizing. For comparison the focal colors (gray filled squares) are shown. The scattering within each color category is due to language variability, task variability, and true individual differences, in addition to the usual sampling variability and error variance.

spatial representation. We obtained the scaling results by stacking the 480 matrices (four tasks for each of the 120 participants), each containing the judged similarity data for a specific task from a specific participant, and applying correspondence analysis (see Romney, Moore & Brazill, 1998, for methodological details). The only novel feature of the representation in Figure 2 is that optimal scores from the correspondence analysis were rotated and rescaled into a best fit to the three OSA dimensions L , j ,

and g of the focal colors for comparison with Boynton and Olson (1987). In Figure 2 the placement of the focal colors is shown as gray squares. The rotation and rescaling were implemented with a Procrustes rotation (Schonemann, 1966) using SYSTAT 8 (Wilkinson, 1999). The L dimension is not discussed in this paper, although the fit is only slightly less than for the two dimensions considered here.

Note that the points representing each color show a clear pattern of placement with considerable variability beyond the main pattern. In the following section we quantify the various sources of the variability of the placement of the points. We calculate the variability associated with the following sources: (1) the central patterning of the points represents a consensus structure shared among the 120 participants and the four tasks completed by each, (2) true individual differences among participants, (3) mode, or the difference in responding to names of colors versus samples of colors, (4) task, or the difference between measures obtained with the triadic task versus the paired comparison task, (5) gender differences in the responses of females and males, (6) language differences between Mandarin and English speakers. In addition to these factors some variability remains unaccounted for and is due to such factors as sampling and error variance.

Figure 3 (A-D) shows visually the effects of mode (A), task (B), gender (C), and language (D) on the mean placement of the terms and samples of colors. The ellipses represent 99% confidence limits around the mean of each distribution on the assumption of bivariate normal distributions. In Figure 3A note that the overall patterns for terms and colors are quite comparable although there are some noticeable differences and regularities. With the exception of yellow most of the positions of the ellipses for names are somewhat on the outside of the picture with respect to those for colors. On average judging similarity of terms of colors was about 5% easier than judging the similarities of color samples. This results in the participants having more agreement on the term tasks than on the color tasks. This phenomena is illustrated in earlier work on kinship structures where we compared the 10 participants with the most agreement with the 10 participants with the least agreement. The positions for those with most agreement were on the outside of the plot relative to those with least agreement (Romney et al., 1996:4702, Fig. 4). As we wrote:

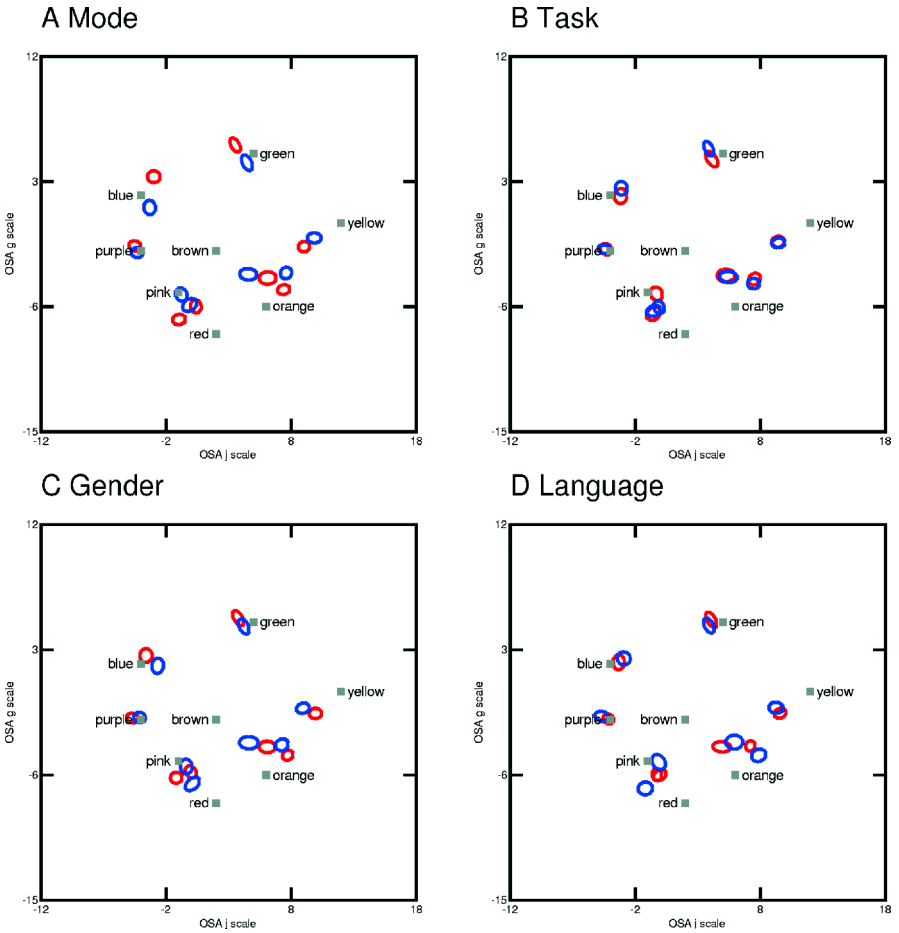


Figure 3. Each panel compares 99% confidence ellipses between two groups: 3A Mode: shows names of colors in red and color samples in blue; 3B Task: shows paired comparisons in blue and triads in red; 3C Gender: shows females in red and males in blue; and 3D Language: shows Chinese respondents in red and English respondents in blue.

An inherent characteristic of the method is that as agreement declines among subjects the ellipses get larger and at the same time drift to the center of the figure. In the case of no agreement, all the ellipses would be large and centered on the midpoint of the diagram. (Romney et al., 1996:4702-4703)

The result shown here is not due to order of presentation as the order of tasks and stimuli were randomized across subjects.

Figure 3B shows the general placement of the ellipses for the triads task and the paired comparison task as very similar since all the pairs of ellipses for the two tasks are overlapping or touching each other. Figure 3C shows a pattern for gender, similar to that seen in Figure 3A, that places the ellipses for females generally further out than for males. The same explanation as given above for the pattern seen in Figure 3A seems appropriate since females have about 4% more agreement than males. In Figure 3D the difference between languages is fairly modest with the Chinese participants making less distinction between pink and red than do the English participants.

Culture Consensus Theory and Cultural Knowledge

Figure 3 presented a visual comparison of the main effects on the semantic and perceptual structure of color. We turn now to an analysis of the similarities and differences among the participants and the four tasks completed by each participant. For example, suppose we ask the question of how similar or different the Chinese participants are compared to the English participants. In order to answer that question we begin with the premise that people in all cultures hold concepts in their minds about a variety of semantic and perceptual domains such as animals, colors, emotions, and kin terms. We further posit that within any given domain, the concepts vary in the extent to which they are similar to each other in meaning. The meaning is defined as the relative location of the concepts in a semantic or perceptual structure (see Romney & Moore, 1998, for theoretical discussion). The structure is empirically defined on the basis of judged similarity on tasks such as triads and paired comparisons. Figure 2 is a representation of 480 semantic and perceptual structures defined in terms of judged similarity data. The aim of this section is to present comparisons among selected subsets of the 480 “pictures” or configurations of the structure of color terms and color samples by language, gender, etc.

In order to compare participants and tasks we need an index or measure of how similar each picture is to each other picture. We use correlation as our basic measure of similarity. The correlation is computed from a vector of 28 interpoint distances obtained from the coordinates

of the three dimensional representations plotted in Figures 2 and 3. This results in a 480 by 480 matrix of correlations among all participants and tasks. The following paragraphs describe how we use this correlation matrix to obtain estimates of the magnitude of the various effects we are interested in, namely, language, gender, mode, task, and individual differences.

We begin with a comparison of the semantic and perceptual structures of color between Chinese and English participants. A recent paper derives methods for such an analysis (Romney, Moore, Batchelder & Hsia 2000). In this section we use mean correlations within and between groups to answer such questions. In this perspective each participant is viewed as having some knowledge (to be estimated from the data) of a cultural pattern, e.g., the meaning of color terms. The pattern of agreement among participants as measured by the mean correlations within a group reflects the degree of knowledge of the group members (for background of culture consensus see Romney, Weller & Batchelder, 1986; Batchelder & Romney, 1988; Romney & Batchelder, 1989; Romney, 1999). We build on assumptions that are widely used in psychometrics (Nunnally, 1994) and that trace back to Spearman's work early in the century (1904). The first assumption is that the magnitude of the mean correlation among a set of participants indicates the extent to which a common shared pattern exists. The second assumption is that the correlation between two subjects, i and j , is the product of the correlation of each subject with the relevant shared cultural pattern, or the "truth": that is, using " t " for the culturally shared pattern, $r_{ij} = r_{it}r_{jt}$. The magnitude of the r_{it} for each subject may be interpreted as measuring his or her cultural knowledge. Our confidence in such an inference increases as the magnitude of the correlations between individuals increases and the variance between individuals decreases.

A consequence of these assumptions is that the square root of the average correlation within any given category of participants (e.g., Chinese or English) is an approximation of the average knowledge of the relevant cultural pattern among participants within the group (Moore et al., 1999; Romney et al., 2000). We can obtain these correlations by characterizing each of the four tasks for each of the 120 participants as a vector of the 28 interpoint Euclidean distances computed from the three dimensions (with the first two being represented in Figures 2 and 3) from the correspondence

analysis. This results in a 480 by 480 correlation matrix among the four tasks for the 120 participants.

The square root of the mean correlation within the English participants is .616 which may be interpreted as the mean agreement or cultural knowledge of the English participants. This means that, on average, each English participant “shares” 61.6% of cultural knowledge with other English speakers. The corresponding figure for the Chinese participants is 60.8%. In order to determine whether the two groups of participants are similar or different, i.e., how much difference does language and culture make in the meaning of colors, we need to look at the mean correlation between participants speaking different languages. If this figure is low then there are large differences between groups whereas the higher it is, the less difference there is between the two groups as a whole. The square root of the mean correlation among participants of different languages is .596, or 59.6%. This is to be compared to the weighted mean of the within language groups which is 61.1%. This may be interpreted as indicating that participants from the different languages share 59.6% of their knowledge while they share 61.1% of their knowledge with participants speaking the same language. The difference of 1.5% is the effect of the language and cultural differences in the way participants judge similarities among terms for color and color samples.

Table 2

Effect size of different sources of cultural knowledge for the average respondent estimated from mean correlations within and between various groups stated as percents.

Source of Knowledge	Chinese	English	Overall
A. Additive factors			
Universally shared			59.6
Language			1.5
Individual differences	13.2	15.7	14.4
Residual error			24.5
Total			100.0
B. Non-additive factors			
Gender	4.0	1.1	3.0
Color vs. words	7.1	1.9	5.2
Pairs vs. triads	2.9	3.7	3.3

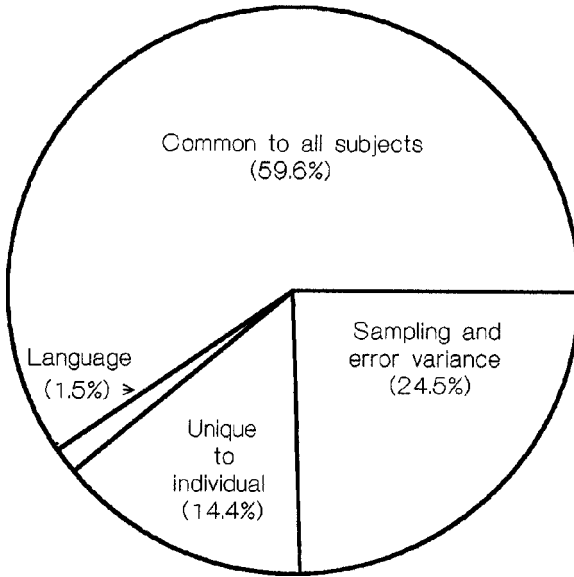


Figure 4. Pie chart showing relative contribution to the knowledge of an average respondent of the following factors: a common universally shared part, that due to a common language, individual differences, and the unexplained sampling and error part.

The square root of the average correlation among the four tasks of all individuals is .755 or 75.5%. The difference between this figure and the 61.1% within language figure given above represents true individual differences, i.e., participants have individual differences in knowledge over and above that shared with other participants. This difference is 14.4% of the total knowledge of each participant, far greater than the effect of language difference alone. We have applied these kinds of calculations to all possible effects in the research design and present the results in Table 2 and the pie chart in Figure 4. The individual differences are larger than all the other difference effects combined.

Analysis of Variance (ANOVA) of Participants

Further information about the statistical significance of the findings may be extracted from the 480 by 480 correlation matrix with an appropriately designed ANOVA. For such an analysis we computed the first three principal components of the correlation matrix among individuals via a

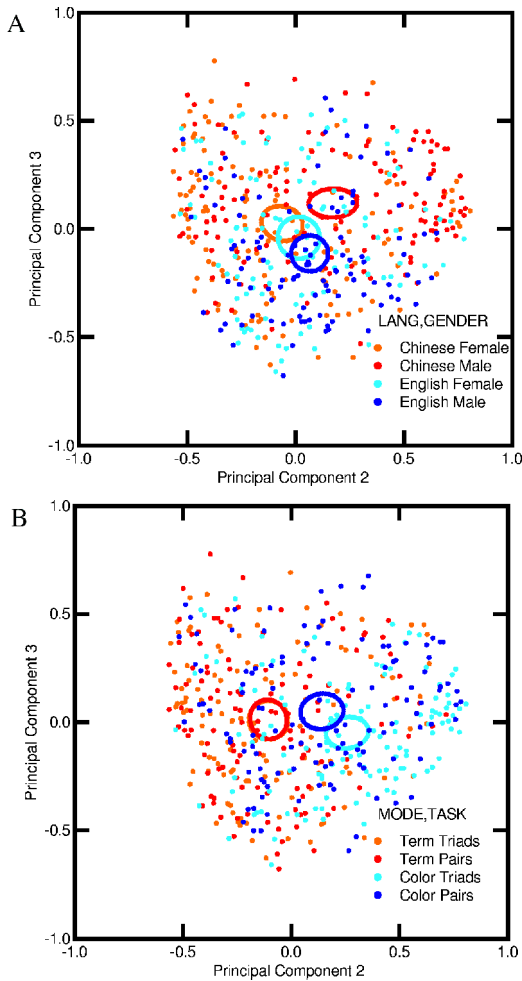


Figure 5. The ellipses are 99% confidence limits of the mean of selected groups. Panel A shows Language and Gender while Panel B shows Mode and Task.

singular value decomposition with components weighted by the square root of the singular values. The singular values of the first three components accounted for 39.2%, 13.4%, and 8.8%, respectively, for a total of 61.5% of the total variance.

Figure 5 (A and B) displays the placement of all tasks and participants on the second and third principal components. Figure 5A gives a visual idea of how much difference there is between participants of different languages

and gender across all tasks. The ellipses indicate the 99% confidence regions around the mean for the specified groups on the assumption of bivariate normality. In Figure 5A there is a noticeable difference between the female and the male Chinese participants on component 2. The difference for the English participants is much smaller although in the same direction. The Chinese participants tend to be above the English participants on component 3. In Figure 5B the color terms task is nearly identical for both methods of comparison and clearly to the left of the color sample comparison on component 2.

The impressions for the visual comparisons just reported were tested for statistical significance with a repeated measures ANOVA (Rosenthal & Rosnow, 1991) with each of the three retained principal components as a dependent variable. The design for the ANOVA treated gender and language as a between subjects source of variance while mode (terms versus sample colors) and task (triads versus paired comparison) were treated as within subjects sources of variance. Table 3 shows the results of the ANOVA. Dimension 1 is basically a knowledge dimension as discussed above. The direction of the differences for gender in Table 3 show that females are significantly higher on cultural knowledge than males; on mode, the judgement of color terms is easier than for color samples; and on task, the paired comparison task is easier than the triad task.

An examination of the results in Table 3 confirm our impressions obtained from Figure 5. Component 2 shows significant effects for all reported variables except language which is significant on component 3. The F value for mode (terms versus colors) is the largest effect found on both component 1 and 2. A careful comparison of the visual plots with the ANOVA is instructive in showing how closely the geometric representation captures the patterns revealed in the ANOVA and visa versa. The magnitude of the effects in Table 2 are estimated with *eta*.¹

As statistically significant as the ANOVA results are, most of the *etas* are rather low. The ANOVA results should be compared with those from

¹There are alternative ways of calculating *eta* and we have used the following formula from Rosenthal and Rosnow (1991, p. 323):

$$eta = \sqrt{\frac{(F \cdot df_{between})}{(F \cdot df_{between}) + df_{within}}}$$

Table 3

Repeated measures analysis of variance results on 120 respondents with gender and language as between subjects and mode and task as within subjects sources of variance. Mode is difference between names of colors and color samples while task is difference between paired comparison ratings and triadic comparisons. The dependent variables are the first three dimensions of a principal components analysis of the similarity among subjects. Non-significant higher order interactions are merged in error term.

Source of Variance	df	MS	F	eta
Between Subjects for Principal Component 1				
Language	1	0.004	0.078	0.026
Gender	1	0.554	10.782**	0.292
Language*Gender	1	0.036	0.708*	0.078
Error	116	0.051		
Within Subjects				
Mode	1	0.263	12.681**	0.309
Error	119	0.021		
Task	1	0.284	11.886**	0.301
Error	119	0.024		
Mode*Task	1	0.089	7.072*	0.237
Error	119	0.013		
Between Subjects for Principal Component 2				
Language	1	0.027	0.107	0.030
Gender	1	2.375	9.519*	0.275
Language*Gender	1	0.996	3.990	0.182
Error	116	0.250		
Within Subjects				
Mode	1	10.234	163.780**	0.762
Mode*Lang.*Gender	1	0.899	14.384**	0.330
Error	118	0.062		
Task	1	0.388	9.863*	0.277
Error	119	0.039		
Mode*Task	1	0.423	9.156*	0.268
Error	119	0.046		
Between Subjects for Principal Component 3				
Language	1	2.596	13.214**	0.320
Gender	1	0.014	0.070	0.025
Language*Gender	1	0.820	4.175	0.186
Error	116	0.196		
Within Subjects				
Mode	1	0.006	0.104	0.030
Error	119	0.054		
Task	1	0.336	8.011*	0.251
Error	119	0.042		
Mode*Task	1	0.278	9.216*	0.268
Error	119	0.030		

Note: ** indicates $P < .001$, * indicates $P < .01$.

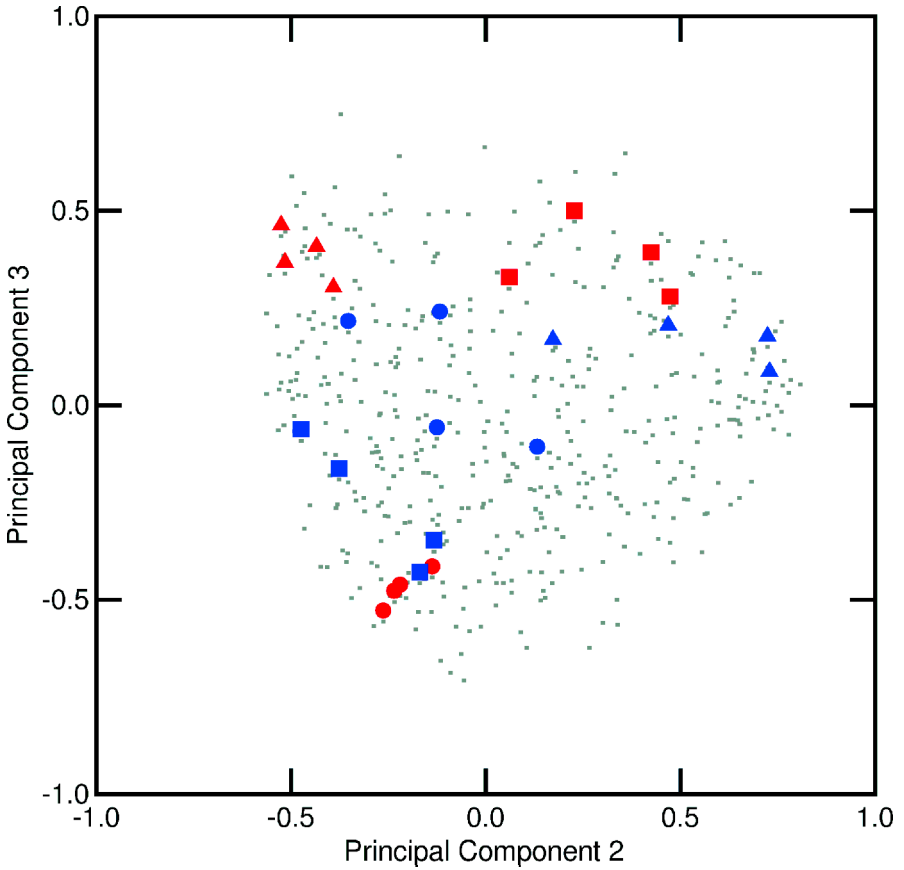


Figure 6. A scatterplot showing a clustering effect among the four tasks for selected Mandarin (red) and English (blue) speaking respondents. Different individuals in each language represented by squares, circle, and triangles.

the mean correlations in the previous section of the paper. The size of the individual differences found there is larger than the combined effects of gender, language, mode, and task. The individual differences, in turn, are small compared to the very considerable magnitude of the universally shared portion.

Figure 6 gives an idea of the individual differences by plotting the four tasks for a small number of selected participants. In Figure 6 all the points are in the same position as in Figure 5A and 5B. Three participants from each language group are highlighted. The aim is to illustrate the fact that

each individual tends to perform in a similar way on all four tasks. Recall that points close together in the figure are more similar in structure than points further from each other. The Chinese participants are shown in red while the English participants are shown in blue. The circles, squares, and triangles represent the three participants from each language.

General Comments about Color, Color Terminology, and Universality

The results discussed above may be summarized very simply as indicating that participants from Taiwan and the United States judge the similarity among colors in very similar ways. It does not make a great deal of difference whether the names of the colors or the color samples are used as stimuli. It does not make a great deal of difference whether one uses triads or paired comparisons to collect the data. It does not make a great deal of difference whether one collects the data from females or males. Finally, there are sizable (bigger than the sum of the task, mode, language, and gender differences put together) differences among individuals. Our previous study on an independent group of Chinese and English females show these same results (Moore, Romney & Hsia, 2000). Our studies have stimulated us to sample the recent literature on linguistic relativity versus universality with respect to color and color terms. The following are our views of the issues and pitfalls in this highly controversial area, with special reference to the Berlin and Kay evolutionary hypothesis and the many research questions it has generated.

We should make clear in advance that in general we believe that Berlin and Kay are fundamentally correct. There may be many errors in the details but the idea that there is an evolutionary sequence of terms, and that, as basic color terms are added to a lexicon, their color foci are in similar parts of the color spectrum in different languages, seems basically correct. The implications of the Berlin and Kay theory are the focus of our discussion. Our views are part of the struggle to eventually reach consensus about these important topics.

First, **languages may or may not have developed an elaborate color lexicon** concerning what we generally mean when we use the concept of color. It is possible that some languages may not have a systematic lexicon of color terms. We are convinced by the several reports

that describe languages that do not elaborate color terminology to a very large extent. This point seems now well established. An example would include Levinson's (2000) recent report on the Yeli. We agree with Levinson's position that some languages may not have an extensive color lexicon. However we believe that the weight of evidence supports certain regularities and constraints on the structure of color lexicons as they develop and evolve.

Second, in our view **not all parts of the color space are equally likely to be named**. One piece of evidence may be found in Figure 7 adapted from MacLaury (1997, p. 202) which shows radically different frequencies of color foci in different Munsell hue areas. The distribution is based on 10,644 color term foci collected from 107 languages. Clearly in the World Color Sample, different parts of the color spectrum are not equally likely to receive names. MacLaury, following Berlin and Kay, used only the "outside" of the color space which included, by and large, the most saturated colors available for each hue and value (brightness) level in the Munsell color space. Had the Berlin and Kay array included interior parts of the color solid the results obtained by MacLaury would be even more striking. One advantage of the Boynton and Olson (1987) naming study that used the OSA system is that both the interior and exterior parts of the color spectrum were sampled with equal coverage. We might note that of the eight focal colors as defined by Boynton and Olson used in this study, seven are on the outside of the color space while the eighth, orange, is only one level in from the outside.

Third, a closely related point to the above is, **not all parts of the color space need to be named at all**, even in languages that have a full complement of color terms. This means that one does not "partition" the color space in a strict mathematical sense of dividing it into mutually exclusive and exhaustive categories. Examples of the tendency to present both theory and data as complete partitions of the Munsell color array that was used in the field trials of the World Color Survey include MacLaury (1997), Hardin (1998), Kay and Maffi (1999), and Kay and McDaniel (1978). For example, Hardin illustrates a schematic representation of the Kay-McDaniel sequence in a figure with eight Munsell color array mappings that are each a complete partition of the sample space (1998:213, Figure 11.2). To explain an evolutionary sequence virtually requires that

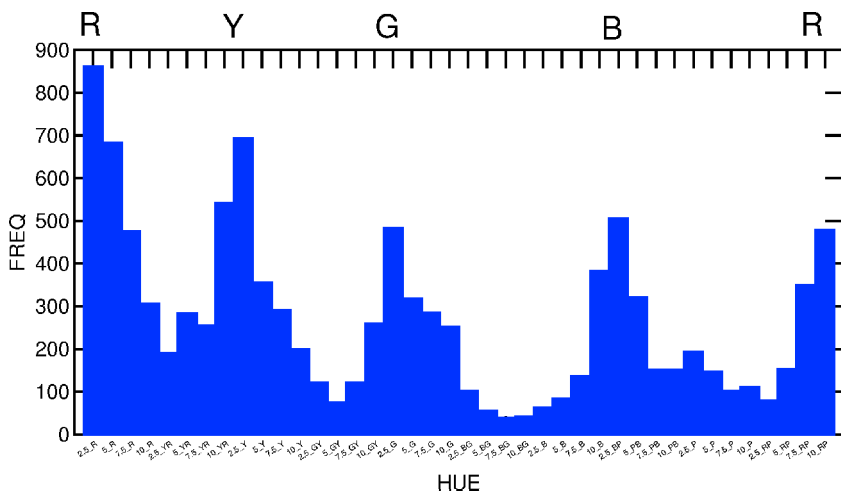


Figure 7. Distribution of 10,644 World Color Sample color terms across Munsell color array adapted from MacLaury 1997, p. 202, Figure 1.

there are some parts of the color solid that somehow stand out more than other parts. For example, saturated colors may be more likely named than unsaturated colors.

These first three observations have been made by many other investigators and are not original with us. Basically they are borrowed from and affirm the work of Kay and Maffi (1999) and Kay (1999). These papers explicitly formulated a hypothesis covering our first two points and urged that it be put to test by objective scientific study. In their “Emergence Hypothesis” Kay and Maffi (1999) clearly elucidate the notions that not all languages have a full lexical coverage of the color space and that some salient areas of the color space are more likely to be named than other less salient areas.

Another unfortunate tendency produced by the notion of partitioning the color space is that it tends to draw attention to the boundaries of the partition rather than to the “prototype” area. This is important enough to deserve emphasis by making an independent point.

Fourth, **it is more productive to focus on color foci than on color boundaries.** Historically the idea of using boundaries goes at least as far back as Ray’s comment: “Comparative work may be facilitated if color units are designated in terms of boundaries rather than midpoints”

(Ray, 1952, p. 255). There are very strong methodological and theoretical reasons that lead us to believe that Ray, and those following his example, may not have followed the most fruitful line of conceptualization. There may be situations in which it is important to focus upon a specific boundary problem, but in general focal points are much easier to quantify than boundaries, especially when dealing with several colors at a time.

From a methodological view it is much easier to measure and quantify a single point in a coordinate system. No one has figured out how to quantify a whole system of boundaries in a two dimensional plane, let alone the whole three dimensional color space where one is dealing with volume rather than areas in a plane. The rather precise measurements of effects in this paper would not be possible using boundaries as the focus. Without quantification we cannot resolve questions involving magnitude of effects, which is the main scientific task. From a theoretical view of prototypes, and other cognitive theories of categories, the focus of attention is on the center rather the boundaries of concepts. Thus if we are to relate anthropological findings of color categories to those in neighboring sciences it will be more likely to be by comparing central locations rather than boundaries.

One simple example of the difference in perspective obtained by looking at the foci of colors or looking at the boundary of colors may be obtained from an examination of the figure in Davidoff, Davies & Robertson (1999, p. 203). Their figure shows the distribution of English and Berinomo color names using the standard Munsell array. They have partitioned the whole set of 160 chips into the eight chromatic color terms for English and into three chromatic color terms for Berinomo (plus words for black and white that extend somewhat beyond the English counterparts, see Roberson, Davies & Davidoff, 2000 for details). The visual comparison of the English and Berinomo figures looks radically different. However, if you look at where the foci (the chip most frequently designated as the best example of the category) of the three chromatic colors are located, they are in the identical position of their English counterparts except for green which is adjacent to its English counterpart. The fact that the three colors come early in the Berlin and Kay scheme and that the foci of all three are so close to English (and many other languages) seem to us to be the salient message of the Berinomo data.

Fifth, **individual differences will become an increasingly important research area.** Individual differences are receiving increasing attention in the vision literature. Recent examples include Cicerone & Nerger, 1989; Jameson, Highnote & Wasserman, 2001; and Kuehni, 2001. In a careful laboratory study with 40 observers Kuehni reports:

It is apparent that when two color normal individuals look at a reflecting sample under identical conditions of viewing, they may not experience the same color. For unique hues this can be shown with minimum ambiguity, because judgments, for example, if a blue contains redness or greenness are relatively easy, especially if a change of hue series of colors is displayed, such as in this experiment. Based on the results, the individual difference can be up to 4 Munsell 40 Hue steps. (Kuehni, 2001, p. 63)

Thus, when Levinson (2000) reports that both intra- and interindividual variation in color naming tasks among Rossel language speakers was very large and seemingly unpatterned, we would argue that all the evidence so far collected (including our own studies of Chinese and English speakers) suggest that this kind of variation is normal and should be expected (e.g., Sturges & Whitfield, 1995). Further, we would argue, unlike Levinson, that this evidence does not, in itself, diminish the general outline of basic color term theory as put forward by Berlin and Kay.

We do not know at this point in time what accounts for the individual differences; they may very well lie at the level of the genetics of the visual system. We note, however, that we don't think, as has been suggested by Furbee et al. (1997) in their study of English speakers, that "eye color," *per se*, is responsible. In both of our recent studies of Chinese and English speakers (the present study and Moore, Romney, and Hsia, 2000), the Chinese, who were all "dark" eyed, show the same order of magnitude of individual differences as do English speakers, who were represented by all colors of eyes, including blue. Our study is directly comparable to Furbee et al.'s in many respects, and it seems unlikely, given our findings, that eye color accounts for the phenomena.

Sixth, **all data should be considered to have a large random error component arising from a variety of random effects.** That means, among other things, that no theory will account for all of the data. To attempt to account for every last case leads one into an endless cycle of accounting for meaningless residuals of random process and pure

error (e.g., a typo by a secretary when entering data). In this study there is a residual of about a quarter of the data that is the result of unknown sampling variability, unknown bias factors, experimenter effects, uncontrolled lighting, unmotivated participants, etc. An examination of our Figure 2 illustrates how much variability exists in empirical data. Although we are using similarity judgments, there is no reason to suppose color naming is any less error free.

Seventh, **focus of study should be on estimating, i.e., quantifying and measuring, the amount or size of the effects rather than taking a polar position or asking if there is an effect.** For example, if one were to simply test whether or not there is a language or gender effect in the present data, they would find that both have significant effects by conventional statistical tests. It would be easy to assume that since language, for example, is highly statistically significant, that linguistic relativity is more strongly supported than linguistic universals. In the present data there is evidence for both a widely shared portion pointing to some universal tendency (about 60%) and a unique portion due to linguistic differences (less than 2%). Instead of arguing an all-or-none polarized position it would be well to measure more and more accurately the size of effect of whatever phenomena we are attempting to account for.

Conclusions

The major findings of this research may be stated rather simply: people speaking different languages classify the eight basic colors and the written counterparts in basically similar fashion although there is a sizable and significant amount of true individual variation within each language. Other statistically significant effects including language, gender, task (triads versus paired comparisons), and mode (color terms versus color samples) all taken together account for a relatively small proportion of the total variability, clearly less than the individual differences.

The original Berlin and Kay (1969) study and the enormous research effort that it has stimulated is clearly one of the most significant scientific contributions of social and cultural anthropology in the last several decades. Though it may need modification in details the major framework they provided has been unusually robust. The number of researchers from a variety of fields (including anthropology, linguistics, psychology, vision,

etc.) working on the problem is impressive. The extent to which these researchers build on the work of each other, bodes well for the future of the field.

REFERENCES

- BATCHELDER, W.H. & ROMNEY, A.K.
 1988 Test theory without an answer key. *Psychometrika* 53:71-92.
 1989 New results in test theory without an answer key. In E.E. Roskam (Ed.), *Advances in Mathematical Psychology*, Vol. 2, pp. 229-248. Heidelberg: Springer-Verlag.
- BERLIN, B. & KAY, P.
 1969[1999] *Basic Color Terms: Their Universality and Evolution*. Palo Alto: CSLI Publications, Stanford University.
- BLOCK, N. & FODOR, J.
 1972 What psychological states are not. *Philosophical Review* 81:159-181.
- BOYNTON, R.M. & OLSON, C.X.
 1987 Locating basic colors in the OSA space. *Color Research and Application* 12:94-105.
- CICERONE, C.M. & NERGER, J.L.
 1989 The relative numbers of long-wavelength-sensitive to middle-wavelength-sensitive cones in the human fovea centralis. *Vision Research* 29:115-128.
- CONKLIN, H.C.
 1955 Hanunoo Color Categories. *Southwestern Journal of Anthropology* 11:339-344.
- DAVIDOFF, J., DAVIES I. & ROBERSON, D.
 1999 Colour categories in a stone-age tribe. *Nature* 398:203-204.
- FURBEE, N.L., MAYNARD, K., SMITH, J.J., BENFER, R.A., QUICK, S. & ROSS, L.
 1997 The emergence of color cognition from color perception. *Journal of Linguistic Anthropology* 6:223-240.
- HARDIN, C.L.
 1998 Basic color terms and basic color categories. In W.G.K. Backhaus, R. Kliegl & J.S. Werner (Eds.), *Color Vision: Perspectives from Different Disciplines*, pp. 207-217. New York: Walter de Gruyter.
- HARDIN, C.L. & MAFFI, L.
 1997 *Color Categories in Thought and Language*. Cambridge: Cambridge University Press.
- HEIDER, E.R.
 1972 Universals in color naming and memory. *Journal of Experimental Psychology* 93(1):10-20.
- HEIDER, E.R. & OLIVIER, D.
 1972 The structure of the color space in naming and memory for two languages. *Cognitive Psychology* 3:337-354.
- ISHIHARA, S.
 1997 *Tests for Colour-Deficiency*. Tokyo: Kanehara.

JAMESON, K.A., HIGHNOTE, S.M. & WASSERMAN, L.M.

2001 Richer color experience in observers with multiple photopigment opsin genes. *Psychonomic Bulletin & Review* 8(2):244-261.

KAY, P.

1999 The emergence of basic color lexicons hypothesis. In A. Borg (Ed.), *The Language of Color in the Mediterranean*, pp. 76-90. Stockholm; Almqvist and Wiksell International.

KAY, P. & BERLIN, B.

1997 Science \neq Imperialism: There are nontrivial constraints on color naming. *Behavioral and Brain Sciences* 20:196-203.

KAY, P. & KEMPTON, W.

1984 What is the Sapir-Whorf Hypothesis? *American Anthropologist* 86:65-79.

KAY, P. & MAFFI, L.

1999 Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist* 101:743-760.

KAY, P. & MCDANIEL, C.K.

1978 The linguistic significance of the meanings of basic color terms. *Language* 54:610-646.

KUEHNI, R.G.

2001 Determination of unique hues using Munsell Color Chips. *Color Research and Application* 26(1):61-66.

LEVINSON, S.C.

2000 Yeli Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology* 10(1):3-35.

MACLAURY, R.E.

1997 Ethnographic evidence of unique hues and elemental colors. *Behavioral and Brain Sciences* 20:202-203.

MAFFI, L.

1999 A bibliography of color categorization research, 1970-1990. In Berlin & Kay (Eds.), *Basic Color Terms: Their Universality and Evolution*, pp. 173-189. Palo Alto: CSLI Publications, Stanford University.

MOORE, C.C., ROMNEY, A.K. & HSIA, T.

2000 Shared cognitive representations of perceptual and semantic structures of basic colors in Chinese and English. *Proceedings of the National Academy of Sciences* 97:5007-5010.

MOORE, C.C., ROMNEY, A.K., HSIA, T. & RUSCH, C.D.

1999 The universality of the semantic structure of emotion terms: Methods for the study of inter- and intra-cultural variability. *American Anthropologist* 101:529-546.

NICKERSON, D.

1981 OSA color scale samples: a unique set. *Color Research and Application* 6:7-33.

NUNNALLY, J.

1994 *Psychometric Theory*. Third edition. New York: McGraw-Hill.

RATLIFF, F.

1976 On the Psychophysiological bases of universal color names. *Proceedings of the American Philosophical Society* 120:311-330.

RAY, V.F.

1952 Techniques and problems in the study of human color perception. *Southwestern Journal of Anthropology* 8(3):251-259.

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* 129:369-398.

ROMNEY, A.K.

1999 Culture consensus as a statistical model. *Current Anthropology* 40:S103-S115.

ROMNEY, A.K., BOYD, J.P., MOORE, C.C., BATCHELDER, W.H. & BRAZILL, T.J.

1996 Culture as shared cognitive representations. *Proceedings of the National Academy of Sciences* 93:4699-4705.

ROMNEY, A.K. & MOORE, C.C.

1998 Toward a theory of culture as shared cognitive structures. *Ethos* 26:314-337.

ROMNEY, A.K., MOORE, C.C., BATCHELDER, W.H. & HSIA, T.

2000 Statistical methods for characterizing similarities and differences between semantic domains. *Proceedings of the National Academy of Sciences* 97:518-523.

ROMNEY, A.K., MOORE, C.C. & BRAZILL, T.J.

1998 Correspondence analysis as a multidimensional scaling technique for non-frequency similarity matrices. In J. Blasius & M. Greenacre (Eds.), *Visualization of Categorical Data*, pp. 329-345. San Diego: Academic Press.

ROMNEY, A.K., MOORE, C.C. & RUSCH, C.D.

1997 Cultural universals: Measuring the semantic structure of emotion terms in English and Japanese. *Proceedings of the National Academy of Sciences* 94:5489-5494.

ROMNEY, A.K., WELLER, S.C. & BATCHELDER, W.H.

1986 Culture as consensus: A Theory of culture and informant accuracy. *American Anthropologist* 99:313-338.

ROSENTHAL, R. & ROSNOW, R.L.

1991 *Essentials of Behavioral Research: Methods and Data Analysis*. Second edition. New York: McGraw-Hill.

SAUNDERS, B.A.C. & VAN BRAKEL, J.

1997 Are there nontrivial constraints on color categories? *Behavioral and Brain Sciences* 20:167-228.

SCHONEMANN, P.H.

1966 The generalized solution of the orthogonal Procrustes problem. *Psychometrika* 31:1-16.

SEARLE, J.R.

1992 *The Rediscovery of the Mind*. Cambridge: MIT Press.

SHEPARD, R.N.

1975 Form, formation, and transformation of internal representations. In R. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium*, pp. 87-122. Hillsdale, NJ: Lawrence Erlbaum.

SHEPARD, R.N. & COOPER, L.A.

1972 Representation of colors in the blind, color-blind, and normally sighted. *Psychological Sciences* 3(2):97-104.

SPEARMAN, C.

1904 "General intelligence," objectively determined and measured. *American Journal of Psychology* 15:201-293.

STURGES, J. & WHITFIELD, T.W.A.

1995 Locating basic colours in the Munsell Space. *Color Research and Application* 20:364-376.

WILKINSON, L.

1999 *SYSTAT 9 Statistics II*. Chicago: SPSS Inc.