Whorf and His Critics: Linguistic and Nonlinguistic Influences on Color Memory

JOHN A. LUCY
RICHARD A. SHWEDER
Committee on Human Development
University of Chicago

The current state of research on language and thought using color as a stimulus is reviewed. Five related experiments that integrate and expand the findings of this previous color research are reported. The relationship between memory and (1) color stimuli discriminability, (2) focality, (3) two-person communication accuracy, (4) group communication accuracy, and (5) referential confusability is assessed. What we discover is that the color array previously used to demonstrate the influence of focality on memory is discriminatively biased in favor of focal chips. Results show that when perceptual discriminability is controlled for, the various linguistic indices are better predictors of memory accuracy in both short-term and long-term recognition memory tasks than is focality. Memory for color stimuli seems to be mediated by basic color descriptions, which may include, but are not restricted to, basic color terms. The discussion takes up the implications of the findings for previous research and for Whorf's hierarchical view of the relationship between language and sublinguistic universals. [Whorfian hypothesis, language and thought, color terminology, memory, culture theory]

JOHN ARTHUR LUCY studied for his B.A. at Pomona College and is presently a doctoral candidate in the Committee on Human Development of the University of Chicago. He has been a lecturer in social sciences in the college of the University of Chicago and has done fieldwork among the Yucatec Maya of Mexico as a Doherty Fellow and as a Social Science Research Council/American Council of Learned Societies International Doctoral Research Fellow. His research focuses on the interrelationship between individual cognitive processing and conventional cultural systems. His current project attempts to bring ethnographic, linguistic, and psychological theories and methodologies together for an integrated understanding of the conceptualization of space and cause among the Maya.

RICHARD A. SHWEDER is Assistant Professor of Human Development, Committee on Human Development, University of Chicago. He received his Ph.D. (1972) from the Department of Social Relations (Anthropology), Harvard University. He has been a research associate of the Department of Anthropology of Utkal University, Bhubaneswar, Orissa, India; a research fellow of the Department of Psychology and Social Relations, Harvard University; a fellow of the Child Development Research Unit, University of Nairobi, Nairobi, Kenya; and a lecturer in the Department of Sociology of the University of Nairobi. His research interests include cross-cultural, developmental, and experimental approaches to modes of thought; cultural schemata and their influence on learning, memory, and judgment; the limits of rational adaptive behavior; the anthropology of thought; and comparative child development.

INTRODUCTION

OVER A PERIOD of 25 years, research using color as a stimulus has produced what is perhaps the most cumulative and systematic corpus of basic data on the language and thought question (see Brown and Lenneberg 1954; Burnham and Clark 1955; Lenneberg and Roberts 1956; Lenneberg 1961; 1971; Lantz and Stefflre 1964; Stefflre, Castillo Vales, and Morley 1966; Heider 1972; also see Brown 1976 for a review).[1] During this period, a dramatic reversal of opinion concerning the relationship between language and thought has occurred. As Brown notes (1976:152), color research "began in a spirit of strong relativism and linguistic determinism [influenced primarily by Benjamin Lee Whorf] and has now come to a position of cultural universalism and linguistic insignificance." This interpretive "about-face" represents to Brown (1976:152) the "fascinating irony" of the color-research tradition; to us it seems especially ironic since the evidence we shall present challenges the empirical basis for this reversal of interpretation.

In the early period, from the early 1950s to the mid-1960s, the guiding assumption of the color-research tradition was that the color spectrum was a continuum, and the basic hypothesis tested was that language shaped thought (viz., recognition memory). Brown himself imagined "a universal law relating referent codability to recognition (and perhaps other aspects of cognition) with each different language having its own codability scores for given referents and the speakers of that language having corresponding recognition skills" (1976:129). Stefflre, Castillo Vales, and Morley (1966:112) argued that their subjects' success in delayed recognition of color stimuli is related to "communication accuracy," that is, the extent to which "the names or descriptions the subjects give for the stimuli allow other native speakers of the language to pick out the particular stimulus encoded." It was during that early period that Brown (1965:334) characterized the process of delayed recognition as follows:

> When the color initially appears you try to give it a distinctive name. . . . When the color is removed the name can be retained, even rehearsed. Somehow, names are responsive to volition in a way that images are not. . . . When the chip is found which best deserves the name, that is recognition.

The latter period in the color-research tradition extends from the late 1960s to the present. As a result of the important work of Berlin and Kay (1969), attention shifted to the psychophysical inequalities of the color spectrum. The hypothesis tested (see Heider 1972) was that certain "focal" areas in the color space are salient to the human perceptual system, form the nucleus of basic color terms, and are inherently more memorable regardless of language and culture. In 1972 Heider concluded that memory for color does not vary with changes in linguistic code, and, on the basis of Heider's evidence, Roger Brown (1976; also see 1977) has recently renounced his earlier linguistic hypotheses. Reflecting on the 25-year history of the color-research tradition, Brown (1976:149) remarks: "From the extreme relativism of Whorf and the anthropologists of his day, we have come to an extreme cultural universality and presumptive nativism."

The goals of the present project are to examine critically the empirical foundation of this reversal in opinion and to introduce new evidence regarding the role of language as a factor in human color memory. We assess the relationship between memory and: (1) color stimuli discriminability; (2) focality; (3) two-person communication accuracy; (4) group communication accuracy; and (5) referential confusability. What we discover is that the color array used by Heider (1972) to establish the translinguistic superior memorability of focal versus nonfocal colors is discriminatively biased in favor of focal colors. Under conditions simulating "perfect memory" (that is, in the presence of visible target probes), focal chips are easier to identify in the array than are nonfocal chips. Other results suggest that when Heider's array is modified to make focal and nonfocal

chips equally discriminable, various linguistic indicators are better predictors of memory accuracy in both short- and long-term memory than is focality. We demonstrate that, for the most part, whatever influence focality has on memory is mediated by language. We introduce the notion of a "basic" *description* for a color (which is not to be confused with a "basic" color *term*). "Basic" descriptions seem to facilitate memory. Finally, in our discussion section, we revive Whorf's (frequently overlooked) hierarchical interpretation of the relationship between sublinguistic universals (whose existence he never denied) and language (whose relationship to sublinguistic universals Whorf viewed as one of "appropriation," not one of "opposition").

THE COLOR-RESEARCH TRADITION: SOME LOOSE THREADS

In this section, we evaluate two representative studies from the color-research tradition. Our goal is to identify several ambiguous issues that motivate the experimental work to follow.

Despite the shift in interpretive framework, the "color-research tradition" has been unified by a more or less consistent methodology. The basic procedure involves using samples of colors to elicit two sets of responses: one "cognitive," usually recognition memory, and the other "linguistic," usually some measure (typically referred to as codability) of the various properties of expressions that people use to denote the sensation of color. The pattern of association between the two types of responses is then taken as evidence for or against some specific hypothesis concerning the relationship between language and thought.

As noted earlier, color research can be divided into two periods or, perhaps more accurately, two schools of thought. To characterize better these two schools of thought, the critical experiment from each period will be described and evaluated. The work of Lantz and Stefflre (1964) brought together into a unified framework a series of somewhat discrepant findings from the first period. Except for the replication of their study in a cross-cultural context (Stefflre, Castillo Vales, and Morley 1966), it represents the last empirical effort of the first period. A study by Heider (1972) is regarded by most researchers as the definitive study of the second period; certainly from an experimental point of view, it is the most sophisticated study in establishing the influence of perceptual saliencies in the color spectrum on linguistic forms. Both studies contain careful reviews of the literature and are quite sensitive to the implications of their work for earlier findings. They thus provide a good experimental and theoretical basis for replication and extension.

Lantz's and Stefflre's study actually consists of two parallel studies: one on ease of encoding and the other on similarity of encoding. We will be concerned only with the first of those: ease of encoding, or codability.

After carefully reviewing previous research in which codability is measured by inter-subject naming agreement and various indices of brevity, Lantz and Stefflre argue that the contradictory findings in previous research may be caused by the inadequacy of the measure of codability. Therefore, they propose a new measure of codability: *communication accuracy*. They argue that this measure will allow a superior prediction of memory because of a formal similarity between the two activities:

> We will view memory as though it were a situation in which an individual communicates to himself through time using the brain as a channel. This communication process can be approximated by having individuals communicate with other people. Items accurately communicated interpersonally would then be predicted to be more accurately communicated intrapersonally as measured by the usual memory tests. . . . Thus, for our measure of codability, people were presented with a stimulus array and asked to make up messages that would enable another person to pick out the stimulus it refers to . . . [Lantz and Stefflre 1964:473].

In their experimental design, Lantz and Stefflre contrast this new measure of codability, communication accuracy, with the two measures used previously: intersubject naming agreement (Brown and Lenneberg 1954) and brevity (Glanzer and Clark 1963a,b). On the cognitive side, they look at recognition memory using three different recognition conditions: (I) 5-second presentation of 1 chip with a 5-second interval before recognition trial; (II) 5-second presentation of 4 chips with a 5-second interval before recognition trial; and (III) 5-second presentation of 4 chips with a 30-second interval before recognition trial. In addition, they employed both sets of color materials previously used in similar research: the Brown-Lenneberg array (1954) and the Farnsworth-Munsell array (Burnham and Clark 1955; Lenneberg 1961). Thus, within the experimental design, they are able to assess the importance of stimuli, procedures, and measures. The overall results show that communication accuracy is the most predictive linguistic index for recognition memory and that the results are strongest in the conditions where there are more chips or longer time intervals. Further, certain contradictory findings available in previous research are shown to be caused by the use of different stimulus materials since the Farnsworth-Munsell colors are poorly related to the naming and brevity measures of codability. For purposes of comparison, Lantz's and Stefflre's basic findings are presented in Table I.

Lantz and Stefflre are able to interpret the significant correlations between the naming and brevity measures and the second recognition condition as being caused by the correlation of each of these measures with communication accuracy. Controlling for this relationship reduces the two correlations to near zero (.06 and .07, respectively), leaving communication accuracy as the only significant predictor of recognition memory.

A subsequent project (Stefflre, Castillo Vales, and Morley 1966) investigated the extent to which communication accuracy and recognition memory varied from language to language (and from culture to culture). Using the Farnsworth-Munsell array, Stefflre, Castillo Vales, and Morley assessed communication accuracy in both Mexican Spanish and Yucatec Maya. Recognition memory was assessed using a 5-second exposure of 1 chip with a 30-second delay before recognition testing. (Note that this condition differs from I, II, and III above.) Stefflre, Castillo Vales, and Morley discovered that within each of the two languages, recognition memory was associated with communication accuracy ($r = .44$, $p = < .05$ for Yucatec; $r = .58$, $p = < .01$ for Spanish). While the worst colors

TABLE I. CORRELATIONS BETWEEN CODABILITY MEASURES AND RECOGNITION CONDITIONS (FROM LANTZ AND STEFFLRE 1964:477).

	Recognition condition		
Codability	I	II	III
Farnsworth-Munsell:			
Communication accuracy	.32	.71***	.66**
Naming	−.18	−.05	−.30
Brevity	−.29	.31	.23
Brown-Lenneberg:			
Communication accuracy	.51**	.86***	.78***
Naming	−.02	.40*	.32
Brevity	.19	.42*	.33

*$p < .05$
**$p < .01$
***$p < .001$

(in terms of communication accuracy) were the same for the two languages (implying some cross-linguistic similarity), the best colors differed between the two language groups (implying some degree of cultural or linguistic relativity). Overall, the communication accuracy scores for Yucatec speakers were unrelated to the communication accuracy scores for Spanish speakers ($r = .11$, n.s.). The same was true of the recognition memory scores ($r = .06$, n.s.). In other words, as Stefflre, Castillo Vales, and Morley note, the Spanish and Maya groups "found different colors easy to remember and easy to communicate" (112).

Two types of criticisms have been directed against Lantz's and Stefflre's (1964) and Stefflre's, Castillo Vales's, and Morley's (1966) research on communication accuracy. Brown (1976:145) has argued that communication accuracy is entirely

> a psychological and not a cultural variable. [For a variable to be "cultural," it must be concerned with "shared knowledge and behavior;" (Brown 1976:145).] It is directly concerned with individual differences. One person might very well be a highly skilled encoder (or decoder) for a given domain and another very unskilled, though both have the same native language. Communicability scores of individual colors would depend partly on the language, but also on salient objects in the environment.

There are grounds for not being completely satisfied with Lantz's and Stefflre's communication accuracy measure, but to us Brown's particular objection lacks force. Certainly *any* measure (latency of response, brevity of the expressions used to label colors, amount of agreement in the use of referential expressions by different subjects) can potentially be examined from the point of view of individual differences, though that hardly seems like a weakness. Moreover, it is *not* individual differences that Lantz and Stefflre investigated; in fact, they control for individual variation in encoding and decoding. What Lantz and Stefflre and Stefflre, Castillo Vales, and Morley do investigate is the overall (or average) referential accuracy of members of a speech community. Quite properly, their concern is with language use as it might occur in ordinary communicative situations. Thus, their view of language is not restricted to dictionary entries (e.g., "blue") but, rather, includes all the expressions, presuppositions, and background knowledge (e.g., "crayon flesh"; are crayons covered with flesh?!) that make it possible for members of a speech community to understand one another. It is noteworthy that recent work in psycho- and sociolinguistics as well as in the philosophy of language has broken down the distinction between the structural aspects of a linguistic code and the contextual and knowledge factors that support comprehension (Searle 1969; Ziff 1972; Bransford and McCarrell 1974; Goffman 1976).

The communication accuracy measure does suffer, however, in that it does not tell us anything about what aspects of language are important in the communication process. Whereas Lantz and Stefflre do control for individual variation in encoding and decoding, they never extract any picture of the patterned regularities that might be responsible for communicative success. We are left with the black box of "communication accuracy," which may rest upon a variety of factors of the speech situation.

A second objection to the research on communication accuracy concerns the finding that the delayed recognition scores for speakers of Yucatec and Spanish are not correlated. Heider (1972) does not dispute the importance of Stefflre's, Castillo Vales's, and Morley's findings but she does point out that the Farnsworth-Munsell color array used by Stefflre et al. is a low- "saturation" array containing no "focal" colors. She argues that both Yucatec and Spanish speakers would probably have had superior memory for focal versus nonfocal colors, regardless of their linguistic and cultural differences. The reasonableness of that prediction is later examined.

The work of Lantz and Stefflre was not continued by others but was supplanted by a new line of research owing much of its impetus to the publication of Berlin's and Kay's

Basic Color Terms (1969). Berlin and Kay argue that there is a universal semantic and evolutionary hierarchy for "basic" color terms (e.g., red is basic; maroon is not)[2] such that if a language has N basic color terms, their referential foci can be predicted simply on the basis of that number. Berlin's and Kay's theory spawned a series of studies into individual color lexicons to see whether they fit the hypothesized hierarchy. That line of research eventually came to consider the possible biological or psychophysical basis for the universal hierarchy. Since the findings of that research resulted in claims about the perceptual unevenness of the type of color array used by Brown and Lenneberg (1954) and Lantz and Stefflre (1964), it eventually began to present a serious challenge to the findings of the earlier period of research. The crucial study that attempts to evaluate the role of perceptual saliences in the color arrays in this later period is Heider's (1972).

Heider is specifically concerned with Berlin's and Kay's "focal" colors—colors that are best examples of basic color terms across most languages. Heider argues that "focal" colors are perceptually salient and that this saliency leads in turn to greater memorability (recognition memory) and greater codability (brevity). Since the perceptual salience is presumably universal, so are the focal colors, and so, in turn, are the referential bases for the basic color terms in all languages. Heider conducts four experiments in support of that hypothesis.

In the Western scientific tradition, hue, value (= brightness), and saturation (= purity) are the three psychophysical dimensions necessary to describe completely the color space. In Experiment 1, Heider shows that there is a relationship between saturation and basic color terms. She first identifies the hue and value combinations for each of Berlin's and Kay's focal colors. (All colors in their array are at maximum saturation.) She presents subjects with all the levels of saturation available for each such hue-value combination and finds that 90–95% of all choices of the best example of the basic color term associated with that focal are from the two highest levels of saturation. In Experiment II, Heider attempts to show that focal colors are most codable across languages. She finds that for 23 languages (one speaker each) the focals are more codable on several indices of brevity (number of words, number of letters, and response latency) than a selection of nonfocals taken from the array. In Experiment III, focals and nonfocals are contrasted in a short-term recognition memory experiment conducted in two vastly different languages—English and Dani (New Guinea). (Since Heider finds no differences between her two subcategories of nonfocals, we will group them all as "nonfocals" throughout this paper.) Her findings (see Table II) show that the category of focal chips is easier to remember than that of nonfocals. Finally, in Experiment IV, the findings for short-term recognition memory are extended to long-term memory. Heider shows that subjects do better with focals than nonfocals in a paired-associates learning task administered over several days; she argues that superior performance on this task is indicative of better long-term memory. Heider concludes at the end of these four experiments that

TABLE II. SUMMARY OF HEIDER'S FINDINGS ON FOCAL/NONFOCAL DIFFERENCES IN RECOGNITION MEMORY (ADAPTED FROM HEIDER 1972:16).

	Focal colors (n = 8)	Nonfocal colors (n = 16)
English: Number correct	5.25 (out of 8)	5.73 (out of 16)
Correct-response latency	6.0	11.15
Dani: Number correct	2.05 (out of 8)	1.18 (out of 16)
Correct-response latency	2.5	3.25

> The same colors were most codable in a diverse sample of languages . . . , and those colors were best remembered even in a culture where the linguistic factors of hue terms and codability were absent . . . [Heider 1972:19].

She attributes this superiority to the inherent saliency of the focal colors.

Heider wavers on whether or not her study is one in language and thought. In her introduction, she says she is studying "the relation between a linguistic domain and the nonlinguistic domain which it encodes" (1972:11) which is not "language and thought." Yet, in her abstract she holds that on the basis of her findings "linguistically causal interpretations of earlier language and cognition studies using color were challenged" (1972:10), and in her final conclusion she writes:

> . . . the color space would seem to be a prime example of the influence of underlying perceptual-cognitive factors on the formation and reference of linguistic categories [1972:20].

Given the latter two comments, there can be little doubt that she intends to criticize earlier research. It is noteworthy, therefore, that she does not test directly to see whether memorability and codability are more highly related to each other than would be predicted on the basis of their individual correlations with focality.

Heider does gather extensive cross-linguistic data showing similar memory patterns (better retention of focals) in two societies with different linguistic codes (Dani and English) and similar encodings (in terms of brevity) across a wide spectrum of languages. This suggests that we must take seriously the claim that focality may be a factor independently regulating codability and memory. Studies from the earlier period do not provide direct information on such a possibility. There is, however, a major weakness in the way Heider makes use of the Dani as a control group.

The Dani have only two "basic" color terms, but, as Berlin and Kay note (1969:5), "every language has an indefinitely large number of expressions that denote the sensation of color." "Basic" color terms (terms such as "white," "red," and "green") comprise but a *very* small subset of the linguistic resources of a speech community for talking about color (see, e.g., Lienhardt 1961). By Berlin's and Kay's criteria, none of the following English expressions would qualify as a "basic" color term: olive, peach, maroon, pea-soup green, salmon, lemon, lilac, lavender, sapphire, emerald, blond, grass green, sky blue, bright navy, pale yellow, aqua, crimson, like an apple, etc.[2] Clearly, the number of basic color terms in a language is not indicative of the ability of a culture to represent a color sensation linguistically. The fact that the Dani have but two "basic" terms tells us little about their ability to generate a linguistic expression for denoting colors. This point takes on special importance when one considers that: (a) the only measure of codability that is known to relate consistently to recognition memory is "communication accuracy" (Lantz and Stefflre 1964; see above); and (b) "communication accuracy" is probably independent of either the brevity of a term or the criteria for identifying an expression for color as a "basic" term (see, e.g., Brown 1965:340; see also Appendix 5 of this study). It would seem that Heider's selection of the Dani as a control group was motivated by a commitment to an inappropriate codability measure (that is, *basic* color *terms*).

We can now see that the empirical basis for the shift in opinion outlined in our introduction is somewhat shaky. The most glaring deficit is the absence of any direct contrast of focality and codability as predictors of recognition memory. Lantz and Stefflre do consider the influence of the type of color array but do not examine the competing hypothesis that focal versus nonfocal properties of the color stimuli might explain their results (see Figure 1). Heider shows that focal colors are more codable on some indices and more memorable also, but she neither tests to see whether memorability and codability relate more strongly than their joint correlations with focality would suggest (see Figure 1), nor does she utilize the one linguistic index (communication accuracy) that is known to be a powerful and stable predictor of recognition memory.

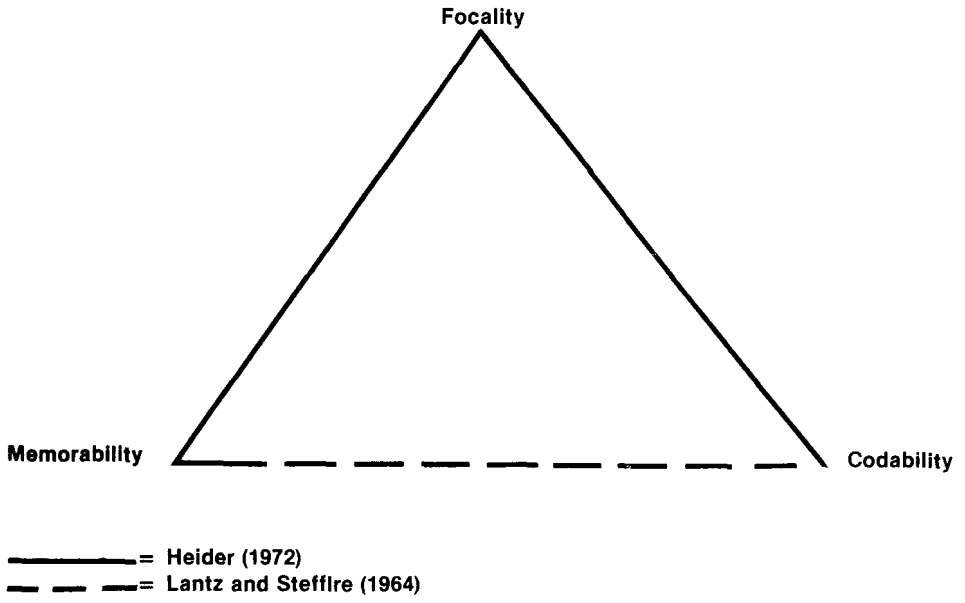

Fig. 1. Triad of relations between focality, codability, and memorability, showing the assessments made in previous research.

The experiments that follow attempt to investigate the interrelations among the triad of variables portrayed in Figure 1. The first three experiments, which retest Heider's findings, involve analyses and modifications of her critical procedure employed with her Dani and American samples. The last two experiments extend and modify Lantz's and Stefflre's approach by investigating communication accuracy in relationship to the focal/nonfocal distinction, by exploring the bases of subjects' performances, and by creating a new, less individual, index of communication accuracy. The entire set of experiments allows an examination of the interrelations among language, thought, and external stimuli.

EXPERIMENTS ON FOCALITY, MEMORABILITY, AND CODABILITY

Experiment I: Discriminability During Perceptual Search, or Recognition Memory?

Purpose

Heider found a difference in recognition memory between focal and nonfocal colors. Both accuracy and latency scores show this effect. Simple inspection of the array, coupled with an informal attempt to match visible probes with their counterparts in the array, suggested that the focals might have an advantage in both speed and accuracy during the process of perceptually searching through the array, independent of any memory effect. While Heider had computed a discriminability score (1972:17) on the target chips relative to their neighbors in the array to assure that this potential biasing factor was *not* influencing her results, informal inspection seemed to belie that conclusion. Moreover, Heider's discriminability scores were ultimately derived from discriminability tasks administered under experimental conditions unlike her own (see Newhall, Nickerson, and Judd 1943). Hence the present experimental test for a possible biasing relationship between focality and discriminability in Heider's "Dani array."

Method

Eighteen university students (12 women and 6 men) were given a perceptual matching task using Heider's materials. (Appendix 1 presents a full list of the color chips used by Heider in her memory experiment, as well as a discussion of our use of those materials.) Each subject in turn was presented with each of 25 probes and asked to find the exact same chip in the full 160 color-chip array. (See Appendix 1 for an explanation of why one extra chip was present at this stage of the experiment. It plays no part in the results discussed below.) Subjects were instructed to match both as quickly and as accurately as they could; both accuracy and latency scores were then recorded. By leaving the probe in full view during the search period, we hoped to stimulate "perfect memory," so that we could assess whether Heider's results might simply come from effects arising during the perceptual search process and not during the memory process.

Results

The evidence on the relative discriminability of focal versus nonfocal chips can be analyzed in either of two ways: by subject or by chip. Analyzed by subject (that is, using paired scores), focal chips were found to differ significantly on both measures of perceptual search. Focal chips were located significantly more accurately ($\bar{X}_f = 78.4\%$, $\bar{X}_{nf} = 64.5\%$, $t_{17} = 3.90$, two-tailed $p = < .001$) and more quickly ($\bar{X}_f = 4.46$ seconds, $\bar{X}_{nf} = 5.34$ seconds, $t_{17} = 3.11$, two-tailed $p = < .01$) *even with the probes visually present at all times.* The analysis of latency scores, however, should be interpreted with caution for several reasons. Our analysis focuses only on latencies for correct choices. Using all latencies (correct and incorrect choices) would be misleading and uninterpretable since many of the figures would represent the time to select nontarget chips. However, although using correct latencies guarantees that each chip receives scores it is entitled to, the possibility remains that the various chips and subjects are unequally represented. Hence, neither measure is satisfactory. Fortunately, the basic measure in most color research, including Heider's, is the accuracy score (number or percent correct); it is this measure that we shall focus upon in our experiments.

Analyzed by chip, the difference in perceptual search accuracy for focal versus nonfocal colors falls just short of significance ($t_{22} = 1.77$ pooled variance analysis, $p = < .10$; $t_{16.7} = 1.79$ separate variance analysis, $p = < .10$). This pattern of findings is summarized in Table III. The data for Experiment I are summarized in Appendix 2.

Discussion

It seems to us that caution is called for in interpreting Heider's Dani study. The overall

TABLE III. ACCURACY SCORES (PERCENTAGE CORRECT) FOR PERCEPTUAL SEARCH (DISCRIMINABILITY) TASK (EXPERIMENT I).[a]

$\bar{X}$ focal	78.4
$\bar{X}$ nonfocal	64.5
Three t-test analyses of focal/nonfocal differences	
t_{17} (by subject - 18 paired scores)	3.90
p (two-tailed)	$<.001$
t_{22} (by chip-pooled variance)	1.77
p (two tailed)	$<.10$
$t_{16.7}$ (by chip-separate variance)	1.79
p (two-tailed)	$<.10$

[a] Accuracy under conditions simulating perfect memory.

pattern of our results suggests that the differences observed by Heider in her short-term recognition memory experiment may be in part because of a difference between focal and nonfocal chips at the point of perceptual search during recall. What this means is that, even if the Dani (or anyone else) had perfect recall, Heider's stimulus array is such that it is more difficult to identify nonfocal than focal target chips. It is important to stress that this perceptual discriminability factor is *not* the same as Heider's "saliency," which is an *intrinsic* property of the color, not a property dependent on neighboring chips. She takes great care to exclude this factor of perceptual discriminability as a possible source of bias in her data since it is known to correlate with recognition accuracy. Our results suggest that the use of discriminability measures computed on the psychophysical properties of the color chips (hue, value, and saturation) is not an adequate control and that direct behavioral measures must be taken.

The next experiment involves controlling for the unequal discriminability of focal and nonfocal chips by equating focals and nonfocals at the point of search so that a proper test for an independent effect because of memory may be administered.

Experiment II: Construction of an Array Controlled for Discriminability

Purpose

In this experiment, we try to equalize the error and latency scores for the focal and nonfocal chips while preserving the maximum comparability with Heider's materials.

Method

Equalization of search times and error rates was accomplished by eliminating selected chips from the existing array.[3] The procedure for elimination involved discarding chips that were erroneously selected as targets two or more times in the original search task. Thirty-one chips met that criterion without complications. Two chips remained that also met the criterion but were themselves target chips (two nonfocals) from the original set of 25. Rather than eliminate these two chips, a third target chip (also a nonfocal) that was the *source* of the errors for the other two was deleted; this brought the other two target chips below the criterion. This deletion was a superior procedure because it equalized the number of targets (at eight each) for the focals and Heider's two categories of nonfocals, and it allowed the construction of a rectangular array. Rearrangement of the array was necessary because removal of 32 chips left "holes" in the original array of colors. Hence, the remaining 128 chips (left after eliminations) were randomly reordered into an 8-by-16 grid. This new array had the advantage of randomizing the effect of neighboring chips on identification of targets; the original array had not provided such randomization, and the very obvious "structure" in that array undoubtedly influenced the search times for the various colors.

The procedure of Experiment I was then replicated using this new array with ten subjects. This new array was not perceptually equal either. With the probe in full view, all ten subjects were more accurate in identifying focal chips ($\bar{X}_f = 91.3\%$, $\bar{X}_{nf} = 73.9\%$, $t_{(9,\ \text{by subject})} = 8.83$, $p = < .001$). Latency scores also continued to favor focal chips ($\bar{X}_f = 7.91$ seconds, $\bar{X}_{nf} = 9.96$ seconds, $t_{(9,\ \text{by subject})} = 2.48$, $p = < .05$). Consequently, the above procedure for making the array discriminitively fair was repeated a second time by discarding a further 8 chips, effectively deleting 1 row (substituted into the 8 openings), leaving a final array of 8 by 15. This array was tested on 15 university students (9 men and 6 women).

Results

The array just described showed no significant difference between focals and nonfocals

either on percent correct ($\bar{X}_f$ = 89.4%, $\bar{X}_{nf}$ = 82.2%, $t_{(15,\ by\ subject)}$ = .34, n.s.) or on latency for correct choices ($\bar{X}_f$ = 7.76 seconds, $\bar{X}_{nf}$ = 8.44 seconds, $t_{(15\ paired\ scores)}$ = 1.49, n.s.). Note that what differences do exist show the focals as faster and more often correct. (The array developed in this experiment is listed in full in Appendix 3.)

Discussion

The results of the experimental test confirm that the final 120-chip array constructed by elimination and randomization showed no statistiscally significant difference between focal and nonfocal chips in error rate or in latency on correct responses. (It is interesting to note that randomization alone would not have created an unbiased array since the bias persisted after the first rearrangement.) This array, then, presents a fair set of materials with which to test for the hypothesized superior memorability of the focal colors.

Experiment III: Recognition Memory for Focal and Nonfocal Colors

Purpose

Experiment III attempts to: (1) replicate Heider's experiment concerning the difference between focal and nonfocal colors in short-term recognition memory by using the new array developed in Experiment II; (2) test for a difference between focals and nonfocals in long-term memory with a more straightforward measure than Heider's paired-associates learning task; and (3) explore through informal observation and questioning the nature of subjects' errors and their subjective understandings of their memory techniques.

Method

There were two experimental conditions for each subject: a short-term one and a long-term one. The basic procedure for both conditions was the same as that employed by Heider: a subject was presented with a chip to view for a set amount of time; then the subject passed a certain length of filler time without the chip in view; and finally, the subject tried to select (recognize) the target out of the full color-chip array when it was uncovered. In the short-term condition, presentation was for 5 seconds and delay for 30 seconds, thus replicating Heider's timing. In the long-term condition, presentation was for 1 minute and delay for 30 minutes. The array and probes used were those developed in Experiment II.

Subjects were 24 students (12 men and 12 women; 21 undergraduates and 3 graduates) residing in a university housing unit (thus making feasible the laborious long-term experiment since the subjects were allowed to do as they wished during the 30-minute intervals). All subjects were tested in a small room (college dormitory decor), seated across a desk from the experimenter with the covered array on the white desk top. Lighting was bright, from multiple sources, and identical at the time of presentation and recall.

For each subject, the short-term condition was administered first and the long-term condition second. Each subject received 18 of the 24 probes in the short-term condition (6 focals and 12 nonfocals) and 6 of the 24 probes in the long-term condition (2 focals and 4 nonfocals); within that presentation frame, the chips were randomized for each subject with the restriction that all chips were exposed an equal number of times across all subjects. Filler time during the 30-second condition was occupied by silence or, more usually, by idle conversation. In the long-term condition, where the filler time interval was 30 minutes, subjects were allowed to leave the experimental area and do as they wished. They had to return after 30 minutes to make an identification and to receive another presentation. At the conclusion of the final recognition trial, or on subsequent occasions, the pur-

pose of the experiment was explained to those who cared to know, and the subjects were asked to comment on the experiment (e.g., difficulties, strategies, etc.).

Results

Table IV presents the accuracy scores for both memory conditions. As Heider notes, the basic measure in this experiment is the accuracy score (number, or percent, correct); problems with the latency score were mentioned in Experiment I. Again, the data were analyzed both by subject and by chip. In the short-term memory condition, the focals and nonfocals are *not significantly different* as measured by the "by-subject" analysis (t_{23} = 1.52, n.s.) or as measured by the "by-chip" analysis (t_{22} = 1.01 pooled variance, n.s., and $t_{14.7/21.5}$ = 1.03 separate variance, n.s.)(see Table IV). Converting the pooled variance analysis to its associated correlation coefficient by the usual formula (see Welkowitz, Ewen, and Cohen 1971) of $r = \sqrt{t^2/(t^2 + df)}$, we derive r = .21 (n.s.). To compare the short-term accuracy results in Table IV with Heider's results in Table II, we need only use our percent correct figure to compute how many chips Heider's subjects would have gotten correct had they succeeded at the same rate as our subjects. To ease the comparison further, since there are twice as many nonfocal as focal probes, the nonfocal score can be scaled down by one-half (see Table V).

In the long-term memory condition, a significant difference between focals and nonfocals emerges for the first time in the "by-subject" analysis (t_{23} = 2.64, p = < .05). However, the two "by-chip" analyses are not significant (t_{22} = 1.45 pooled variance, n.s. and $t_{14.7/21.5}$ = 1.86 separate variance, p = < .10) (see Table IV). Converting the pooled variance analysis to its associated correlation coefficient by the method discussed earlier, we derive r = .29 (n.s.).

A separate t test (paired scores) indicated that the long-term condition was significantly more difficult (t_{23} = 5.15, two-tailed p = < .001) than the short-term condition. (Summary data for Experiment III are given in Appendix 4.)

Discussion

Looking at the short-term memory results, we see that Heider's finding of superior memorability for focals *is not replicated* once the perceptual discrimination for the two categories of chips is equalized. The comparisons in Table V allow us to see that the major difference between our findings and those of Heider resides in the markedly improved accuracy scores for the nonfocals. This suggests that the main effect reported by Heider is

TABLE IV. ACCURACY SCORES (PERCENTAGE CORRECT) FOR RECOGNITION MEMORY EXPERIMENT (EXPERIMENT III).

	Short-term memory	Long-term memory
$\bar{X}$ focal	62.5	58.3
$\bar{X}$ nonfocal	53.7	41.6
Three t-test analyses of focal/nonfocal differences		
t_{23} (by subject - 24 paired scores)	1.52	2.64
p (two-tailed)	(n.s.)	<.05
t_{22} (by chip-pooled variance)	1.01	1.45
p (two-tailed)	(n.s.)	(n.s.)
$t_{14.7/21.5}$ (by chip-separate variance)	1.03	1.86
p (two-tailed)	(n.s.)	<.10

TABLE V. COMPARISON OF SHORT-TERM MEMORY SCORES IN EXPERMENT III AND IN HEIDER (1972).

Accuracy (Number correct)	Heider	Experiment III
$\bar{X}_f$	5.25	5.00
$\bar{X}_{nf}$	2.87	4.30

because of the superior discriminability of the focals in the array and not to any "saliency" inherent in the colors themselves. Once the discriminability of the nonfocals is brought up to the level of the focals, the recognition scores become roughly equivalent also. The focal chips in Heider's array profit in recognition memory from superior discriminability from their neighbors.

Under the long-term memory condition, the significant advantage of focals in accuracy scores expected by Heider does emerge on the unbiased array (although in a marginal way). However, the relationship of focality to codability and codability to recognition memory remains to be analyzed in Experiments IV and V.

In an attempt to discover what factors might be involved in the memory task, clues were sought in the comments and incidental behaviors of the subjects. Three methods emerged as the dominant strategies among our subjects.

One strategy was basically verbal. Subjects generally labeled the color chips with one of the primary color terms (e.g., "red," "blue") and then supplemented this with other stored information to distinguish the particular variant at hand. One method of supplementation involved simply storing "visually" the idiosyncratic aspects of the chip. It is noteworthy that few subjects reported using pure visualization without the verbal element. Two who did (one reporting that she tried to get the "feel" of each color) were among the most accurate. A second method of supplementation was to append to the primary term a more involved verbal modification (e.g., "pea-soup green but a little lighter"). Different subjects made these added descriptors more or less complex; some subjects varied their own complexity in light of their general memory for how many similar confusable colors they recalled seeing in the array. Thus, for example, knowing that there were many greens and few reds, the subjects would make more complex descriptions for the greens than for the reds. These latter subjects were among the most accurate. An interesting aspect of this strategy (emerging especially under the long-term condition) was the disastrous effect of forgetting the primary term. Without memory for the basic term (e.g., "it was a blue"), subjects had absolutely no idea which color, or type of color, they had seen. Subjects just stared hopelessly at the array. Subjects who had the primary term were quite comfortable and much less often had a sense of not remembering the specific chip, even when they were quite far off in their matches.

A second strategy was to relate the color to some colored object stored in permanent memory (e.g., "color of my mother's lipstick," "color of my bedroom"). These objects were often verbally encoded, and it appeared that the verbal name for the object was used in memory. Subjects using this strategy, then, were often using a verbal code to elicit a stored object memory from which they derived the needed color. The approach seemed distinct from working completely within a semantic color system and from using free-floating visualization. Subjects apparently felt fairly confident that these stored object colors were reliable for repeated comparison as required in the memory experiment. In many cases, the matches to famous colors (e.g., "Kodak yellow," "Pepto-Bismol pink") were quite compelling in their aptness.

A final strategy involved visually matching the probe with some object present in the

room (e.g., an article of clothing). Then, at time of recall, the subject would try to find a color in the array that was either identical to the object or that differed in some way from the object that they had also stored (perhaps verbally) (e.g., "a little bluer than the green of my shirt"). Under experimental conditions, that strategy is not terribly helpful and was not frequently employed. Whenever a subject did mention using this method en route, the use was noted: more often than not, the subject made an incorrect selection. Nonetheless, this is undoubtedly one of the more common everyday techniques for remembering a color—similar to taking a swatch of cloth to the store to match better a dress, curtains, upholstery, etc. Experiments conducted under highly artificial conditions (that is, with all-neutral room and clothing) or without careful observation and questioning of subjects might not detect this strategy. (Note that, for focals to outperform nonfocals with this strategy, they would have to be more available in the environment.)

If we conceive of a color as a bundle of information to be stored, the subjects appear to utilize a "standard" (a verbal code, a remembered object, a present object) to encode in memory the bulk of the information. Then they only need to actively store the much smaller quantity of information necessary to distinguish the color at hand from the standard. We may, after Stefflre, regard memory to be much like communication except that the information is not conveyed from speaker to hearer but from time of memory to time of recall/recognition. Since conventional forms (culturally organized natural language) are dominant in interpersonal communication, it is natural to ask to what extent such conventional forms are important in intrapersonal communication, that is in memory. That would amount to assessing the extent to which each of the three standards is likely to be conventional. Present objects are not likely to be conventional as the strategy emerges here. Remembered objects are quite likely to have a normative aspect although whether we can call that conventional or not is open to question. A verbal code, of course, would immediately implicate conventional forms. Further, verbal supplements are apparently used in conjunction with the other two standards as well. In short, there is good reason to suppose that the search for an influence of language on memory for color would produce positive results in light of our subjects' reports. Experiments IV and V explore that possibility.

Before leaving our analysis of subjects' self-reports, two more items deserve mention although they are not taken up further in what follows. First, subjects had strong emotional or subjective reactions such as "ugly," "pretty," "dull," etc. to specific colors. Further, many subjects responded to the color-memory procedure as if it were a test of intellectual or artistic giftedness. That is perhaps understandable given the university student population and the general association of psychology with abilities testing. It is not clear to what extent such emotional and motivational factors may influence the experimental results.

Second, some aspects of subjects' decision making are worthy of note. Subjects did not appear to be more likely to be correct in their selections simply because they felt "sure" of the color. No evidence of a reliable relationship between feelings of "sureness" and performance was evident in accuracy although subjects did tend to respond more quickly. It was also very common for subjects to narrow the field to two colors as possible correct choices and then to select the wrong one of the two. It may be that a more sensitive measure, such as one indexing perceptual distance of the selection from the correct choice or asking for a second choice, would produce more interesting results than a simple correct/incorrect scoring. (An approach using that notion in a different setting is undertaken in Experiment V.)

Experiment IV: Memory and Two-Person Communication Accuracy

Purpose

This experiment explores the possible influence of linguistic encoding on memory for

color. The linguistic measure employed is the accuracy of communication using the color chips in a two-person communication task. The specific goal is to see whether an index of this sort correlates with memory performance and, in particular, whether it is superior to focality as a predictor of such performance.

Method

The basic design involved having one subject (an encoder) describe verbally a color chip so that a second subject (a decoder) could select the chip from the array. Encoder and decoder sat on opposite sides of a table with the array between them. The experimenter indicated the target to the encoder, while the decoder turned away for a moment. Then the encoder had up to five opportunities to describe the chip in such a way that the decoder could locate it in the array. Encoders turned away while decoders indicated their choices. Encoders were instructed not to use locative descriptions and not to make references to any objects of similar colors in the room. The target chips and array were those used in Experiments II and III. A record was kept of the number of trials necessary for successful location of each chip; in some cases, a record was also kept of the verbal descriptions used. Twenty-two pairs of subjects were tested.

Results

Since we were interested not in a dichotomous variable (such as focality) but in a hypothesized continuum of communicability, correlation coefficients were used to evaluate the results. These correlations have the advantage of indicating the strength of the relationship as well as its significance. Such an index of relative strength will be important in comparing the various predictors of memorability.

Two scores were computed for each chip: one based on the number of times the decoder selected the correct chip on the first try (number of first hits) and the other on the mean number of tries necessary for identification (with a maximum of 5) across all 22 pairs. Each of these communication accuracy scores is significantly correlated with the short-term and long-term memory accuracy scores from Experiment III, with correlations ranging from .54 to .58. These figures are presented in Table VI. Since, even in our equalized array, a spurious correlation might have existed between communication accuracy and perceptual search accuracy (see Experiment II), partial correlations were computed for each accuracy value controlling for this factor. In no case in any of our experiments did the partial correlations substantially reduce the size of our original correlations or change the significance values. (Summary data for Experiment IV are available in Appendix 6.)

Discussion

Whereas focality did not relate at all to short-term memory in Experiment III, com-

TABLE VI. TWO-PERSON COMMUNICATION TASK: CORRELATION COEFFICIENTS (R) BETWEEN MEASURES OF TWO-PERSON COMMUNICATION ACCURACY AND MEMORY ACCURACY (N = 24 CHIPS).

Communication accuracy	Short-term memory	Long-term memory
First hits	.58***	.54**
Mean tries (−r)	.54**	.58***

** two-tailed $p < .01$
*** two-tailed $p < .001$

munication accuracy does in Experiment IV. In fact, in both memory conditions there is a significant relationship between communication accuracy and recognition memory. (Of the two measures of communication accuracy, first hits are the more valid in comparison with memory strategies and performance since in the memory task, second choices were not elicited.) This suggests either that language itself is involved in memory or that the factors that influence communication accuracy also influence memory. The significance of that finding is discussed after we present the following results of Experiment V.

Experiment V: Memory, Group Communication Accuracy, and Referential Confusability

Purpose

In Experiment III, subjects utilizing a linguistic description in memory commented that much of their difficulty in remembering certain chips had to do with the greater number of alternatives in the array for certain hue names. For example, an encoding such as "light green" might be applicable to many more chips in the array than an encoding such as "dark orange."[4] More formally, a single linguistic description used in memory can have several appropriate referents at the point of recall. It was our hunch that the linguistic descriptions typically used by subjects would vary in the range of their applicable referents in the array and that this difference in range, the degree of *referential confusability,* would be highly correlated with the memory results of Experiment III. Experiment V investigates that possibility. Further, by a slight addition to the procedure, we are able to get an alternative measure of communication accuracy. This new measure of communication accuracy is derived from the descriptions typically used by subjects and is therefore more clearly independent of the skills of individual encoders and decoders and the immediate context.

Method

The first step in assessing referential confusability was to find out exactly what descriptive phrases were used by subjects. To gather a collection of descriptions, 29 university subjects (15 men and 14 women) were asked to describe the target colors so that another person could identify the color if he had to select it from an (unspecified) array of colors. The only restriction on subjects was that they limit their descriptions to three words or less.[5] Having in hand this large collection of descriptions, some method was needed to derive from them a basic description for each target so that we could assess how many chips in the full array might meet the description. Alternative procedures, such as giving subjects each individual description to decode and computing mean values for each target, were deemed too tedious, given the 696 individual descriptions involved. Even some method of counterbalanced subsampling, such as that used by Lantz and Stefflre, would have required about 180 decodings for each subject, thus making the task extraordinarily lengthy and boring. Both of those methods also camouflage the linguistic variable at work.

Preliminary attempts to reduce the 29 descriptions available for each chip to a single basic description by simple mechanical procedures (e.g., adding up the number of usages of each lexical item) produced meaningless results. It became obvious that some account would have to be taken of the syntactic relations among the elements within a description and that a certain amount of subjective judgment would have to be exercised in certain instances (e.g., Are *army green* and *military green* the same? Does *pea-soup green* involve one modifier of *green* or two?).

The following procedure was used for the set of 29 descriptions associated with each

chip. The last term of each description was taken as the basic term (e.g., if the description was *royal blue, blue* was the basic term). Once the basic terms had been listed, the modifiers were examined. If there was no modifier (e.g., the subject simply said "blue"), then one point was scored for that basic term in unmodified form. If there was a single modifier (e.g., "royal blue"), then one point was scored for the basic term with that modifier. If two (or more) modifiers were present, a decision had to be made about whether the two formed a compound modifier (e.g., *pea-soup green*), scored as a unit, one point for a single modifier; formed two independent modifiers (e.g., *greenish turquoise blue*), scored one-half point each; or formed embedded modification (e.g., *light royal blue*), scored the same as without the outer modification (i.e., *light royal blue* = *royal blue*).[6] (In some cases, there was ambiguity (e.g., *deep sea blue* could be interpreted as *deep-sea blue* or as *deep sea-blue*), which was resolved case by case. In most instances, the scores that mattered were not changed.) *Any basic term-modifier combination accumulating four points or more was counted as a component of the chip's "basic description."* Many chips had basic descriptions with more than one such component. Once there was a basic description for each target chip, we assessed how many other chips in the array could meet each description. Ten university subjects (six men and four women) were presented with the basic descriptions and were asked to pick out the best exemplar of the description and any other chips in the array that could be accurately described by the description. Rows and columns in the array were numbered so that each chip in the array was uniquely identifiable. The task was self-administered by some subjects and experimenter-administered to others. Three scores were available: (1) the mean number of alternatives applicable to the basic descriptions for a target chip, that is, what S's considered to be "reasonable choices" ("number of reasonable choices"); (2) the absolute number of different chips considered the best exemplar of the basic description, that is, number of different chips across subjects selected as "best example" ("number of best examples"); (3) the number of times subjects correctly selected the original target chip as the best example of the basic description ("target as best example"). This third score may be considered an alternative measure of communication accuracy—what we will call "group communication accuracy." Notice that, unlike the two-person communication accuracy measure, our "group communication accuracy" indicator reflects the efficacy of basic descriptors developed from one group's cumulated descriptions in guiding a second group's cumulative decoding efforts. A little later we discuss the empirical relationship between the two-person and group communication accuracy scores.

Results

Basic descriptions and their component modifier scores are listed in Appendix V. The 40 component descriptions used in forming the 24 basic descriptions account for 272 encoder responses out of the total of 696—that is, 39% of all encodings were represented in the basic descriptions given to the decoders. If we take the component descriptions with the highest modifier scores from each basic description and use them as an index of intersubject agreement on a particular chip, we find that intersubject agreement (a measure of codability often used in the past) does not correlate with either short-term memory ($r = -.20$, n.s.), long-term memory ($r = -.12$, n.s.), or focality ($r = .10$, n.s.). Thus, agreement on how a chip is best described does not assure high recognition memory.

The decoders selected the original targets 34% of the time in their indications of the best example of the basic description. This group-communication accuracy score correlated significantly with both short-term memory ($r = .59$, $p < .001$) and long-term memory ($r = .63$, $p < .001$) (see Table VII).

The mean number of reasonable choices for the basic descriptions ranged from 1.2 to 4.2. This score was correlated with the memory scores for the target chip that had been

TABLE VII. CORRELATION COEFFICIENTS (R) FOR MEASURES OF GROUP-COMMUNICATION ACCURACY AND REFERENTIAL CONFUSABILITY WITH MEMORY ACCURACY (N = 24 CHIPS).

Group-communication accuracy		
Number of times correct target selected as best example	.59***	.63***
Referential confusability		
Number of different targets selected as best example (−r)	.55**	.69***
Number of reasonable choices (−r)	.57**	.48**

** two-tailed $p < .01$
*** two-tailed $p < .001$

the original basis for the description. As predicted, these correlations were significant for both short-term memory ($r = .57$, $p < .01$) and long-term memory ($r = .48$, $p < .01$) (see Table VII). The number of different chips selected as best example provided an alternative index of referential confusability (that is, how many really close competitors there were—close enough to be selected as best example). The number of best examples ranged from two to eight and correlated significantly with both short-term memory ($r = .55$, $p < .01$) and long-term memory ($r = .69$, $p = < .001$) (see Table VII). However, this latter measure is independent with the number of correct selections of the target. (Summary data for all three of these indices based on decoder performance are presented in Appendix 6.)

Discussion

High intersubject agreement on basic descriptions is not in and of itself predictive of recognition memory. Nonetheless, the 34% success rate of the decoders indicates that distinctive information concerning the targets was maintained under the procedure for producing basic descriptions. This success rate compares favorably with the 40% success rate under the two-person communication task (where the array was available to the encoder) and with the 57% and 47% success rates under short- and long-term memory conditions respectively (where the decoders were working with their own personal memory encoding of the chip). That these subjects did so well with a standardized linguistic description formulated independently of the array is truly remarkable.

Previous findings of a relationship between communication accuracy and recognition memory (e.g., Lantz and Stefflre 1964) have been regarded to be unduly dependent on the abilities of individual encoders (or decoders) or to be dependent on linguistic descriptions formulated specifically for the color array at hand. Notice that the group communication accuracy score escapes those problems and yet produces the same general pattern of results as the more traditional two-person communication accuracy procedure. The correlation between the two measures of communication accuracy is substantial ($r = .81$, $p = < .001$). Previous findings have also been criticized because the color arrays utilized lacked focal chips; both of our experiments use an array with focal chips. Communication accuracy proves to have a far stronger relationship to recognition memory than does focality.

Figure 2 presents, in pictorial and numerical form, the triad of relations among focality, communication accuracy (both measures), and recognition memory (short- and long-term). In every case, the strongest leg of the triangle is the language-and-memory correlation. Further, even when the influence of focality is removed by taking a partial cor-

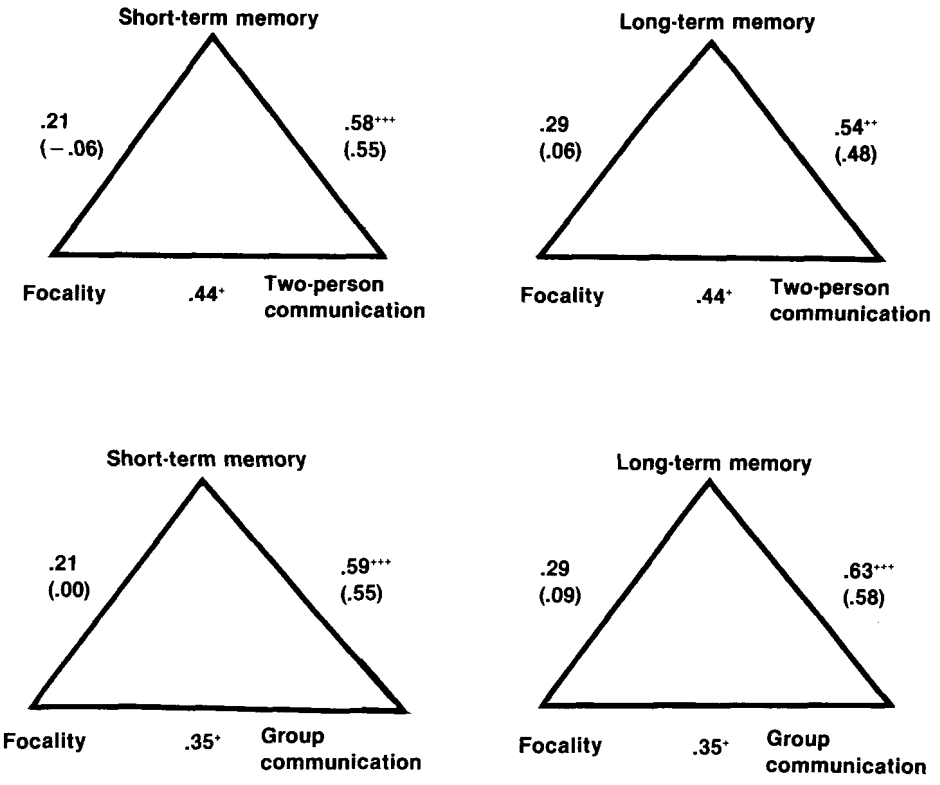

Fig. 2. Relationship between language indicators and focality to short- and long-term memory accuracy with the language variable or focality partialed out (in parentheses) (N = 24 chips).

relation (numbers in parentheses in Figure 2), the four correlations between language and memory retain their basic strength and significance. However, when the linguistic influences are partialed out, the predictive utility of focality reduces to close to zero. These results imply that linguistic encodings are used in color memory and that their efficacy is not dependent on focality.[7]

It would be a mistake to conclude that language is the only significant factor in color memory. The conclusions we stress here are only that language can serve as a highly effective vehicle for color memory, that it appears to operate independently of "focality," and that it need not be dependent on individual encoder or decoder abilities but may reflect regularities at the level of the speech community.

Our expectations regarding the confusability of a description are borne out: the size of a basic description's referential range (the number of chips in the array to which it could be reasonably applied) is a good predictor of both short- and long-term memory performance with the encoded chip. That is reasonable since, no matter how apt a description

is, it will be of little use if it is equally applicable to several other chips or more applicable to a nontarget chip.

In actual practice, a memory technique such as linguistic encoding is probably adaptive to variations in the nature of the stimuli to be encoded; thus, in the long run, it will be unprofitable to try to account for subject performance simply in terms of some property of the stimuli (e.g., focality) or some property of the memory vehicle (e.g., brevity of description). Communication accuracy is an effective predictor because it reflects *language as applied* to a particular set of stimuli, thus encompassing both the vehicle and the object within a single measure. If either the linguistic descriptions or the stimuli were changed, one would expect the communication accuracy scores (and by implication the memory scores) to change also (cf. Lantz and Stefflre 1964).

To indicate the operation of these linguistic variables in a more concrete way, it will be useful to examine in greater detail three of the most telling cases.

Chip 8 (a nonfocal) has the basic description "flesh/peach." Neither word is a "basic" color term in English; neither is the chip particularly perceptually salient in Heider's terms with its saturation level of 6, nor is it likely to be universal across languages, cultures, or races. Yet, this color was the single best-remembered chip in long-term memory and tied with three other chips as second best-remembered in short-term memory. Clearly this nonfocal color is well remembered. Our linguistic measures detail the potential sources of its strength. It has the most-agreed-upon basic description of any chip; this description has the second-fewest competing reasonable choices and the second-fewest number of chips selected as best example. The success of this chip apparently lies in its cultural and linguistic saliency to Americans (conventionalized by the Crayola Company) and in the lack of competition for its basic description in the array.

Chip 14 (also a nonfocal) has the basic description "light yellow/pale yellow." This color chip is not salient in terms of its saturation, which is low (six), nor would its description be cited as a basic term of English although it contains such a term in "yellow." Unlike "flesh," however, the description does not have any obvious cultural or racial significance and has only moderate intersubject agreement. Nonetheless, this chip is the best remembered of any chip under short-term memory and ties for second-best-remembered in long-term memory. Our measures reveal that its strength lies in the lack of competition in the array for this description. It has the highest number of correct choices of the original target as best example; it also has the fewest number of chips listed as reasonable alternative choices.

Finally, for contrast, we can look at Chip 22 (a focal) with the basic description "pink." Although the chip has low saturation (six), it is one of the focal chips and has as its description a basic color term on which there was high intersubject agreement. Despite those attributes, it has the second-worst score on short-term memory and is easily the worst-remembered focal on either long- or short-term memory. Its basic description exhibits a high number of reasonable alternative choices in the array and led only one subject to the original target. (The other subjects distributed their choices for best example among four other chips). Despite high intersubject agreement and a target that was a focal, this chip simply suffered from too much competition in the array—too many chips to which the description "pink" applied.

GENERAL DISCUSSION—WHORF AND THE COLOR-RESEARCH TRADITION

Our examination of the empirical basis for the reversal of opinion in the color-research tradition finds that empirical base deficient. Reviewing the literature, we find no direct contrast of the two views. Reanalyzing Heider's critical experiment, we find it to be based in part on artifact, and nonreplicable in the short-term memory condition. Conducting a

direct contrast of communication accuracy and focality as predictors of recognition memory, we find the linguistic measures to be far superior and the predictive value of focality to be near zero once these linguistic measures are taken into account. Given this evidence, the view of the earlier era regarding the importance of language in recognition memory seems more sound.

Further, we have introduced new evidence regarding the role of language as a factor in human color memory. Subjects' introspective accounts indicate that language is one type of "standard" used in memory and it is a useful supplement to other standards, such as permanently stored objects and images. Our second measure of communication accuracy indicates that this "standard" need not be idiosyncratic. Finally, it appears that the referential range of the linguistic description used is the critical element affecting performance. On balance, it seems that the metaphor forwarded by Lantz and Stefflre, that memory is analogous to intrapsychic communication with oneself over time, is especially apt in that the vehicle for this communicatin is often linguistic and in that memory success relates to the referential precision of the linguistic description at the time of recognition (decoding). We have also suggested how other vehicles for memory may operate in a similar fashion; certainly, language is not the only factor involved. To the extent that language *is* involved, however, it remains likely that different languages will induce different patterns of recognition. That conclusion is certainly supported by Stefflre, Castillo Vales, and Morley (1966) and by our evidence of group-communication accuracy.

It remains for us to reassess the relationship of color research to Whorf's original conception of the relationship between language, thought, and external stimuli. There are many ironies to be found in the color-research tradition. One is that color researchers (and others) have assumed that Whorf's linguistic relativism is incompatible with the existence of sublinguistic universals such as focality. It is ironic because Whorf not only knew better but explicitly cautioned against that misleading assumption:

> There is a universal *Gefühl*-type way of linking experiences, which shows up in laboratory experiments and appears to be independent of language—basically alike for all persons.
>
> Without a serial or hierarchical order in the universe it would have to be said that these psychological experiments and linguistic experiments contradict each other [1956:267, emphasis in original].

Whorf was referring to universal phonetic symbolism, the sublinguistic universal receiving attention in his day. Certain nonsense sound patterns, e.g., "queep," seemed to elicit a universal set of associations, e.g., sharp (versus dull), quick (versus slow), hard (versus soft), bright (versus dark), etc., regardless of one's language or culture.

Whorf did not deny that phonetic symbolism might be a sublinguistic universal, nor did he feel he had to choose between the implied nativism of the phenomena and his own linguistic relativism. He did not have to choose because, from Whorf's perspective, the relationship between biological universals (what he referred to as "lower-psyche facts") and language and culture was not one of contrast but rather one of "appropriation." He points out that "There is a yogic mastery in the power of language to remain independent of lower-psyche facts, to override them, now point them up, now toss them out of the picture . . . " (Whorf 1956:267). In the case of phonetic symbolism, this means that any language is free

> . . . to mold the nuances of words to its own rules, whether the psychic ring of the sound fits or not. If the sounds fit, the psychic ring of the sounds is increased [and is noticed, as in poetry]. . . . If the sounds do not fit, the psychic quality changes to accord with the linguistic meaning . . . [Whorf 1956:267].

As an example, Whorf pointed to the sound pattern "deep." Phonetically, "deep" is akin to "queep"; speakers of languages other than English associate "deep" with sharp, quick,

bright, hard, etc. But "deep" is not meaning free to anyone reading this essay; its meaning overrides our sublinguistic affective sensitivity to its phonetic pattern. As Whorf remarks, "[the] linguistic meaning [of 'deep'] in the English language happens to refer to the wrong sort of experience for such an association. This fact completely overrides its objective sound, causing it to 'sound' subjectively quite as dark, warm, heavy, soft, etc. as though its sounds really were of that type" (268).

Once Whorf's hierarchical view is acknowledged (see 1956:267-268; see also 1956:239), it becomes possible to interpret more properly those paragraphs in his writings, which are so widely quoted yet so frequently misunderstood, in which he introduces his

> principle of relativity, which holds that all observers of the universe are not led by the same physical evidence to the same *picture* of the universe unless their linguistic backgrounds are similar, or can in some way be calibrated. . . . The categories and types we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary *the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds*—and this means largely by the linguistic systems in our minds. We cut up nature, *organize it into concepts,* and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community and is codified in the patterns of our *language* [1956:214-215; emphases added].

What Whorf says in these remarks is that the world *presents* itself in a "kaleidoscopic flux of impressions." Nowhere does he suggest that the world is actually *perceived* as a continuum; he acknowledges that the world is perceived categorically and by reference to types. Whorf's point is a *logical* one; from the point of view of the way the world *presents* itself, all things are equally alike and equally different, that is, the number of true things one can say about any two things (the number of predicates that apply) are equal, and perhaps infinite. This notion of logical (or meta-ontological) equidistance has been more recently discussed by Goodman (1968, 1972) and Watanabe (1969:376-388; also see Shweder, Bourne, and Miyamoto 1978). Watanabe provides a formal proof that "insofar as the number of shared predicates is regarded as a measure of similarity and the number of predicates that are not shared is regarded as a measure of dissimilarity . . . there exists no such thing as a class of similar objects in the world," that is, "the world would look 'gray,' amorphous, and meaningless. . . ." Or, as Goodman (1968:32) remarks, "to admit all classifications on equal footing amounts to making no classification at all."

Echoes of Whorf can be found throughout Goodman's (1968, 1972) discussions of "representation." Notice that Whorf's primary concern was with the relationship between our "picture" of the world and the world as it presents itself. What he suggested throughout his writings were the principles that: (a) there is no such thing as a single copy of the way the world actually is; and (b) the relationship between a representation (e.g., a "picture") and the thing represented (e.g., the "universe") is not one of resemblance. Both principles have been recently defended by Goodman (1968:6), who remarks:

> 'To make a faithful picture, come as close as possible to copying the object as it is.' This simple-minded injunction baffles me; for the object before me is a man, a swarm of atoms, a complex of cells, a fiddler, a friend, a fool and much more. If none of these constitute the object just as it is, then none is *the* way the object is. I cannot copy all these at once; and the more nearly I succeeded, the less would the result be a realistic picture.

To those principles Whorf adds a third: that the way we group things into categories has its most important source in our particular speech community or culture.

In the passages quoted earlier, Whorf refers to our "linguistic" background (or "language"), on the one hand, and the way we categorize or organize nature into "concepts," on the other hand. He does not suggest that the relationship between language and categorization is to be thought of as the relationship between an independent and dependent

variable (although that is the way this passage is often interpreted). Quite the contrary, there is good reason to believe that Whorf viewed the relationship between a people's language and their category system as the relationship of a whole to a part. For example:

> . . . language first of all is a classification and arrangement of the stream of sensory experience which results in a certain world-order. . . . [and] . . . every language is a vast pattern-system, different from others, *in which* are culturally ordained the forms and categories by which the personality not only communicates, but also analyzes nature, notices or neglects types of relationships and phenomena . . . [Whorf 1956:55, 252; emphasis added].

In other words, an adequate description of a language must include an account of the way a people classifies its world; ontology is a cultural inheritance reflected in the way members of a speech community talk to one another.

In short, then, Whorf did not deny the existence of sublinguistic universals (although he thought language "free" to ignore them, override them, or make use of them), nor did he claim that the world was perceived as chaotic (although, on logical grounds, he believed it must be considered as presented in that way), nor did he try to define language independently of a culture's category system. It is in the light of those three points that the true ironies of the color-research tradition can be appreciated.

In the early era of research on language and thought, color was emphasized because of its supposed freedom from intrinsic organization (it was seen as a continuum). Even if that were true (and we have a long line of evidence indicating that it is not), it would be an unnecessary restriction since, from Whorf's perspective, any sublinguistic structure that exists could be used by language as it will. In the latter era, the perceptual irregularities in the color spectrum were seized upon to demonstrate the influence of sublinguistic universals on language. That position was seen as a critique of Whorf's assertions about a "kaleidoscopic" universe, which we now see were misinterpreted. Whorf's image of a "kaleidoscopic flux" is based upon a principle of *logical* equidistance; it is not to be analogized to a psychophysical continuum. For Whorf (and Goodman 1968 and Watanabe 1969 as well) two Munsell color chips are no more like each other than focal red is like the Eiffel Tower; that is, from a logical point of view, the number of true things that can be said about each pair (the number of attributes shared) is equal and infinite.

A second irony of the color-research tradition is that what Whorfian wisdom there was in the early era was subsequently ignored. The communication-accuracy measure introduced by Lantz and Stefflre, for example, was based on a view of language broad enough to include a culture's entire system of color categories, not simply their *basic* color *terms*. If one remains committed to this Whorfian conception of language, it becomes possible to draw two conclusions about the relationship among language, thought, and external stimuli in the color domain. First, a linguistic measure (communication accuracy) is predictive of recognition memory for color regardless of array, and, as we have shown, this applies to arrays containing focals as well as to those studied by Lantz and Stefflre. Second, the claim that the sublinguistic universal "focality" exercises a causal influence over language is quite tenuous. On the one hand, most color categories (e.g., blonde, sky blue, lavender, etc.) have no reference to focal areas at all; focality is not a necessary condition for forming a color category. On the other hand, focality is not a sufficient condition for the development even of basic color terms; cultures that have had equivalent amounts of time to evolve a color lexicon vary enormously in the number of basic color terms (not to be confused with color categories) in their linguistic repertoire. Berlin and Kay (1969) have made a significant contribution to our understanding of the evolution of basic color terms; nonetheless, their work was not intended to illuminate (and it does not illuminate) the way people make use of all their linguistic resources to encode color stimuli in practical cognitive tasks such as recognition memory.

What we discover in our various experiments on language, focality, and memory is

consistent with Whorf's hierarchical view of the relationship between language (including cultural category systems) and sublinguistic universals. While focality may occasionally emerge as an important variable in the organization of behavior, the availability of a referentially precise basic color description (e.g., "flesh," "pale yellow") and not focality is the decisive element in recognition memory. One is certainly free to appropriate linguistically the perceptually salient properties of focal chips (e.g., "the reddest red"), and there is a tendency for focal chips to be easier to describe than nonfocal chips (see Figure 2). Nonetheless, there are many influences other than focality on the ease with which a referentially precise description can be formulated, and the influence of focality on recognition memory is mediated almost entirely by basic color descriptions, a linguistic factor.

In summary, language appears to be a probable vehicle for human color memory, and the views developed by Whorf are not jeopardized by the findings of any color research to date.

NOTES

Acknowledgments. The research reported in this study was made possible by a grant from the Division of Social Sciences of the University of Chicago. We express our thanks to Michele Bograd, Mark Busse, and Suzanne Gaskins for assistance on several aspects of the project. Paul Kay's comments on an earlier version of this manuscript were much appreciated.

[1] Cross-cultural research on both color perception and color lexicons extends, of course, back to the turn of the century (e.g., Woodworth 1910; Rivers 1901) although that work has not been very influential in contemporary research. A variety of more linguistically oriented studies conducted during the 1950s considered the variability in color lexicons without directly assessing any cognitive variable (such as memory) (e.g., Conklin 1955) or merely assessed referential range (e.g., Ray 1953).

[2] "Basic" color terms are a subset of the expressions for denoting color. According to Berlin and Kay (1969:6), the subset is identified by a number of criteria. Basic color terms must be monolexemic (blue is basic; sky blue is not), must not be included in the signification of any other color term (red is basic; crimson, a "kind of" red, is not), must not be restricted in their application to a narrow class of objects (yellow is basic; blonde is not), and must not be the name of an object having that color (yellow is basic; gold is not).

[3] Several considerations favored this method of equalization. This array retains the majority of the chips in Heider's array, thus facilitating comparison; it randomizes the possible effects of immediate neighbors; and it shows behavioral evidence of successful equalization. Alternative methods of equalization include numerical weighting and construction of a new array in which chips are perceptually equidistant. Numerical weighting, as conducted by Heider, failed to prove effective since significant perceptual differences were found in Experiment I. Further, in weighting to compensate for accidental influences because of neighboring chips, real effects might have been obscured. Construction of a completely new array is not only very tedious but would have eliminated most of the focals from the array since they are extremes and cannot be made perceptually equipotent. This would also have destroyed comparability with Heider's array. On balance, then, the procedure of controlled modification of the existing array with direct behavioral assessment of focal/nonfocal equity seemed the best approach.

[4] A rough preliminary count on our part indicated that four of the eight basic color terms of English (blue, green, purple, pink) were applicable to 75% of the chips in the array whereas the remaining four terms (brown, yellow, red, orange) were applicable to only 25%. This inequity might not be a problem if it could be shown that it represented the actual distribution of colors in the natural world or if the distribution of the probes were proportioned in a similar manner. We cannot ascertain the actual distribution of colors, but we can examine the applicability of the eight terms to the probes; here, we find that the nonfocal probes, with their 81/19 division, approximate the 75/25 division in the array, but that the focals, with their 50/50 division, do not. This means that

the focal set contains a disproportionate number of easy-to-identify chips if the subjects used basic color terms in memory, with their attendant potentials for referential confusability. For example, there are five nonfocal probes that might be termed "green," in contrast to one focal probe; "green" is one of those terms with many alternates in the array; thus, a subject's nonfocal accuracy score is likely to be lower simply because he must deal more than twice as often with the large number of alternative "greens" in the nonfocal set of probes.

[5] This restriction was not imposed on the first few subjects. When it became clear that the overwhelming majority of the descriptions were three words or fewer, the three-word limit was then imposed on subsequent subjects to facilitate scoring.

[6] This scoring of embedded modifications makes sense in light of the alternatives. For "light royal blue," giving one point to *light* and one to *royal* effectively equates the occurrence of *light* in this case with that in "light blue," which is misleading. Splitting the one-point total into one-half for *light* and one-half for *royal* is no better for it demotes this *royal* relative to the *royal* of simple "royal blue," and there are no grounds for that since the embedding implies the primacy of *royal blue* as the basic description. Considering *light-royal* as a compound with a joint value of one point also obscures the syntactic primacy of *royal* and, as a practical matter, dramatically eliminates modifier matches of any sort in the 24 targets. It is worth noting that this procedure tended to eliminate mostly such all-purpose modifiers as *light, dark,* and *medium.* (Very frequently, different subjects would use the opposite terms on the same color!) Other modifiers eliminated from descriptions were of much lower frequency but were of the same type: *pale, bright, very, deep, almost, sort of.*

[7] In our examination of linguistic indicators, we were concerned with individual chips; we wanted to know whether the most communicable chips were also the most memorable. Thus, for the sake of comparability, the focality-memory correlation coefficients were also calculated "by chip." The most generous interpretation we can give to the relationship between focality and memory is to convert the "by-subject" t-test results in Experiment III into their associated correlation coefficients by the method discussed earlier. By doing that, we lose comparability with the linguistic indicators, but it is worth presenting the results nonetheless. r(focality − STM) = .30 (n.s.) (controlling for two-person communication accuracy, it reduces to .06). r(focality − LTM) = .48 ($p = < .05$) (controlling for two-person communication accuracy, it reduces to .32). Even under these generous conditions, the relationship between the linguistic indicators and memory is more substantial. r(two-person communication accuracy − STM) = .58 ($p = < .001$) (controlling for focality, it reduces to .52. r(two-person communication accuracy − LTM) = .54 ($p = < .01$) (controlling for focality, it reduces to .41).

REFERENCES CITED

Berlin, Brent, and Paul Kay
1969 Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press.

Bransford, John D., and Nancy S. McCarrell
1974 A Sketch of a Cognitive Approach to Comprehension: Some Thoughts About What It Means To Comprehend. *In* Cognition and the Symbolic Processes. W. B. Weiner and D.S. Palermo, eds. pp. 189-229. New York: Wiley.

Brown, Roger
1965 Social Psychology. New York: Free Press.
1976 In Memorial Tribute to Eric Lenneberg. Cognition 4:125-153.
1977 In Reply to Peter Schönbach. Cognition 5:185-187.

Brown, Roger, and Eric H. Lenneberg
1954 A Study in Language and Cognition. Journal of Abnormal and Social Psychology 49: 454-462.

Burnham, Roger W., and Joyce R. Clark
1955 A Test of Hue Memory. Journal of Applied Psychology 39:164-172.

Conklin, Harold
1955 Hanunoo Color Categories. Southwest Journal of Anthropology 11: 339-344.

Glanzer, Murray, and William H. Clark
1963a Accuracy of Perceptual Recall: An Analysis of Organization. Journal of Verbal Learning and Verbal Behavior 1:289-299.
1963b The Verbal Loop Hypothesis: Binary Numbers. Journal of Verbal Learning and Verbal Behavior 2:301-309.
Goffman, Erving
1976 Replies and Responses. Language in Society 5:257-313.
Goodman, Nelson
1968 Languages of Art. New York: Bobbs-Merrill.
1972 Seven Strictures on Similarity. *In* Problems and Projects. N. Goodman, ed. pp. 437-450. New York: Bobbs-Merrill.

Heider, Eleanor Rosch
1972 Universals in Color Naming and Memory. Journal of Experimental Psychology 93:10-20.

Lantz, DeLee, and Volney Stefflre
1964 Language and Cognition Revisited. Journal of Abnormal and Social Psychology 69:472-481.
Lenneberg, Eric H.
1961 Color Naming, Color Recognition, Color Discrimination: A Reappraisal. Perceptual and Motor Skills 12:375-382.
1971 Language and Cognition. *In* Semantics. D. Steinberg and L. Jakobovits, eds. pp. 536-557. Cambridge: Cambridge University Press.
Lenneberg, Eric H., and John M. Roberts
1956 The Language of Experience: A Study in Methodology. International Journal of American Linguistics Memoir 13 (Supplement to Vol. 22, [2]): 1-33.
Lienhardt, Godfrey
1961 Divinity and Experience: The Religion of the Dinka. Oxford: Oxford University Press.

Newhall, Sidney M., Dorothy Nickerson, and Deane B. Judd
1943 Final Report of the O.S.A. Sub-Committee on the Spacing of the Munsell Colors. Journal of the Optical Society of America 33:385-418.

Ray, Verne F.
1953 Human Color Perception and Behavioral Responses. Transactions of the New York Academy of Sciences 16:98-104.
Rivers, W. H. R.
1901 Primitive Color Vision. Popular Science Monthly 59:44-58.

Searle, John R.
1969 Speech Acts. Cambridge: Cambridge University Press.
Shweder, R. A., E. J. Bourne, and J. M. Miyamoto
1978 Concrete Thinking and Category Formation: A Cultural Theory With Implications for Developmentalists. Unpublished manuscript. University of Chicago.
Stefflre, Volney, Victor Castillo Vales, and Linda Morley
1966 Language and Cognition in Yucatan: A Cross-Cultural Replication. Journal of Personality and Social Psychology 4:112-115.

Watanabe, Satosi
1969 Knowing and Guessing. New York: Wiley.
Welkowitz, Joan, Robert B. Ewen, and Jacob Cohen
1971 Introductory Statistics for the Behavioral Sciences. New York: Academic Press.
Whorf, Benjamin Lee
1956 Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf. J. Carroll, ed. Cambridge, Mass.: M.I.T. Press. (1st eds. 1936, 1940, 1941, 1942).
Woodworth, R. S.
1910 The Puzzle of Color Vocabularies. Psychological Bulletin 7:325-334.

Ziff, Paul
1972 Understanding Understanding. Ithaca, N.Y.: Cornell University Press.

Submitted 31 May 1978
Revised version submitted 22 September 1978
Accepted 19 October 1978
Final revision received 5 November 1978

APPENDIX 1

Discussion of Heider's Array and Probes

In her codability experiment, Heider used Berlin's and Kay's array, which contained the maximum saturation level for each hue and value combination. In the memory experiment, she modified their array since subjects found the memory task too difficult with 320 chips in the array. She eliminated every other hue (the 2.5 and 7.5 columns), making a 160-chip array (with just the 5.0 and 10.0 columns). This array is reproduced in the accompanying diagram. Scores in the grid represent the saturation level associated with each hue and value. Letters on the hue axis refer to approximate hues (e.g., RP = red-purple). The chips were mounted on the buff-colored mounting sheets provided with the chips by the Munsell Corporation. The probes were mounted on 5.1- by 7.6-cm. cards of a matching color. This differed from Heider's array and probes, which were mounted on white.

In eliminating every other hue, Heider also eliminated many of the targets she had used in the codability study, including four of the eight focals. She substituted chips that were similar from the 5.0 and 10.0 columns left in the array. In making those substitutions, Heider is not explicit about which chip is substituted for which, so we have simply marked the new set of probes in the array. (The subscripts denote as follows: F = focal, B = boundary [near a focal], I = internominal [between focals].) There are several other problems with replicating her list of probes. In one case, she lists a nonexistent chip (10GY 8/10), for which a plausible substitution was made (10GY 8/8). For another, she simply indicates "maximum saturation" (5YR 7/max), for which the maximum value available at the time of this research was substituted (5YR 7/14). Finally, Heider lists more substitutions than there are 2.5 and 7.5 values in the original list of boundary chip probes. Therefore, we included an extra boundary chip, making the total number of probes 25. One of these was eliminated in Experiment II (10P 7/6), returning the total to 24 for all experiments subsequent to Experiment I.

HUE

VALUE	10RP	5RP	10P	5P	10PB	5PB	10B	5B	10BG	5BG
2	8	8	6	8	10	8	6	6	6	6
3	10	10_I	10	10_F	10	10	6_B	8	8	8
4	14	12	12	12	12	12	10_F	10	6_I	8
5	14	12	12	10	10	12	12	10	10	10
6	12	12	10	8_B	10	10	10	8_B	8	10
7	8	10	6_B	6_I	8	8	8	8	6_I	8
8	6	6	6	4	4	6	6	4	4	4
9	2	2	2	2	2	2	2	2	2	2

Appendix 1 continued

HUE

VALUE	10G	5G	10GY	5GY	10Y	5Y	10YR	5YR	10R	5R	
	6	6	4	2	2	2	2	4	6	8_B	2
	4_B	8	6	6	4	4_B	6	6_F	10	10	3
	10	10	8	8	6_I	6	8	8	12	14_F	4
	10	10	12_F	10	8	8	10	12	16	14	5
	10	10	12	10	10_I	10	12	14	14	12	6
	8	10	8_B	12	12	12	14	14_F	10	10	7
	6	6	8_I	10	12	14	14_F	6_I	6	6_F	8
	2	2	4	4	6_B	6	4	2	2	2	9

APPENDIX 2

Summary data for Experiment I. Perceptual search (discriminability) scores reported by chip and by subject.

By subject (percentage correct and number correct out of 8 focals and 16 nonfocals with visible probe)		By chip (number of subjects out of 18 who correctly located the chip in the array with visible probe)	
Subject number		Chip number	
1	0.75 (6)	0.31 (5)	1	13
2	0.62 (5)	0.75 (12)	2	12
3	0.75 (6)	0.75 (12)	3	17
4	0.75 (6)	0.43 (7)	4	7
5	0.87 (7)	0.69 (11)	5	10
6	0.62 (5)	0.43 (7)	6	12
7	0.37 (3)	0.31 (5)	7	12
8	0.62 (5)	0.62 (10)	8	15
9	0.75 (6)	0.69 (11)	9	5
10	0.87 (7)	0.87 (14)	10	14
11	1.00 (8)	0.69 (11)	11	13
12	0.87 (7)	0.87 (14)	12	12
13	0.75 (6)	0.69 (11)	13	8
14	1.00 (8)	0.69 (11)	14	14
15	0.75 (6)	0.56 (9)	15	14
16	0.87 (7)	0.75 (12)	16	8
17	0.87 (7)	0.69 (11)	17	17
18	1.00 (8)	0.81 (13)	18	18
$\bar{X}$	= 0.78	0.64	19	14
	$t_{17} = 3.90$ ($p = < .001$)		20	13
			21	17
			22	10
			23	13
			24	11

APPENDIX 3

Array developed in Experiment II by removing selected chips and randomizing (location of probes marked by subscripts).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	5G 4/10	5Y $3/4_B$	5P 5/10	5R $8/6_F$	5RP 6/12	10YR 2/2	5PB 8/6	10BG 9/2	10PB 4/12	5R 6/12	10G 4/10	10YR 5/10	10P 5/12	10G $3/4_B$	10B 8/6	1
2	10BG $4/6_I$	10R 6/14	10Y $9/6_B$	5R 5/14	10P 9/2	10B $4/10_F$	5B 9/2	10BG 5/10	5PB 6/10	5P 4/12	5RP 7/10	10B 7/8	5BG 2/6	10P 4/12	5BG 3/8	2
3	10G 5/10	5G 8/6	10R 5/16	10G 6/10	10G 9/2	5RP 4/12	5P 2/8	5P $6/8_B$	10B $3/6_B$	10G 7/8	10GY $7/8_B$	5R $4/14_F$	5BG 9/2	5BG 5/10	10RP 4/14	3
4	10R 4/12	5YR $3/6_F$	10P 6/10	10GY $5/12_F$	10BG 8/4	5Y 7/12	5P $3/10_F$	5RP 2/8	10BG 2/6	10YR 7/14	10YR 4/8	5GY 8/10	10YR $8/14_F$	10Y $6/10_I$	5GY 7/12	4
5	10PB 5/10	10P 2/6	10B 9/2	5R 7/10	5Y 5/8	5RP $3/10_I$	5PB 9/2	5B $6/8_B$	10Y 2/2	5YR 6/14	5BG 8/4	10YR 9/4	5R 3/10	5G 6/10	10B 2/6	5
6	5PB 2/8	10RP 6/12	10PB 9/2	5BG 7/8	5P 9/2	10Y $4/6_I$	10YR 6/12	5Y 6/10	5P $7/6_I$	10RP 7/8	5GY 5/10	5YR $8/6_I$	5RP 5/12	5YR 5/12	10R 3/10	6
7	5G 9/2	5B 2/6	5R $2/8_B$	10R 7/10	10G 8/6	5G 5/10	5YR $7/14_F$	5PB 5/12	10R 2/6	5Y 8/14	10Y 8/12	10BG $7/6_I$	10GY 4/8	10GY $8/8_I$	5YR 2/4	7
8	5Y 2/2	10PB 8/4	5B 8/4	5GY 2/2	10RP 5/14	5RP 9/2	10R 9/2	5P 8/4	5YR 9/2	10P 8/6	10RP 9/2	5BG 6/10	5R 9/2	5GY 9/4	10GY 9/4	8
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

APPENDIX 4

Summary data for Experiment III. Recognition memory scores reported by chip and by subject.

		By chips		By subjects (percentage correct)				
					Short-term memory		Long-term memory	
		Short-term memory (number correct)	Long-term memory (number correct)		Non-focals	Focals	Non-focals	Focals
1.	10Y 6/10	6	2	1 (F)	42	50	25	50
2.	10Y 4/6	13	3	2 (F)	75	100	50	100
3.	10GY 8/8	10	4	3 (M)	75	67	50	50
4.	10BG 7/6	4	2	4 (F)	67	50	25	50
5.	10BG 4/6	6	0	5 (M)	33	67	0	50
6.	5P 7/6	12	3	6 (M)	25	100	50	100
7.	5RP 3/10	10	0	7 (M)	42	17	50	100
8.	5YR 8/6	15	6	8 (F)	83	67	25	50
				9 (M)	67	83	75	100
9.	10B 3/6	9	4	10 (F)	58	83	25	50
10.	5P 6/8	11	1	11 (F)	42	67	50	50
11.	5R 2/8	8	1	12 (F)	42	67	25	50
12.	5B 6/8	13	1	13 (M)	33	83	25	50
13.	5Y 3/4	10	5	14 (F)	58	67	75	50
14.	10Y 9/6	16	5	15 (M)	50	0	50	100
15.	10GY 7/8	4	2	16 (M)	92	83	50	50
16.	10G 3/4	8	1	17 (F)	58	50	50	50
				18 (F)	67	67	25	50
17.	10YR 8/14	15	4	19 (M)	33	67	50	100
18.	5YR 7/14	12	4	20 (M)	33	0	25	0
19.	10GY 5/12	15	4	21 (F)	42	67	100	50
20.	10B 4/10	12	4	22 (M)	58	83	50	100
21.	5R 4/14	13	4	23 (F)	83	67	0	0
22.	5R 8/6	5	2	24 (M)	33	50	50	0
23.	5YR 3/6	10	3		$\bar{X} = 53.8$	$\bar{X} = 62.6$	$\bar{X} = 41.7$	$\bar{X} = 58.3$
24.	5P 3/10	8	3					

APPENDIX 5

Basic descriptions developed in Experiment V, with their basic terms and their modifiers.

Chip	I.D. number	Basic description	Basic term	Modifiers	Points[a]
1	10Y 6/10	olive green	green	olive	4.5
		yellowish green		yellowish	4.5
		avocado green		avocado	4.0
				lime	3.0
				chartreuse	2.0
				pea	2.0
				misc.	9.0[b]
2	10Y 4/6	olive green	green	olive	10.5
				avocado	3.0
				military/army	2.0
				dark	2.0
				misc.	11.5
3	10GY 8/8	light green	green	light	6.5
		pastel green		pastel	4.0
				yellowish	2.0
				whitish	2.0
				lime	2.0
				mint	1.5
				misc.	11.0
4	10BG 7/6	light blue	blue	light	7.5
		turquoise		sky	3.5
				aqua	2.5
				baby	2.0
				misc.	9.5
			turquoise	unmodified	4.0
5	10BG 4/6	blue green	green	blue	9.0
		greenish blue		misc.	1.0
			blue	greenish	4.0
				dark	3.0
				misc.	6.0
			turquoise	unmodified	2.5
			aqua	unmodified	3.5
6	5P 7/6	lavender	lavender	unmodified	7.5
		violet		misc.	1.0
			violet	unmodified	7.0
			purple	whitish	3.5
				light	3.0
				misc.	4.0
			lilac	unmodified	3.0

[a] There are 29 total points for each chip, one for each subject. Half-points resulted when a subject gave two independent modifiers to describe one chip.

[b] The miscellaneous category includes all modifiers or basic terms for a given chip that received only one point.

Chip	I.D. number	Basic description	Basic term	Modifiers	Points[a]
7	5RP 3/10	purple	purple	unmodified	5.0
		reddish purple		reddish	4.0
		violet		dark	2.0
				misc.	2.5
			violet	unmodified	4.0
			magenta	unmodified	2.5
			burgundy	unmodified	2.0
			maroon	misc.	4.0
			misc.		1.0
8	5YR 8/6	flesh	flesh	unmodified	14.0
		peach		misc.	1.0
			peach	unmodified	5.5
			salmon	unmodified	2.5
			beige	unmodified	2.0
			misc.		4.0
9	10B 3/6	dark blue	blue	dark	8.0
				deep	3.0
				unmodified	2.0
				royal	2.0
				gray	2.0
				misc.	10.0
			gray	blue	2.0
10	5P 6/8	lavender	lavender	unmodified	5.5
		violet		light	3.0
				misc.	1.0
			violet	unmodified	4.0
				misc.	.5
			purple	medium	3.0
				whitish	2.5
				light	2.5
				misc.	5.0
			lilac	dark	2.0
11	5R 2/8	maroon	maroon	unmodified	12.0
				dark	2.0
				misc.	1.0
			burgundy/wine	unmodified	3.0
				dark	1.5
			purple	misc.	2.5
			red	misc.	2.0
			misc.		4.0
12	5B 6/8	sky blue	blue	sky	10.0
		light blue		light	4.5
				sea	2.0
				bright	2.0
				robin egg	1.5
				misc.	6.0
			turquoise	misc.	2.0
			misc.		1.0

Chip	I.D. number	Basic description	Basic term	Modifiers	Points[a]
13	5Y 3/4	olive green	green	olive	8.0
		brownish green		brownish	4.5
				moss	2.0
				misc.	7.5
			brown	greenish	2.0
				misc.	2.0
			misc.		3.0
14	10Y 9/6	light yellow	yellow	light	6.5
		pale yellow		pale	5.0
				greenish	1.5
				misc.	9.0
			green	yellowish	3.0
				misc.	1.0
			white	yellowish	3.0
			misc.		1.0
15	10GY 7/8	light green	green	light	7.0
				yellowish	2.5
				apple	2.0
				pea	2.0
				whitish	1.5
				misc.	13.0
			misc.		1.0
16	10G 3/4	dark green	green	dark	9.0
				evergreen/pine	3.0
				forest	2.0
				sea	2.0
				unmodified	2.0
				grayish	1.5
				misc.	7.5
			misc.		2.0
17	10YR 8/14	yellow	yellow	unmodified	5.0
				bright	3.0
				orangish	2.5
				sunshine	2.0
				rich-looking	2.0
				misc.	8.5
			orange	yellowish	2.0
			gold	misc.	2.0
			misc.		2.0
18	5YR 7/14	orange	orange	unmodified	7.0
		yellowish orange		yellowish	6.0
				bright	3.0
				light	2.0
				orange	2.0
				fluorescent	2.0
				misc.	3.0
			apricot	misc.	2.0
			misc.		2.0

Chip	I.D. number	Basic description	Basic term	Modifiers	Points[a]
19	10GY 5/12	green	green	bright	6.0
		bright green		unmodified	5.5
				grass	3.0
				kelly	2.0
				yellow	1.5
				misc.	11.0
20	10B 4/10	royal blue	blue	royal	4.5
		medium blue		medium	4.5
				unmodified	2.5
				light	2.0
				sky	2.0
				bright	2.0
				rich	1.5
				misc.	9.0
21	5R 4/14	red	red	unmodified	8.0
				bright	2.5
				fire engine	2.0
				pinkish	2.0
				misc.	8.5
			magenta	unmodified	3.0
			misc.		3.0
22	5R 8/6	pink	pink	unmodified	13.5
				bright	3.0
				cosmetic	2.0
				medium	2.0
				misc.	5.5
			flesh	misc.	2.0
			misc.		1.0
23	5YR 3/6	brown	brown	unmodified	11.5
				yellowish	3.0
				shit	2.5
				medium	2.5
				dark	2.0
				misc.	6.5
			misc.		1.0
24	5P 3/10	purple	purple	unmodified	13.0
		dark purple		dark	4.0
				misc.	6.0
			violet	misc.	2.0
			misc.		4.0

APPENDIX 6

Summary data for Experiments IV and V. Two-person communication, group communication, and referential confusability.

		Experiment IV		Experiment V		
		Number of 1st hits	$\bar{X}$ number of tries	Number of targets chosen	Number of best examples	Number of reasonable choices
Chip	I.D. number					
1	10Y 6/10	6	2.5	3	8	2.8
2	10Y 4/6	13	1.5	7	3	2.3
3	10GY 8/8	8	2.2	3	4	3.3
4	10BG 7/6	4	3.5	3	7	3.0
5	10BG 4/6	4	3.1	0	8	3.7
6	5P 7/6	4	2.7	2	6	4.2
7	5RP 3/10	11	2.0	2	5	3.0
8	5YR 8/6	13	1.7	6	3	1.8
9	10B 3/6	6	2.3	0	2	3.5
10	5P 6/8	9	2.2	2	8	3.2
11	5R 2/8	7	2.3	2	5	2.2
12	5B 6/8	3	3.4	0	5	3.5
13	5Y 3/4	9	2.2	4	3	2.8
14	10Y 9/6	15	1.4	9	2	1.2
15	10GY 7/8	6	2.6	0	6	4.2
16	10G 3/4	4	2.8	0	4	2.5
17	10YR 8/14	9	1.9	1	3	2.0
18	5YR 7/14	12	1.6	8	3	2.3
19	10GY 5/12	12	2.0	7	4	2.4
20	10B 4/10	13	1.7	7	4	2.0
21	5R 4/14	17	1.3	7	4	2.6
22	5R 8/6	11	1.7	1	5	4.0
23	5YR 3/6	11	1.8	6	3	2.5
24	5P 3/10	5	2.2	2	3	2.2