MBS TECHNICAL REPORT 16-02

A Numerical Approach for Defining Basic Color Terms and Color Category Best Exemplars

Nicole Fider[1], Louis Narens[2,3], Kimberly A. Jameson[3], Natalia L. Komarova[1,3,*]

[1] Department of Mathematics, University of California Irvine, Irvine CA 92697-3875.
[2] Department of Cognitive Sciences, University of California Irvine, Irvine CA 92697-5100.
[3] Institute for Mathematical and Behavioral Sciences, University of California Irvine, Irvine CA 92697-5100.
[*] To whom correspondence should be addressed. E-mail: komarova@uci.edu

Abstract

A new method is presented for the identification of basic color terms (BCTs) from a language's color-naming data. A function is defined to measure how well a term is understood in a communicating population. A threshold value is then constructed for this function to separate basic terms from other terms. Also, a new mathematical method is proposed and analyzed for determining the best exemplar associated with a basic color term. Using data from 110 societies provided by the World Color Survey, comparisons are made between our methods and other known studies. The comparisons suggest a new definition of "basicness" which mostly agrees with the typical definition due to Kay and colleagues (B. Berlin, T. Regier, and other colleagues) found in the color categorization literature, but which has the advantage of not relying on syntactic or semantic features of languages or color lexicons, and therefore permits a methodology that is generalizable to other category domains for which a construct of "basicness" plays a role.

Cognitive psychology theory suggests that humans conceptualize worldly objects using relational frameworks that resemble hierarchical models of classification. These intuitive and compelling hierarchies provide nested ways of classifying and organizing things in the world according to varying levels of specificity. Examples of different relational hierarchy levels are: *superordinate*

categories, which are a general way of classifying items in a domain; *basic* categories, which are more specific in the sense that they have more features than superordinates; *subordinate* categories, which have even more specific features [1]. Such hierarchical classification schemes, also called *natural kind categorization systems*, serve as models of human conceptual organization and categorical similarity across many domains, and for category structures found both across individuals and across human societies. Examples include facial expression of emotion [2], kinship relations [3], perceptual color appearance [4]; other examples include more common categories that can vary across societies, such as birds [5], furniture [6], and plants [7]. The belief is that such natural kind classification systems differ from purely conventional classification systems (e.g., classification conventions such as partitioning distances using inches, feet and miles; or partitioning the world into different timezones) in that they draw conceptual boundaries using principles of classification that divide stimuli according to their real resemblances and real differences, and correspond to real distinctions in the nature of real-world things and in their cognitive processing [8].

One domain for which natural kind categorization has been extensively investigated, and which helped establish early natural categorization ideas, is exemplified by the hierarchical structures of color categorization systems that are represented in the world's languages [4, 9, 10, 11].

As a natural category domain, perceptual color appearance is universally prevalent. All humans with normal color vision can discriminate on the order of 10^6 different colors, which is many more colors than any individual can name reliably. Each of the approximately 7,000 living languages has some form of color lexicon, suggesting categorical color appearance is a widely held concept. Prior to global socialization most ethnolinguistic societies included individuals with only knowledge of the one color categorization system specific to their native language. Starting in the 1950's [12], considerable interest developed in comparing concept formation of native-language speakers from a range of cultures to determine whether individuals from different ethnolinguistic groups similarly conceptualized natural categorization domains, and to detail the features such categorization systems had in common [13]. Berlin and Kay [4] provided a seminal study for the large-scale investigation of color categorization across languages, followed by an extensive empirical survey known as the World Color Survey (WCS) [11, 14]. The now extensive literature on the topic suggests, among other things, that Basic Color Term (or BCT) categories are found across languages, supporting

an original conjecture of Berlin and Kay that BCTs are widely held cognitive constructs [4]. Subsequent investigations and a variety of theories have since been formulated to extend and account for these empirical findings [4, 15, 16, 17, 18, 19, 20, 11, 21, 22, 23, 24, 25, 26, 14].

Despite good empirical support across many language groups for the widespread existence of relational hierarchy structures that include a behavioral and sometimes linguistic superordinate "color" concept together with subordinate levels of "basic" and "nonbasic" color categories, subsequent research also suggests that a number of linguistic societies apply meaning to perceived color in ways that differ from the familiar associations learned by English language speakers [16, 27, 28, 29, 30, 18, 31]. These empirical results, and a reconsideration of the defining criteria, raise questions concerning the most objective way to define hierarchical levels such as BCTs and other nested color concepts such as *non-basic terms*, color category *best exemplars*, and others.

Berlin & Kay [4] fully developed the concept of BCTs as part of a highly influential theory for classifying color lexicon regularities in terms of color lexicon evolution. A BCT is a color term that at a minimum statisfies four criteria: It is (i) monolexemic, or monomorphic, (unlike *light blue* and *bluish*), (ii) with its meaning not included in the meaning of any other color term (unlike *crimson* which is a type of *red*), (iii) applicable to a wide class of objects (unlike *blonde* which applies only to beer, hair and wood), and (iv) it must be "psychologically salient", e.g., used by reliably native speakers (unlike *chartreuse*). The languages of modern industrial societies typically have thousands of color words, and a much smaller number of basic color terms. English has eleven BCTs: *red, yellow, green, blue, black, white, gray, orange, brown, pink, and purple.* Slavic languages have twelve, with separate basic terms for light blue and dark blue [32]. These ideas have been subsequently enriched by Kay and colleagues (e.g., [33, 34, 14, 17, 14, 35], and others), although the definition of BCTs remains largely unmodified from the original formulation detailed above. In sections below we refer to the above definition of Berlin, Kay and colleagues as the *B&K definition.*

Although the idea of BCTs has been shown to be empirically robust [20, 36, 14, 37, 38, 22, 24, 34], some aspects of the definition are problematic, and some are regularly violated by empirical data. For example, in Vietnamese color naming, color terms are commonly not monomorphic or monolexemic, in violation of (i) above [27, 28, 29]. And in English, *orange* is considered a basic color term even though it shares a name with an object (violating

B&K's subsidary criterion (vi), [4], p.6). Exceptions of this sort suggest that the current working definition of BCT could be improved.

Previous investigations have applied numerical methods to the study of BCTs and basic color regions. Most notably, Regier, Kay and colleagues have suggested that the extensions of color categories are constrained by the functional need for color naming systems to provide informative partitions of color space, and that category best-exemplars can be predicted using a model of representativeness [17, 11, 20, 36, 39, 35]. Brown and Lindsey (2009) used K-means clustering analysis to observe trends (or "motifs") in individual's color categorization patterns [22, 24]. Bimler (2007) defined boundary density maps for distinct languages for the purpose measuring similarities between languages based on their color-naming rules [37, 38]. And Chuang and Hanrahan (2009) confirm the B&K-definition by observing term deviation measurements of color terms of various languages [40]. Many of these empirical investigations show that in most linguistic societies, even when individuals vary in how they categorize color appearances, color terms are found to index specific color regions along a continuum of reliability. Thus some color words and their corresponding appearances have robust meaning while others have less clear semantic associations (both within and across individuals) which reflects a gradient of meaning in a population's color category system. As the above mentioned investigations attest, such variation creates difficulties for determining the exact denotative range of color terms, and for differentiating BCTs from other color terms that are present in a lexicon.

The goals of this paper is to provide a context for redefining the B&K "basicness" construct, and to propose numerical methods for identifying focal colors, by developing a quantitative, data-based methodology that objectively measures color term strength and identifies a language's BCTs from its nonbasic color terms. Our previous work on color categorization systems [41, 42, 43, 44, 45] in artificial agent populations aimed to formally capture the pragmatic communication features of color categories. Consistent with those studies we develop here an alternative view for BCTs emphasizing principles of categorization as a drive toward cognitive economy and communication combined with structure in the perceived world (cf. [46]).

Results

Basic Color Terms

Here we propose a quantitative algorithm to identify which color terms in a given language are basic. This algorithm relies on a notion of color term semantic strength. Therefore, in the rest of the paper we will refer to the new algorithm and the resulting color basicness definition as CS (for color strength). The main idea is as follows.

For a fixed language L we define a function Q_L, which assigns to each color term w a strength value in $[0, 1]$ based on how well the population of L uses the term. For a given color word w we know which colors, or colored stimuli, each person described with w. $Q_L(w)$ is calculated using this data (see SI for the details of the formula) such that any terms with low or inconsistent usage earn low strength values, while terms with high and consistent usage earn high strength values. We then choose a value t_* in $[0, 1]$ and postulate that a color term w should be considered to have enough strength to be deemed basic, or "CS-basic", with respect to t_* if $Q_L(w) \geq t_*$. We refer to t_* as the CS-basicness threshold value. In general, raising the threshold value t_* will cause fewer color terms to be considered CS-basic, while lowering t_* will cause more color terms to be considered CS-basic.

The strength function Q_L gives us an objective way to compare the color terms for language L, and imposing certain requirements on our threshold value allows us to find a range of reasonable values. Specifically, we require that our threshold values (1) admit only *modal* color terms of any language, (2) always admit at least two color terms per language, and (3) are relatively *stable*. These requirements produce values in the range $R = (0.168, 0.3343)$ as reasonable threshold candidates.

There are 110 languages represented in the WCS. In our analysis, two of the 110 languages (Mampruli and W.Taramuhara) are omitted because of irregularities in the data format. We use the data from the WCS's "color-naming" task to calculate the color strength of different color terms. Of the 2,244 color terms present across the 108 languages, 546 terms have strengths at or above 0.3343 and will be considered CS-basic regardless of the choice of t_* in the range R; 1578 terms have strengths at or below 0.168 and will be considered CS-nonbasic regardless of the choice of t_*. There are 123 terms whose classification will depend on the choice of t_*, so we classify these terms as CS-"potentially basic" for our discussion.

Table 1: Term classification: B&K vs CS

	CS-BCTs	CS-PBCTs	CS-NBCTs
B&K-BCTs	532	85	27
B&K-PBCTs	6	26	40
B&K-NBCTs	6	11	1511

Table 2: The number of WCS color terms classified as BCTs, "potentially basic" color terms (PBCTs), or nonbasic color terms (NBCTs) according to the compared B&K- and CS- definitions.

In analyses of WCS data [47], similar color term categorization results for each language were seen using the typically employed B&K-definition of basicness. Table 1 compares B&K and CS classifications of terms. The two definitions classify 2,068, or about 92%, of the terms in the same way. If we exclude all terms that are classified as non-basic by either of the definitions (which comprise more than 71% of all terms), more than 85% of the remaining terms are classified by the two methods in the same way. There are 46 terms which are B&K-"potentially basic" but are considered to be either CS-basic or CS-nonbasic. Similarly, there are 97 terms which are CS-"potentially basic" but considered to be either B&K-basic or B&K-nonbasic. The largest mismatch category is the terms that are classified as B&K-basic, but "potentially basic" by the CS measure. Many of these apparent mismatches can be eliminated by lowering the chosen threshold value, t_*. A particular choice of t_* is explored below and in the SI.

For a great majority of languages, the CS classification appears "monotonic" with respect to the B&K classification. That is, B&K-basic terms have a higher strength compared to B&K-nonbasic, see figure 1(a). There are, however, several exceptions. 33 terms are classified as basic with respect to one definition, but nonbasic with respect to the other; see table 1. These terms are distributed across 14 languages, displayed in figure 1(b). Terms that are B&K-basic but are CS-nonbasic are marked green, and terms that are CS-basic but B&K-nonbasic are marked blue. For two of the languages in this set (languages 53 and 70) experimenters in the field reported problems with data collection [47]. For three other languages (12, 45, and 92), terms of strength zero were classified as B&K-basic. In addition, for languages 3 and 106, there were pairs of terms such that the term with a lower strength

was classified as a B&K-basic, while the term with a higher strength as a B&K "potentially basic".

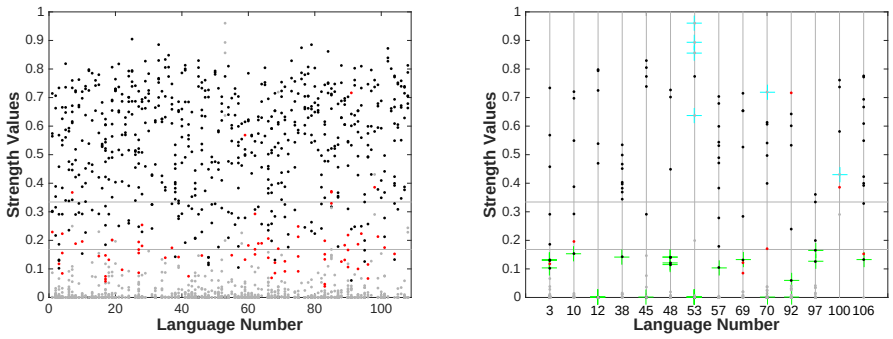

Figure 1: Visual representation of the strengths, B&K-basicness, and CS-basicness of color terms reported by all WCS languages. Points plotted in black correspond to B&K-basic color words, points plotted in red correspond to B&K-"potentially basic" color words, and points plotted in gray correspond to B&K-nonbasic words. The upper and lower bounds for the range of reasonable threshold values R are shown as gray horizontal lines. (a) All languages. (b) The 14 languages where the two definitions do not agree. Mismatches are marked by blue and green.

By fixing a specific basicness threshold value t_*, it is possible to eliminate classifying any terms as CS "potentially basic"; instead, each color term will either be CS-basic with respect to t_* or CS-nonbasic with respect to t_*. As an example we choose the threshold value $t_* = 0.17$; consequently, 665 of the 2,244 terms are identified as CS-basic. Figure 2 shows how many WCS languages were identified as having 2, ..., 10 basic color terms. The SI presents detailed analysis of BCTs for eight WCS languages studied in [40] and also presented by Kay et al. [33]. In sum, this analysis both supports the utility of our parameter-free CS-basicness modeling and largely confirms the results found by Kay et al. [33].

Best exemplars

In addition to quantitatively defining basicness, we also developed methods to objectively identify category best-exemplars, or "focal colors," based strictly on empirical data. Like BCTs, "focal color" is also a mainstream construct,

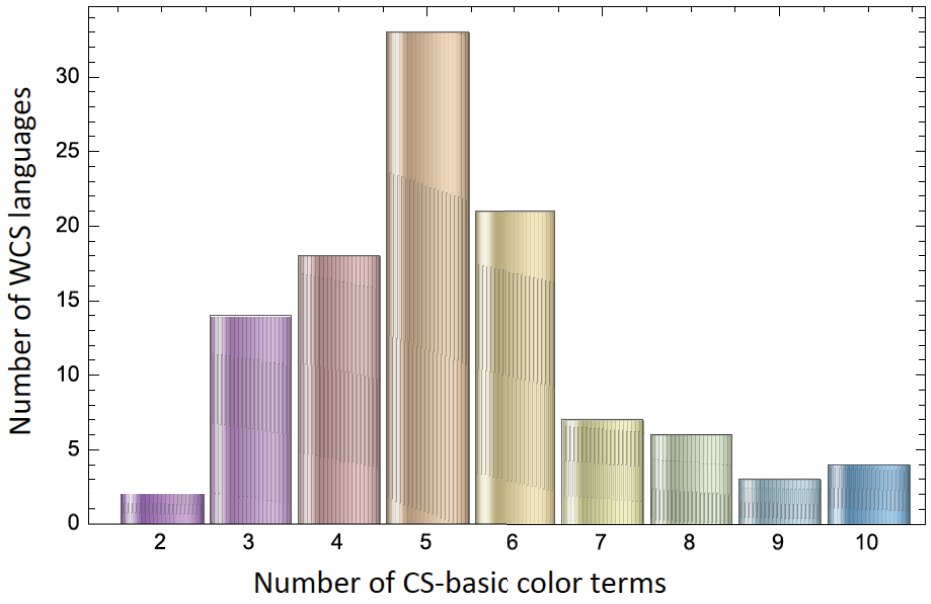

Figure 2: Application of the CS algorithm to WCS languages. Histogram depicts the distribution of WCS languages identified as having 2,. . ., 10 basic color terms. $t_* = 0.17$.

originally aimed at identifying stimuli with features most characteristic of a particular color term, concept, or category, based on data of a surveyed language group [12, 13, 4]. Our revised "best exemplar" definitions satisfy this original "focal colors" aim, using only information directly available in surveyed individual's data.

The WCS included a "focus (or focal) choice" task designed to identify surveyed individual's "focal color" stimuli for each color category, which is separate from the "color-naming" task mentioned above. It is clear that individual focal choice data can be used to identify each category's shared focal colors for a sample population assessed. To this end, for each individual using color word w, we compute the centroid of the set of "Focus Maps" (FM) given by all the individuals of the population. This centroid color, which we denote $F_{FM}(w)$, is the focus of w's category based on a focus-choice data. It seems obvious by construction that F_{FM} should give the best exemplar of w's category. Unfortunately, in the case of the WCS, two complications associated with the focus-choice task data constrain F_{FM} utility: (1) WCS focus-choice task data may contain some amount of empirical irregularity —

as suggested by experimenter reports describing variation in the ways the empirical task was presented to surveyed respondents — which most likely contributed to individual variation in the data depending on how participants may have interpreted the aim of the empirical task. And, (2) focus-choice data are not available for all terms elicited from participants because experimenters often decided to only collect focus-choice data for specific subset of elicited terms.

Thus, an issue here, and in the literature, is how to best identify focal colors when focal-choice data are unavailable or unreliable. For a solution, we turn to the "color-naming" data. Revisiting the mainstream "focal color" concept, it makes sense to assume that a color c is focal to the category corresponding to the word w if c is strongly associated with other members of w's category. Specifically, c should be strongly related to word w; as indexed by the rate at which c is called w, which should be larger than the rate at which any other color is called w. So we define $F_{HC}(w)$ (here the subscript stands for "High Consensus") to be the color stimulus/stimuli on which w attains the highest rate of use. Importantly, the construction of F_{HC} addresses the above mentioned issue by defining "focal colors" based strictly on individual's "color-naming" data (see SI for the precise mathematical formulations, and for a graphical comparison of the two best exemplar algorithms).

In general, it can be argued that F_{HC} is a strong exemplar candidate, with or without the availability of F_{FM}. For "CS-basic" and "potentially CS-basic" color terms, the colors which F_{HC} selects as focal exemplars tend to have high rates of use for a specific color word; an average of 85% of the population calls the color $F_{HC}(w)$ by the color word w. In fact, 29.7% of the studied best exemplars are identified by their corresponding color word by all participants from the population, see SI for a graphical representation and additional best exemplar algorithms, where we show that F_{HC} gives good results in the context of communication accuracy.

By comparison, analyses on focus-choice data F_{FM}'s derived best exemplars do not yield similarly strong results. On average only 76.5% of the population identifies color stimulus $F_{FM}(w)$ by color term w, and less than 16% of all assessed best exemplars are identified by a corresponding color term by all participants surveyed. By the argument used to analyze F_{HC}, this indicates that F_{FM} is suboptimal, despite being based directly on data from the focus-choice task. This is probably due to the above mentioned inconsistencies in the WCS focus-choice data collection. In sum, in addition to concluding that F_{HC} is a reasonable definition for category focal exem-

plars, we can also see that in analyses of WCS color-naming data, F_{HC} gives a better set of focal colors than F_{FM} based on WCS focal-choice data.

Discussion

Human categorization behavior underlies many cognitive functions, including concept formation, decision making, learning, and communication. Semantic categories, their formation, best-exemplars and boundaries, and cultural influences of these on human behavior, have been the topic of much empirical study. The cross-cultural literature on color categorization suggests that there are universal trends across different systems of color representation. Typically this is explained using some form of the B&K concept of "basicness". The B&K concept emphasizes linguistic criteria to define basicness, and for several languages (e.g., Vietnamese, English) it needs to be relaxed to take into account the everyday uses of language and color terminology. Here we develop an alternative definition of "basicness", using a new quantitative method rooted solely in color naming data, that identifies "basic categories" which, perhaps surprisingly, closely resemble those given by the B&K definition. The present method is arguably more objective than the standard B&K definition because it eliminates the need for investigators to impose idiosyncratic constraints or subjective evaluations based on their understanding of semantics, syntax, or culture.

Our approach is based on a strength function that measures how well the psychological representation of a color term w is shared. Operationally, this is done in terms of color chips: for each color term, we determine the degree of agreement among the observers as to which set of chips is described by color term w; this is a measure of color strength that varies between 0 and 1, 1 being the strongest. A color term is considered basic if its color strength exceeds a threshold. A reasonable value for such a threshold, t_*, is determined objectively by requiring that (1) t_* admits only modal color terms of any language, (2) t_* always admits at least two color terms per language, and (3) t_* is relatively stable. Note that conditions (1), (2), and (3) can be generalized so that they do not specifically apply to color, but also to other domains. For our analysis we used data provided by the WCS online data archives, which has color-naming information from an average of twenty-four individuals from each of 110 languages from around the world. Mathematically, these requirements specify the interval (0.168,0.3343) for

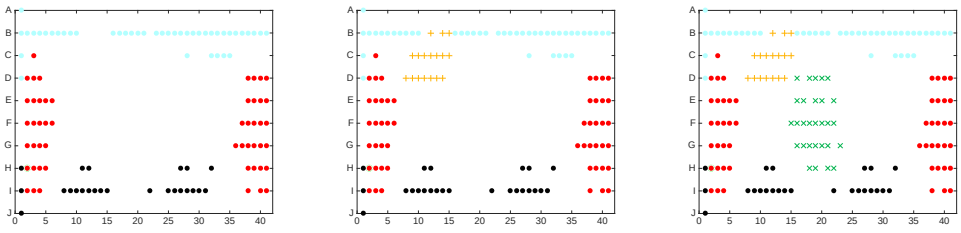

Figure 3: Using the static data in WCS our methodology permits reconstruction of color category evolution, and, for example, predicts the appearance/disappearance of color terms from a language. In this example (language 48), there are 3 color categories under threshold $t_* = 0.17$; lowering the threshold, more categories become visible. The stronger category, corresponding to the English *yellow*, appears at strength 0.141; the weaker category, corresponding to English *green*, appears soon after at strength 0.140.

threshold values. For our example, we chose the relatively low value 0.17. We did not tune the threshold value for the WCS or for our focus subset of its languages by selecting a best fitting value from the interval.

Note that we define t_* to be a constant independent of language. We could obtain a better fit with B&K basicness by varying t_* with language; we are investigating this possibility in an objective way using a characterization of population naming agreement and having t_* depend on such a characterization. We could also narrow the threshold interval by including additional requirements. For example, conditions could be added that relate threshold values to the distribution of modal naming of terms within a population. Further, in addition to (or instead of) the principles used here, one could implement an alternative measure based on intercorrelation matrices of individual choices that can be obtained from the data. This flexibility can facilitate the applicability of our approach to other cognitive objects that allow natural kind categorization.

Not only is the numerical method presented useful for studying individual languages, it can be also be of applied to the comparison of color categorization across societies. BCTs across linguistic societies are considered translation equivalents if they at least have similar best exemplars. Because our identification of basicness is based on a strength function, this suggests another route for comparing color categorization schemes: Compare them in terms of properties of their strength functions. Though this has not been

done in this article, it is certainly a possibility.

It is worth noting that for shared categorization within a communicating society, strength functions share similar mathematical properties as the psychological concept of "generalization gradient" used in concept learning. A natural feature of shared category structures and their formation can be gradual (e.g., medical diagnosis categories), especially when they are initially developing, and can change dramatically across time. We expect that as a category is introduced–first as a vague distinction against existing categories, and eventually evolving into a highly salient concept in the language and culture–the strength of the category will be directly correlated with its stage of evolution. This gradual path of a gradient of basicness is not something that the mainstream definition of "basic" color categories captures. In contrast, our method is based on the notion of color term strength, which lends itself naturally to studies of categorization dynamics. Using WCS language 48, Ifugao, Figure 3 illustrates how, by changing the threshold value, t_*, one can evaluate the kind of dynamics that underlie emerging, or vanishing, color categories. Figure 3's left-most panel shows only 3 BCT categories (approximating English *white*, *black*, and *red*) are identified for Ifugao when threshold t_* is set within the interval R. In Figure 3's middle and right-most panels thresholds are successively lowered, identifying one additional BCT (roughly *yellow*), followed by another (roughly *green*). This series of categorization solutions involving 3, 4 and 5 Ifugao BCT categories, tracks the dynamics of color term strength for this language based on the temporally fixed color lexicon "snap-shot" represented in the collected data. Thus, one can infer from such investigations the presence of two nascent categories in figures 3(b,c), even though not enough information exists in the data to determine whether Ifugao's candidate BCTs 4 or 5 are emerging and in the process of becoming truly basic color categories, or if they are in the process of dropping out of this evolutionary-sample of the Ifugao color categorization system. This approach provides the ability to investigate degrees of "basicness" which is a novel, and potentially useful tool for understanding the evolutionary dynamics inherent in color categorization systems.

Apart from identifying BCTs, this article also provides several methods (two main and two supporting methods, see SI) for mathematical identification of "best exemplars" for basic categories. To quantify "best exemplars" of color categories we first acknowledge that individual responses from a focus-choice task should ideally yield the best exemplars for a population's color categories. We therefore define one focal candidate, F_{FM}, based exactly on

such data. However, it is possible for focus-choice data to be inconsistent due to missing information, imposed biases, confusion of participants, etc. This can result in best exemplars based on F_{FM} as appearing "weak". Here we propose an alternative method to identifying a focal candidate F_{HC} which is instead based on individual responses from a color-naming task. This alternative makes it possible to extract data-based exemplars even when focus-choice data are not available or reliable. F_{HC} is a strong candidate for defining category best-exemplars because (i) it is intuitively in-line with our understanding of focal colors, (ii) the methodology used to construct the algorithm is sound in the sense that it does not impose parameters or rely excessively on non-uniform metrics, as found in CIEL*a*b* distance measures, and (iii) its results can be numerically verified because it consistently provides a color stimulus with the highest communication accuracy, as is desired for a focal color. When checked against F_{FM} using the WCS data, we can see that F_{HC} gives better results in terms of consistency of usage and population agreement.

Materials and Methods

Data Set

The examples and analyses presented in this paper use data provided by the WCS Data Archives available at http://www.icsi.berkeley.edu/wcs/data.html. All languages and color words in this paper are numbered according to the WCS document 'dict.txt' from the archives [14].

Term Strength and Threshold values

For a fixed language L and color term w, we quantify how well the population of L uses w, by creating a measure of how frequently and consistently term w is used. Define by $TM(w,p)$ the set of stimuli which participant p labeled with color term w. So $TM(w,p_\alpha) \cap TM(w,p_\beta)$ gives the set of stimuli which p_α and p_β both label with term w. To measure how well two people agree on how to use the word w, we use $\mathcal{T}(w,p_\alpha,p_\beta) = \frac{|TM(w,j_\alpha)\cap TM(w,j_\beta)|}{|TM(w,j_\alpha)|} + \frac{|TM(w,j_\alpha)\cap TM(w,j_\beta)|}{|TM(w,j_\beta)|}$ (this quantity is set to zero if any of the denominators are zero; $|.|$ is used for the number if elements in a set). Thus, to measure the entire population's use of the color word w we define

color strength $Q_L(w) = [P(P-1)]^{-1} \sum_{p_\alpha \neq p_\beta} \mathcal{T}(w, p_\alpha, p_\beta)$, where P is the total number of participants from language L. A color term is defined as CS-basic if its color strength is higher than a threshold, t_*. A color word w is "modal" if and only if there exists some color c which the population calls by w more than any other color word. We require that any word which is CS-basic should be the most popular name for some color; that is, all CS-basic words should be modal. We therefore do not consider any basicness threshold value which is so low that it admits non-modal color words. It is also unrealistic to say that a language only uses fewer than two BCTs to describe the color space, so we avoid any basicness threshold values so high that some languages are defined to have fewer than two CS-BCTs. Finally, modifying threshold values should not result in large numbers of terms in a dataset (e.g., WCS) to change basicness status; this gives the stability requirement. Further mathematical details are provided in the SI.

Acknowledgements. Portions of this work were funded by the NSF #SMA-1416907. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the official policy or position of any agency of the University of California or the National Science Foundation.

References

[1] Rosch E (1978) Principles of categorization in *Cognition and categorization*, eds. Rosch E, Lloyd BB. (Lawrence Erlbaum, Hillsdale, NJ), pp. 27–48.

[2] Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17(2):124.

[3] Romney AK, d'Andrade RG (1964) Cognitive aspects of english kin terms. *American Anthropologist* 66(3):146–170.

[4] Berlin B, Kay P (1969) *Basic color terms: Their university and evolution.* (California UP).

[5] Wierzbicka A (1984) Apples are not a ?kind of fruit?: The semantics of human categorization. *American Ethnologist* 11(2):313–328.

[6] Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P (2004) Basic objects in natural categories. *Cognitive psychology: Key readings* 448.

[7] Berlin B, Breedlove DE, Raven PH (1973) General principles of classification and nomenclature in folk biology. *American anthropologist* 75(1):214–242.

[8] Hossack K (2007) *The metaphysics of knowledge*. (Oxford University Press on Demand).

[9] Heider ER (1972) Universals in color naming and memory. *Journal of experimental psychology* 93(1):10.

[10] Kay P, Kempton W (1984) What is the sapir-whorf hypothesis? *American anthropologist* 86(1):65–79.

[11] Cook RS, Kay P, Regier T (2005) The world color survey database: History and use in *Handbook of categorization in cognitive science*, eds. Cohen H, Lefebvre C. (Elsevier).

[12] Brown RW, Lenneberg EH (1954) A study in language and cognition. *The Journal of Abnormal and Social Psychology* 49(3):454.

[13] Mervis CB, Rosch E (1981) Categorization of natural objects. *Annual review of psychology* 32(1):89–115.

[14] Cook R, Kay P, Reiger T (2012) World color survey data archives (http://www1.icsi.berkeley.edu/wcs/data.html).

[15] MacLaury RE (1997) *Color and cognition in Mesoamerica: Constructing categories as vantages*. (University of Texas Press).

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

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

[18] Alvarado N, Jameson KA (2005) Confidence judgments on color category best exemplars. *Cross-cultural research* 39(2):134–158.

[19] Belpaeme T, Bleys J (2005) Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior* 13(4):293–310.

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

[21] Griffin L (2006) The basic colour categories are optimal for classification. *Journal of the Royal Society: Interface* 3(6):71–85.

[22] Lindsey DT, Brown AM (2006) Universality of color names. *Proceedings of the National Academy of Sciences* 103(44):16608–16613.

[23] Dowman M (2007) Explaining color term typology with an evolutionary model. *Cognitive Science* 31(1):99–132.

[24] Lindsey DT, Brown AM (2009) World color survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences* 106(47):19785–19790.

[25] Alvarado N, Jameson KA (2011) Shared knowledge about emotion among vietnamese and english bilingual and monolingual speakers. *Journal of Cross-Cultural Psychology* 42(6):963–982.

[26] Puglisi A, Baronchelli A, Loreto V (2008) Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences* 105(23):7936–7940.

[27] Jameson KA, Alvarado N (2003) Differences in color naming and color salience in vietnamese and english. *Color Research & Application* 28(2):113–138.

[28] Jameson KA, Alvarado N (2003) The relational correspondence between category exemplars and names. *Philosophical psychology* 16(1):25–49.

[29] Jameson KA (2005) Why grue? an interpoint-distance model analysis of composite color categories. *Cross-cultural research* 39(2):159–204.

[30] Jameson KA (2005) Culture and cognition: What is universal about the representation of color experience? *Journal of Cognition and Culture* 5(3):293–348.

[31] Jameson KA (2010) Where in the world color survey is the support for the hering primaries as the basis for color categorization in *Color Ontology and Color Science*, eds. Cohen J, Matthen M. (MIT Press), pp. 179–202.

[32] Hardin C (2015) Berlin and kay theory in *Encyclopedia of Color Science and Technology*, ed. Luo R. (Springer, Berlin, Heidelberg) Vol. 10, pp. 978–3.

[33] Kay P, Berlin B, Maffi L, Merrifield W, et al. (1997) Color naming across languages in *Color categories in thought and language*, eds. Hardin C, Maffi L. (Cambridge University Press), pp. 21–56.

[34] Regier T, Kemp C, Kay P (2015) Word meanings across languages support efficient communication in *The handbook of language emergence*, eds. MacWhinney B, William O, et al. (John Wiley & Sons), p. 237.

[35] Abbott JT, Griffiths TL, Regier T (2016) Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences* 113(40):11178–11183.

[36] Regier T, Kay P, Khetarpal N (2007) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104(4):1436–1441.

[37] Bimler D (2015) Psychological color space and color terms in *Encyclopedia of Color Science and Technology*, ed. Luo R. (Springer, Berlin, Heidelberg) Vol. 10.

[38] Bimler D (2015) From color naming to a language space: An analysis of data form the world color survey in *Encyclopedia of Color Science and Technology*, ed. Luo R. (Springer, Berlin, Heidelberg) Vol. 10.

[39] Abbott JT, Regier T, Griffiths TL (2012) Predicting focal colors with a rational model of representativeness in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, eds. Miyake N, Peebles D, Cooper R. pp. 60–65.

[40] Chuang J, Hanrahan P (2009) Statistically identifying basic color terms, (Purdue University), Technical report.

[41] Komarova NL, Jameson KA, Narens L (2007) Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology* 51(6):359–382.

[42] Narens L, Jameson KA, Komarova NL, Tauber S (2012) Language, categorization, and convention. *Advances in Complex Systems* 15(03n04):1150022.

[43] Steingrimsson R (2012) Evolutionary game theoretical model of the evolution of the concept of hue, a hue structure, and color categorization in novice and stable learners. *Advances in Complex Systems* 15(03n04):1150018.

[44] Jameson KA, Komarova NL (2009) Evolutionary models of color categorization. i. population categorization systems based on normal and dichromat observers. *JOSA A* 26(6):1414–1423.

[45] Jameson KA, Komarova NL (2009) Evolutionary models of color categorization. ii. realistic observer models and population heterogeneity. *JOSA A* 26(6):1424–1436.

[46] Rosch E (1999) Principles of categorization in *Concepts: core readings*, eds. Margolis E, Laurence S. (Mit Press), pp. 189–206.

[47] Kay P, Berlin B, Maffi L, Merrifield WR, Cook R (2009) *The world color survey.* (CSLI Publications Stanford, California).

A Numerical Approach to Defining Basic Color Terms and Color Category Best Exemplars

Supporting Information

Nicole Fider, Louis Narens, Kimberly A. Jameson, Natalia L. Komarova

Contents

1 The data

1.1 The World Color Survey

The World Color Survey (WCS) began in the 1970's to provide a wider and more complete empirical base for the study of color-naming and color category evolution. An average of twenty-four participants from 110 different languages were asked to respond to two activities: the color-naming of 330 carefully chosen color chips and the choice of best exemplar color chips for each color name.

In the first of the two WCS activities, each participant was shown 330 colored chips in a fixed order and asked to provide a word describing each chip's color, obtaining for each participant p from language L and for each colored stimulus c, a color word w was used by p to describe c. This allows the definition of a *term map* $TM(w, p)$ for each person and color word in language L, where $TM(w, p)$ is the set of stimuli called w by p. We refer to $TM(w, p)$ as a "map" because for each p and w we can highlight the corresponding WCS

array of stimulus chips to obtain a two-dimensional representation of p's definition of the word w (see Figure S1).

In the second of the two activities, each participant was given the list of terms obtained during his or her first activity, and was asked to identify the WCS chips that best represented each color word. Thus, for each participant p from language L and for each color word w, we know what stimuli p judged to be best described by w. This permits a *focus map* $FM(w, p)$ to be defined for each person and color word in language L, where $FM(w, p)$ is the set of stimuli p considers to be best described by w.

The 330 Munsell chips used as stimuli are shown in figure S1. We refer to this as the WCS grid. The WCS grid is essentially the surface of a color solid which has been discretized, plus a discretized gray axis that cuts through the color solid. To get a rough idea of the shape of its color space, note that without column 0, the grid can be wrapped into a cylinder so that column 1 is adjacent to column 40. Then column 0 corresponds to the center of the color space cylinder. Furthermore, the tiles in row B are all perceptually close to each other, as are the tiles in row I. So the tiles in row B can be brought together near tile AC, and the tiles in row I can be brought together near tile J0. Thus, one can think of the WCS tiles as a discretized version of a color solid's surface, plus the gray axis cutting through the center.

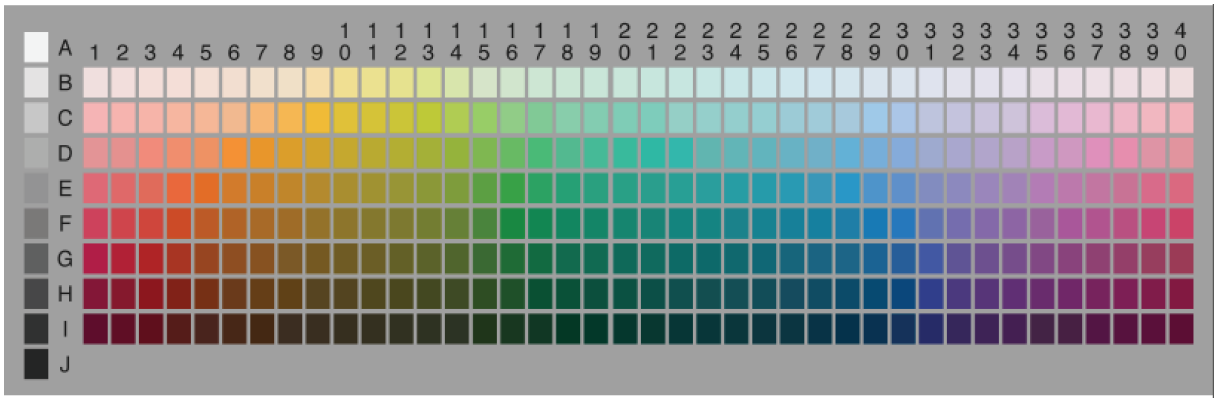

Figure S1: Munsell chips used by experimenters of the WCS, organized according to general location on the three-dimensional color solid.

1.2 The Color Space

While it is convenient to visualize full perceptual color space as a sphere, in reality color perceptual similarity relations and color printing constraints instead give rise to an oddly-shaped Munsell color solid. Furthermore, the WCS sample colors were not taken in a perceptually uniform manner from the Munsell color space. That is, a given pair of colors that are directly adjacent on the WCS stimulus grid might actually be more perceptually different than another pair of directly adjacent colors from a different color region. For evaluating a perceptual "distance" between two colors, or in our case, two color chips, we employ the standard CIE L*a*b* color space metric. Using this metric it is possible to talk about the *center of*

mass of a collection of colors by looking at the average of colors' CIE L*a*b* coordinates. This CIE L*a*b* center of mass is computed in a three-dimensional context and can fall inside the color solid and away from the set of colors on the WCS grid. In order to keep our data in the context of the WCS stimuli, this center of mass is projected back onto the set of WCS colors. For the purposes of this paper, when we discuss the *center* of collection of colors, we are actually referring to this projected CIE L*a*b* center of mass.

We compute distance in a two-dimensional context using a modified taxicab metric; that is, we measure the distance between two colors by the minimum number of vertical, horizontal, and diagonal steps it takes to go from one colored tile to the other. We account for the spherical nature of the color space by saying that tile A0 is adjacent to each tile in row B, and the tile J0 is adjacent to each tile in row I, and X40 is adjacent to X1 for all X in {A,...,J} .

1.3 Center of Mass using CIELAB space distances

We now explain the algorithm used to compute the "center" of a set of color stimuli; it is independent from all other sections, so the variables and functions used here are unrelated to similar variables used elsewhere (L, a , s, f, g, etc).

The World Color Survey experimenters used 330 colored chips in two activities. Because the chips are organized in a tabular grid, each chip c is referenced by its grid coordinate, $X_h N_c$. Instead, we use the notation (X_h, N_h) for this. The color of each chip can be thought of as an element of the color solid in the context of the CIE L*a*b* color space. We use CIELAB measures provided in "cnum-vhcm-lab-new.txt" from

http://www1.icsi.berkeley.edu/wcs/data.html#wmt,

thus each chip can also be reference by its CIE L*a*b* coordinate (L_h, a_h, b_h).

Let f be the function that inputs WCS grid coordinates and outputs CIE L*a*b* coordinates; g be the function that inputs CIE L*a*b* coordinates and outputs WCS grid coordinates; and S be a set of colors represented on the WCS grid.

We define the center of S to be the color t in S minimizes the following quantity,

$$\left(L_t - \frac{\sum_{s\in S} L_s}{|S|}\right)^2 + \left(a_t - \frac{\sum_{s\in S} a_s}{|S|}\right)^2 + \left(b_t - \frac{\sum_{s\in S} b_s}{|S|}\right)^2 .$$

2 Defining the Color Strength Function Q:

For a fixed language L and color term w, we would like a sense of how well the population of L uses w; that is, we need to create a measure of how frequently and consistently term w

is applied. We expect color terms with a high degree of shared agreement — i.e., terms that speakers generally agree on concerning the labeling of color stimuli — to have high strength, and terms with a low degree of shared agreement to have low strength. To accomplish, this we define,

$$TM(w,p) := \text{the set of stimuli which participant } p \text{ labeled with color term } w.$$

Thus, $TM(w,p_\alpha) \cap TM(w,p_\beta)$ is the set of stimuli that p_α and p_β both label with term w. Furthermore, $|TM(w,p)|$ is the number of stimuli that p labels with term w, and $|TM(w,p_\alpha) \cap TM(w,p_\beta)|$ is the number of stimuli that p_α and p_β both label with term w. To measure how well two people agree on how to use the word w, we use the quantity,

$$\mathcal{T}(w,p_\alpha,p_\beta) = \begin{cases} \frac{|TM(w,p_\alpha)\cap TM(w,p_\beta)|}{|TM(w,p_\alpha)|} + \frac{|TM(w,p_\alpha)\cap TM(w,p_\beta)|}{|TM(w,p_\beta)|} & |TM(w,p_\alpha)| \neq 0, |TM(w,p_\beta)| \neq 0 \\ \\ 0 & otherwise\,. \end{cases}$$

$\mathcal{T}(w,p_\alpha,p_\beta)$ has maximum value 2 when the two participants agree perfectly on the use of the word w, and minimum value 0 when their uses of the color word w do not coincide at all. Thus, to measure the entire population's use of a color word w we define

$$Q_L(w) = \frac{1}{P(P-1)} \sum_{p_\alpha \neq p_\beta} \mathcal{T}(w,p_\alpha,p_\beta), \tag{1}$$

where P is the total number of participants from language L.

The function Q_L assigns to each color word w a value in the interval [0,1] based on a population's use of the term. By its construction, terms that many pairs of people consistently use will receive higher strength values compared to those found for terms used by fewer pairs of people.

2.1 A Supporting Function $\tilde{Q}$ for Q

In order to create a function which measures how well the population of a language L agrees on the meaning of its color words, we also constructed a second function $\tilde{Q}_L$. First, we define the set

$$A_n(w) := \text{the set of colors which at least } n \text{ people called by color term } w,$$

where n can range from 1 to P. Because $A_1(w)$ can include tiles that exactly one person called by term w, we can assume that such a word w would not be a good descriptor for any tile in $A_1(w)$. On the other hand, $A_P(w)$ will include only those colors which every person called w, so the word w is a perfect descriptor for all tiles in $A_P(w)$. Unfortunately, it is often the case that $A_P(w)$ is empty; that is, there will be some $\mathcal{M}$ between 1 and P such that $A_\mathcal{M}(w)$ is not empty but $A_{\mathcal{M}+1}(w)$ is empty. The value $\mathcal{M}$ can vary with language, making it difficult to decide which set $A_n(w)$ should be taken as "the best set corresponding

to w."

Because of the above, we take all of the sets into account by allowing those with larger n values to provide more weight to a color term's strength, that is, we take

$$\tilde{Q}_L(w) = \frac{2}{P(P-1)} \frac{\sum_{n=1}^{P} (n-1)|A_n(w)|}{\sum_{n=1}^{P} |A_n(w)|}, \tag{2}$$

where the sum in the numerator demonstrates the kind of weighting we desire. All the other components serve to normalize $\tilde{Q}_L$ to be between 0 and 1.

The function $\tilde{Q}_L$ assigns to each color word w a value in the interval [0,1] based on how well the population of L uses the term. By its construction, terms that many people use consistently earn high strength values, while terms that few people use consistently earn low strength values.

Functions Q_L and $\tilde{Q}_L$ are closely related, even though they are constructed independently. This is shown in the following theorem.

THEOREM

$$\tilde{Q}_L(i) = \frac{1}{P(P-1)} \sum_{p_\alpha \neq p_\beta} \frac{|TM(w,p_\alpha) \cap TM(w,p_\beta)|}{\sum_{p=1}^{P} |TM(w,p)|}. \tag{3}$$

Thus, we know that Q_L is related to $\tilde{Q}_L$ in the same way that $\sum_{p_\alpha \neq p_\beta} \frac{|TM(w,p_\alpha) \cap TM(w,p_\beta)|}{|TM(w,p_\alpha)|}$ is related to $\sum_{p_\alpha \neq p_\beta} \frac{|TM(w,p_\alpha) \cap TM(w,p_\beta)|}{\sum_{p=1}^{P} |TM(w,p)|}$. We would therefore expect these two functions to behave in the same way and indeed we see that by applying Q_L to all color terms present in the WCS, similar results are obtained as when $\tilde{Q}_L$ os applied (see Figure S2).

2.2 Proof: $\tilde{Q} \sim Q$

To prove the theorem, note that we need to show that

$$\frac{2}{P(P-1)} \frac{\sum_{n=1}^{P} (n-1)|A_n(w)|}{\sum_{n=1}^{P} |A_n(w)|} = \frac{1}{P(P-1)} \sum_{p_\alpha \neq p_\beta} \frac{|TM(w,p_\alpha) \cap TM(w,p_\beta)|}{\sum_{p=1}^{P} |TM(w,p)|}. \tag{4}$$

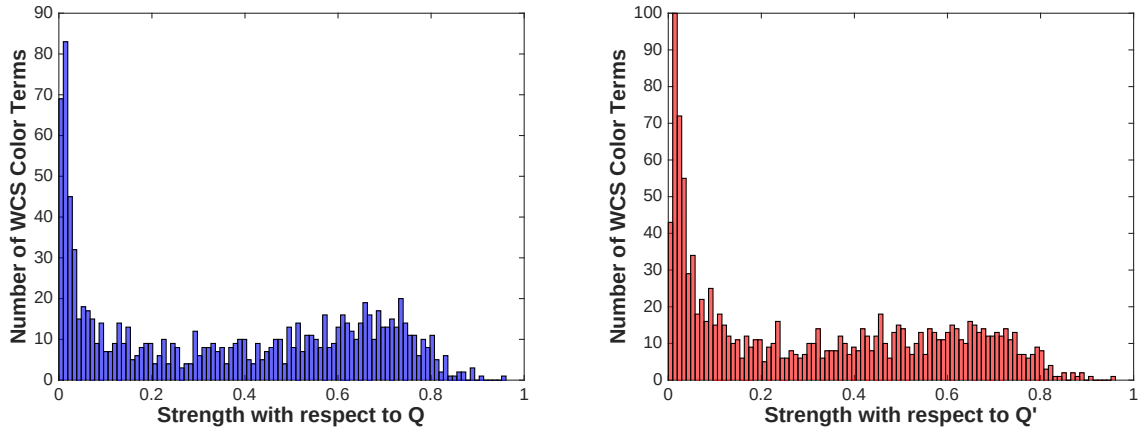

Figure S2: Histograms showing number of terms per given strength value range, where strength is measured using (a) Q_L and (b) $\tilde{Q}_L$.

Consider the following function:

$$B(w,p,c) = \begin{cases} 0 & \text{if person } p \text{ did not call tile } c \text{ by color term } w\,, \\ 1 & \text{if person } p \text{ called tile } c \text{ by color term } w\,. \end{cases} \tag{5}$$

Then rewrite $TM(w,p)$ in terms of R as follows:

$$TM(w,p) = \{c \mid B(w,p,c) = 1\}\,. \tag{6}$$

Thus, we have

$$|TM(w,p)| = \sum_{c=1}^{330} B(w,p,c) \tag{7}$$

and

$$TM(w,p_\alpha) \cap TM(w,p_\beta) = \{c \mid B(w,p_\alpha,c) \cdot B(w,p_\beta,c) = 1\} \tag{8}$$

$$|TM(w,p_\alpha) \cap TM(w,p_\beta)| = \sum_{c=1}^{330} B(w,p_\alpha,c) \cdot B(w,p_\beta,c)\,. \tag{9}$$

Consider the function:

$$C_n(w,c) = \begin{cases} 0 & \text{if less than } n \text{ people called } c \text{ by } w\,, \\ 1 & \text{if at least } n \text{ people called } c \text{ by } w\,. \end{cases} \tag{10}$$

So we can rewrite $|A_n(w)|$ in terms of the C_n's:

$$|A_n(w)| = \sum_{c=1}^{330} C_n(w,c)\,. \tag{11}$$

Define a label-counting function

$$LC(w,c) = \#\text{ of people who call tile } c \text{ by } w\,. \tag{12}$$

Write $LC(w,c)$ in terms of B:

$$LC(w,c) = \sum_{p=1}^{P} B(w,p,c)\,, \tag{13}$$

and write $LC(w,c)$ in terms of the C_n's:

$$LC(w,c) = \sum_{n=1}^{P} C_n(w,c)\,. \tag{14}$$

Then the theorem has the following proof.

PROOF

The left-hand side of equation (4) becomes

$$\frac{2}{P(P-1)}\frac{\sum_{n=1}^{P}\left[(n-1)|A_n(w)|\right]}{\sum_{n=1}^{P}|A_n(w)|} = \frac{2}{P(P-1)}\frac{\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330}C_n(w,c)\right]}{\sum_{n=1}^{P}\left[\sum_{c=1}^{330}C_n(w,c)\right]} \tag{15}$$

$$= \frac{2}{P(P-1)}\frac{\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330}C_n(w,c)\right]}{\sum_{c=1}^{330}\left[\sum_{n=1}^{P}C_n(w,c)\right]} \tag{16}$$

$$= \frac{2}{P(P-1)}\frac{\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330}C_n(w,c)\right]}{\sum_{c=1}^{330}\left[LC(w,c)\right]} \tag{17}$$

$$= \frac{2}{P(P-1)}\frac{\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330}C_n(w,c)\right]}{\sum_{c=1}^{330}\left[\sum_{p=1}^{P}B(w,p,c)\right]}\,, \tag{18}$$

where (15) is obtained by using (11); (16) is obtained by switching the order of the sums; (17) is obtained using (14); (18) is obtained using (13).

The right hand side of equation (4) becomes

$$\frac{1}{P(P-1)}\frac{\sum_{p_\alpha \neq p_\beta} |TM(w,p_\alpha) \cap TM(w,p_\beta)|}{\sum_{p=1}^{P} |TM(w,p)|} = \frac{1}{P(P-1)}\frac{\sum_{p_\alpha \neq p_\beta}\left[\sum_{c=1}^{330} B(w,p_\alpha,c)\cdot B(w,p_\beta,c)\right]}{\sum_{p=1}^{P}\left[\sum_{c=1}^{330} B(w,p,c)\right]} \tag{19}$$

$$= \frac{1}{P(P-1)}\frac{\sum_{p_\alpha \neq p_\beta}\left[\sum_{c=1}^{330} B(w,p_\alpha,c)\cdot B(w,p_\beta,c)\right]}{\sum_{c=1}^{330}\left[\sum_{p=1}^{P} B(w,p,c)\right]} \tag{20}$$

$$= \frac{2}{P(P-1)}\frac{\sum_{p_\alpha > p_\beta}\left[\sum_{c=1}^{330} B(w,p_\alpha,c)\cdot B(w,p_\beta,c)\right]}{\sum_{c=1}^{330}\left[\sum_{p=1}^{P} B(w,p,c)\right]}, \tag{21}$$

where (19) is obtained using (7) and (9); (20) is obtained by switching the order of the sums; (21) is obtained using the symmetry of the sum across $p_\alpha \neq p_\beta$.

It is sufficient to show

$$\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330} C_n(w,c)\right] = \sum_{p_\alpha > p_\beta}\left[\sum_{c=1}^{330} B(w,p_\alpha,c)\cdot B(w,p_\beta,c)\right]. \tag{22}$$

Working from the left-hand side of equation (22), notice that as n increases, $C_n(w,c)$ is non-increasing for fixed k. So there is some $\mathcal{M} \leq P$ such that $C_n(w,c) = 1$ but $C_n(w,c) = 0$ when $n > \mathcal{M}$; for fixed c, the value $\mathcal{M}$ is exactly the number of people who agree to call tile c by color term w. Thus, we have:

$$\begin{aligned}
\sum_{n=1}^{P}\left[(n-1)\sum_{c=1}^{330} C_n(w,c)\right] &= \sum_{c=1}^{330}\sum_{n=1}^{P}[(n-1)C_n(w,c)] \\
&= \sum_{c=1}^{330}\sum_{n=1}^{\mathcal{M}}[(n-1)C_n(w,c)] + \sum_{c=1}^{330}\sum_{n=\mathcal{M}+1}^{P}[(n-1)C_n(w,c)] \\
&= \sum_{c=1}^{330}\sum_{n=1}^{\mathcal{M}}[(n-1)1] + \sum_{c=1}^{330}\sum_{n=\mathcal{M}+1}^{P}[(n-1)0] \\
&= \sum_{c=1}^{330}\sum_{n=1}^{\mathcal{M}}(n-1) \\
&= \sum_{c=1}^{330}\frac{\mathcal{M}}{2}(\mathcal{M}-1) \\
&= \mathcal{M}(\mathcal{M}-1)\frac{K}{2}.
\end{aligned}$$

Working from the right-hand side of equation (22), notice that $B(w, p_\alpha, c) \cdot B(w, p_\beta, c)$ tells us whether or not p_α and p_β both call tile k by color term i, so $\sum_{p_\alpha > p_\beta} [B(w, p_\alpha, c) \cdot B(w, p_\beta, c)]$ tells us how many pairs of people call tile k by color term i. Thus, we have:

$$\begin{aligned}\sum_{p_\alpha > p_\beta} \left[\sum_{c=1}^{330} B(w, p_\alpha, c) \cdot B(w, p_\beta, c) \right] &= \sum_{c=1}^{330} \left[\sum_{p_\alpha > p_\beta} B(w, p_\alpha, c) \cdot B(w, p_\beta, c) \right] \\ &= \sum_{c=1}^{330} \frac{\mathcal{M}(\mathcal{M}-1)}{2} \\ &= \mathcal{M}(\mathcal{M}-1)\frac{K}{2}\,.\end{aligned}$$

The left-hand side of equation (22) equals the right-hand-side of equation (22), completing the proof. ■

3 Finding the Strength Threshold Range

By construction of Q, it is easy for color words to have strength exceeding small values (values near 0), but more difficult for color words to have strength exceeding large values (values near 1). Thus, letting t_* be the strength threshold which separates CS-basic and CS-nonbasic terms, it follows that an increase in t_* causes fewer terms to be classified as CS-basic.

A color word w is *modal* if there exists some color c which the population call by w more than any other color word. We believe it is reasonable to demand any word that is CS-basic is also the most popular word for some color; that is, all CS-basic colors words should be modal. We therefore do not want to choose a threshold so low that it admits non-modal color words. It is also unrealistic to say that a language uses less than two basic color words to divide and describe the color space (see [2], p. 319). Thus, we do not want to choose a threshold so high that some languages are forced to have less than two CS-basic color terms. Analysis of the WCS data shows that these two requirements allow our threshold to be restricted to values in the interval $(0.168, 0.3343)$. Figure S3(a) shows nonmodal terms for all languages with the two strongest terms for each language; ideally a threshold t_* falls above all nonmodal terms and below all strongest pairs of terms.

For any fixed threshold value t_*, we can find a fixed set of basic color terms for each language, which gives us a way to classify each language according to the number of recognized CS-basic color terms. Ideally our threshold value should be *stable* in the sense that a small perturbation will not cause the reclassification of many languages. To find the range of values which satisfy this requirement, we define the function

$$f(t) = \#\text{ of CS-basic color terms present across all languages, with respect to } t.$$

Since f is clearly discontinuous, we take $g(t)$ to be a polynomial approximation of f; since we are only interested in perturbing the threshold value by at most 0.001, we look for values of t in $[0, 1]$ where $|g'|$ is bounded by 1000. This analysis shows that we can restrict our threshold values to the range $[0.1242, 0.4923] \cup [0.7675, 1]$. Figures S3(a-c) shows these analyses.

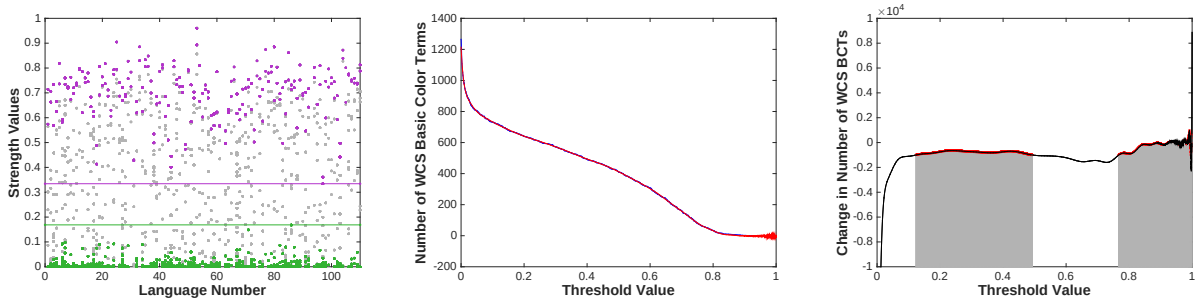

Figure S3: Analysis for threshold range. (a) All color terms are plotted as dots above their languages numbers at heights corresponding to their strengths. Colors corresponding to purple points are modal; colors corresponding to green points are nonmodal. (b) The function $f(t)$ and its polynomial approximation $g(t)$. (c) The derivative of $g(t)$; we have highlighted where $|g'(t)| \leq 1000$.

Combining all three requirements

1. modality,
2. at least two CS-BCT's per language,
3. stability

our analysis tells us that values in the range $R = (0.168, 0.3343)$ are reasonable candidates for the CS-basicness threshold value.

Note that if we assume our strength threshold is restricted to R, any color term with strength less than or equal to 0.168 will never be considered CS-basic, regardless of where the threshold set; any term with strength greater than or equal to 0.3343 will always be considered CS-basic, regardless of where the threshold set. All other terms are potentially CS-basic in the sense that they may or may not be classified as basic, depending on where in the range R the threshold is set.

4 Basic Color Terms: An Example with $t_* = 0.17$

To evaluate our definition in the context of the B&K-definition of basicness, we compare our results with those for eight WCS languages studied by Chuang and Hanrahan (2009) [4]. These eight languages were among those presented by Kay et al. [1], see Table S1. One can see that the two definitions are well matched on these sample languages. They agree perfectly on 6 out of the 8 languages. For the other two languages, the set of CS-basic color terms is close to the set of B&K-basic color terms. In fact, the only differences occur because

the B&K-definition is ambiguous for one term of each of the Kwerba and Martu Wangka languages. If further analysis with respect to the B&K definition classifies these terms as B&K-basic, then the definitions will also match perfectly on these two languages.

Language (L)	**B&K**	**CS**	**Result**
16: Buglere (Panama/Costa Rica)	6	6	
20: Candoshi (Peru)	7	7	$(+1)(-1)$
51: Kalam (Peru)	6	6	
56: Konkomba (Ghana/Togo)	4	4	
60: Kwerba (Indonesia)	4?	4	(+1?)
64: Martu Wangka (Brazil)	5?	5	(+1?)
74: Mura Piraha (Australia)	4	4	
87: Siriono (Bolivia)	5	5	

Table S1: Brief comparison of B&K-basic color terms and CS-basic color terms on eight languages from the World Color Survey. For each language, the number of B&K-basic color terms are listed, as well as the number of CS-basic color terms. A mark of $(+n)$ indicates that our method identified n terms which are not B&K-basic; a mark of $(-m)$ indicates that our method did not identify m B&K-basic terms as CS-basic.

Figure S4 provides a representation of the strengths of the color terms for the eight sample languages. Points plotted in black correspond to B&K-basic color words, and points plotted in gray correspond to words that are not B&K-basic. Terms which may or may not be B&K-basic are plotted in red. The range of reasonable threshold values is bounded by the two gray horizontal lines, and the chosen threshold value $t_* = 0.17$ is shown as a yellow horizontal line. According to the working definition, a color term w is CS-basic if and only if its corresponding point falls above the line $t_* = 0.17$. Changing the threshold value translates to changing the position of the threshold line. We can judge how well the B&K definition and CS definition agree on a given language by determining whether the threshold line perfectly or almost-perfectly separates the red and blue points.

Our choice of t_* determines how well the B&K definition and our CS definition agree: Raising t_* will remove CS-basic terms, while lowering t_* will create CS-basic terms. It is apparent from figure S4 that allowing the threshold to vary within the interval R will impact Candoshi (labeled L20 according to the WCS data archives), Kalam (L51), Konkomba (L56), and Kwerba (L60), and Martu Wangka (L64). The choice of threshold value will have the greatest impact on Martu Wangka (L64); choosing a sufficiently large threshold value in R can cause up to two B&K-basic terms to be lost. The two definitions will always coincide for the Bulgere (L16), Mura Piraha (L74), and Siriono (L87) languages, no matter what value in R we set as the basicness threshold.

It should be noted that no choice of threshold value can make the B&K- and CS-definitions agree on all eight sample languages. Any threshold which allows all B&K-basic color terms of Kalam to be considered CS-basic will also allow a B&K-nonbasic term of Candosi to be CS-basic.

Table S2 gives a numerical breakdown of the B&K-basic color terms and CS-basic color terms of the sample languages. The table shows that for several languages, there is a large,

Language 16: Buglere

Color terms (c)	B&K	CS	$Q_{16}(c)$
1. jutre	✓	✓	0.623
2. jere	✓	✓	0.537
3. dabe	✓	✓	0.528
4. moloin	✓	✓	0.532
5. lere	✓	✓	0.644
6. leren	✓	✓	0.383

Language 20: Candoshi

Color terms (c)	B&K	CS	$Q_{20}(c)$
1. borshi	✓	✓	0.764
2. chobiapi	✓	✓	0.725
4. kamachpa	✓	✓	0.404
5. kantsiripi	✓	✓	0.730
6. kavabana	✓	✓	0.723
12. pozani	×	✓	0.200
13. ptsiyaro	✓	✓	0.730

Language 51: Kalam

Color terms (c)	B&K	CS	$Q_{51}(c)$
1. muk	✓	✓	0.472
2. minj-kimemb	✓	✓	0.448
3. likan	✓	✓	0.584
4. tund	✓	✓	0.732
5. mosimb	✓	✓	0.177
9. walin	✓	✓	0.579

Language 56: Konkomba

Color terms (c)	B&K	CS	$Q_{56}(c)$
1. pipin	✓	✓	0.712
2. bombon	✓	✓	0.638
3. maman	✓	✓	0.732
4. yaankal	✓	✓	0.245

Language 60: Kwerba

Color terms (c)	B&K	CS	$Q_{60}(c)$
1. asiram	✓	✓	0.623
6. icem	✓	✓	0.618
11. kainanesenum	?	✓	0.182
17. nokonum	✓	✓	0.785

Language 64: Martu Wangka

Color terms (c)	B&K	CS	$Q_{64}(c)$
10. karntawarra	?	✓	0.189
25. maru-maru	✓	✓	0.633
26. miji-miji	✓	✓	0.621
38. piila-piila	✓	✓	0.317
48. yukuri-yukuri	✓	✓	0.574

Language 74: Mura Piraha

Color terms (c)	B&K	CS	$Q_{74}(c)$
1. bii sai	✓	✓	0.829
2. biopaiai	✓	✓	0.656
3. ahoasaaga	✓	✓	0.723
4. kobiai	✓	✓	0.768

Language 87: Siriono

Color terms (c)	B&K	CS	$Q_{16}(c)$
2. echo	✓	✓	0.521
4. eirei	✓	✓	0.748
7. erondei	✓	✓	0.530
8. eruba	✓	✓	0.495
9. eshi	✓	✓	0.721

Table S2: Detailed comparison of B&K-basic color terms and CS-basic color terms on eight languages from the World Color Survey. For each language, the corresponding table shows which color terms are B&K-basic and which color terms are CS-basic; the strengths of each color term are also provided, rounded to the nearest thousandth. Color terms which do not satisfy either definition and also have strength less than 0.16 have been omitted for brevity. The numbering assigned by the WCS Data Archives is used for the languages and color terms.

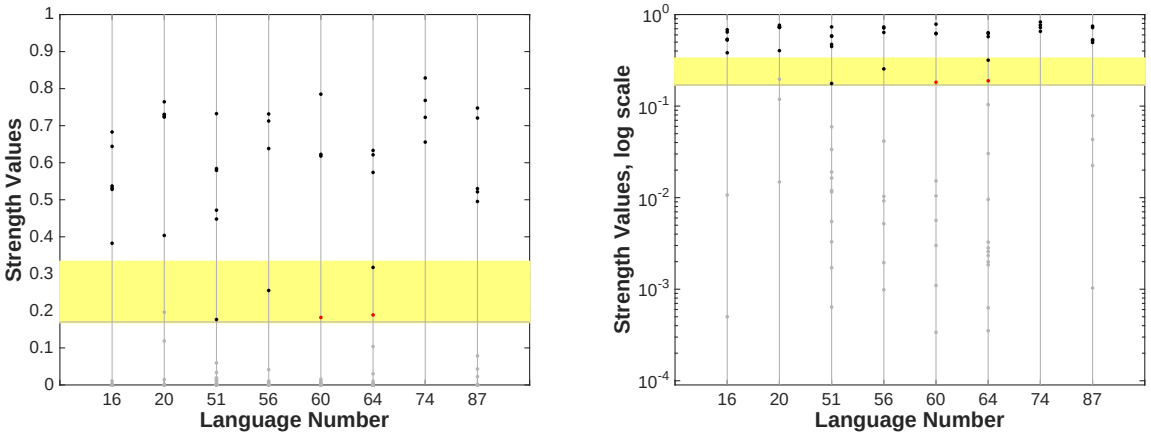

Figure S4: Visual representation of the strengths, B&K-basicness, and CS-basicness of the color terms on eight languages from the World Color Survey. B&K-basic color words are shown in blue and B&K-nonbasic terms in red. Our range of reasonable threshold values, R, is highlighted in yellow; the chosen threshold value $t_* = 0.17$ is shown as a black horizontal line. (Left) Linear scale; (Right) Logarithmic scale, for better resolution of the non-basic terms.

empty "distance" that separates the weakest CS-basic terms from the rest of the CS-basic terms. This large gap might be taken to indicate that some terms which are traditionally considered basic are actually not used by many people, or perhaps are not used consistently by the population. If this were the case, there would be a valid argument against classifying them as basic. Using our method, the removal of such terms as basic could be accomplished by simply raising the threshold value to an appropriate value.

The color term of the Konkomba language which is B&K-basic but not CS-basic is *diyun*, which has a very low strength value of 0.0414. Because of this, it can be argued that *diyun* is not well understood by the Konkomba population and thus should not be considered basic. In the Candoshi language, the terms *tarika* and *pozani* have interesting properties. The strength value of *tarika* is 0.119, while the strength value of *pozani* equals 0.200. By our construction of Q_L, it could be argued that *pozani* is used more consistently or more frequently than *tarika*, so *pozani* should be considered the "more basic" term. But *tarika* is B&K-basic while *pozani* is not B&K-basic, which suggests the B&K definition does not always consistently capture the notion of "basicness".

5 Best Exemplars

We adopt the convention that category best exemplars are specific color appearances. Therefore, in the present approach, consistent with existing investigations of World Color Survey data, the best exemplar of a color category is constrained to a single colored chip from the WCS stimulus grid. We developed two main algorithms which extract best exemplar candidates from the WCS data: F_{FM} (the subscript refers to "Focus Map"), which is based on the focus-choice task data, and F_{HC} (for "High Consensus"), which is based on the color naming task data. To simplify notations and eliminate the need for subscripts that differentiate languages, let us suppose we are working with a fixed language L and a fixed color word w.

5.1 Defining F_{FM} and F_{HC}

Focal Map (FM)Algorithm For each participant p we have a the focus map $FM(w, p)$, which gives person p's choice(s) for the best exemplar for color w. Because participants were sometimes allowed to select multiple colors as the best exemplar of a category, the focus-choice data are normalized by collapsing each participant's focus map into a single color. Specifically, for each participant p, we let $\overline{FM}(w, p)$ be the center of $FM(w, p)$. We define $F_{FM}(w)$ to be the center of the set $\{\ \overline{FM}(w, p) \mid p$ a participant from language $L\}$.

According to data-collection instructions given to WCS experimenters, focus-choice data were only collected for those terms which, based on B&K theory, were likely to be classified as basic. For this reason, there are many term-participant pairs that have no focus map, and in fact there are many terms which have no focus maps at all. Due to this absence of observed data, $F_{FM}(w)$ cannot be defined for these terms. Therefore, we restrict our initial focal analyses to color terms that are CS-basic or potentially CS-basic.

High Consensus (HC) Algorithm For each color word w and color c, recall the label-counting function,

$$LC(w, c) = \#\text{ of participants who labeled color } h \text{ with } w\,,$$

and recall the set

$$A_n(w) = \{c \mid LC(w, c) \geq n\}\,.$$

By the nature of how we identify the color words used by a language, $A_1(w)$ is nonempty for all w. However, it is possible that there is no color stimulus called w by the entire population. Thus, $A_P(w)$ can be empty. Because

$$A_1(w) \supseteq A_2(w) \supseteq \cdots \supseteq A_{P-1}(w) \supseteq A_P(w)$$

there exists a maximal index $\mathcal{M}$, such that $1 < \mathcal{M} \leq P$, and$a(\mathcal{M}, w)$ is nonempty where $a(\mathcal{M}, w)$ to be the region of highest consensus for w. As defined earlier, $F_1(w)$ as the center of $a(\mathcal{M}, w)$.

It turns out that F_{HC} is, for the most part, not very informative for color terms that are non-CS-basic. This is because many non-CS-basic color terms observed in WCS data were used by only one surveyed participant. For such color words, the region of highest consensus $a(\mathcal{M}, w)$ will be the term map of one individual, and thus, F_{HC} has no real meaning with respect to a population's naming agreement. Therefore, we only restrict our analysis to color terms that are CS-basic or potentially CS-basic.

5.2 Supporting Focus Candidates F_{MM} and F_{TM}

In our efforts to identify the best exemplars of color categories, we also constructed two more data-based candidates. $F_{MM}(w)$ is found by taking the center w's modal map, and $F_{TM}(w)$ is found using the term maps for w across L's population.

Modal Map (MM) Algorithm For any color word w, the set $MM(w)$ is defined to be the set of colors that are labeled w more than any other color term. We call the set $MM(w)$ the modal map of w. Note, that $MM(w)$ is nonempty if and only if w is modal. We define $F_{MM}(w)$ to be the center of $MM(w)$. Note that $F_{MM}(w)$ is only defined for modal color terms.

Term Map (TM) Algorithm For each participant p a term map $TM(w, p)$, is interpreted to be person p's color category for color w. If, as suggested by Jameson & D'Andrade [3], we accept the premise that a category exemplar should be as perceptually different, or spatially distant in the color space, from other category best-exemplars as possible, then, generally, a category best exemplars should occur central to a category's entire stimulus region. A natural candidate for the best exemplar of w according to p would therefore the center of $TM(w, p)$. This single color is denoted by $\overline{TM}(w, p)$. We define $F_{TM}(w)$ to be the center of the set $\{\ \overline{TM}(w, p) \mid p$ a participant from language $L\}$.

5.3 Absence of focus-choice data for several basic color terms in WCS

Table S3 shows languages and color terms that are classified as "CS-basic" and "potentially CS-basic" but which are missing focus-choice data in the WCS archive. In five out of the six languages, the color terms missing focal data are classified as BK-basic. The sixth language (L53) was identified earlier as being one of the languages where the CS- and BK-definitions give very different results.

Language	Color term	Strength of color term	BK-basic
3	15. "pensaal"	0.8161	yes
53	1. ikura	0.1993	no
53	2. iura	0.7741	yes
53	3. ilyby	0.8565	no
66	4. canga/cangu	0.7446	yes
78	7. istak	0.7799	yes
80	3. koomagi	0.2325	yes
80	10. wigium	0.2636	yes
97	5. matak	0.1993	yes

Table S3: "CS-basic" and "potentially CS-basic" color terms which do not have WCS focus-choice data. Languages and color terms are numbered according to the WCS Data Archive.

5.4 Discussion of Four Methods for defining Category Best-Exemplars

The four algorithms for identifying category foci presented here are all strictly based on observed data. F_{FM} uses data provided by the WCS focus-choice task, while F_{HC}, F_{MM} and F_{TM} use data from the WCS color-naming task. It is import to compare these four methods to (a) objectively determine which algorithm provides the most appropriate method for identifying category best-exemplars, and (b) show how their results compare with existing methods for identifying category foci.

As discussed in the main text, it can be argued that F_{HC} is a strong candidate method for identifying category best-exemplars, with or without the availability of F_{FM}. For "CS-basic" and "potentially CS-basic" color terms, color stimuli that F_{HC} identify as best-exemplars tend to be strongly associated with one specific color term. Indeed, on average 85% of the WCS population refer to a given color stimulus $F_{HC}(w)$ usnig the single color word w. In fact, 29.7% of studied best exemplars are identified by their corresponding color word by all participants sampled from a population. The images in figure S5 illustrate the percentage of "correct" use observed for each best-exemplar given by F_{HC}'s algorithm. These figures show that F_{HC} gives good results in the context of communication accuracy.

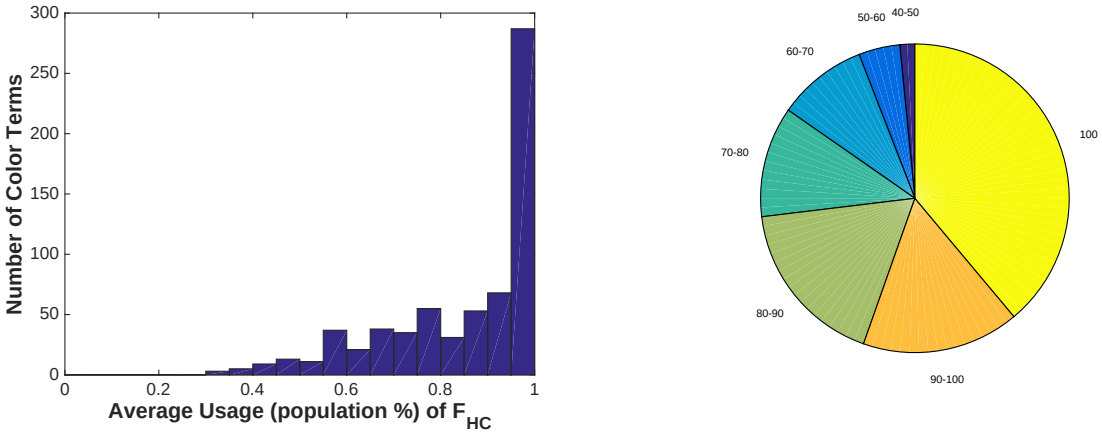

Figure S5: Analysis of F_{HC}: (a) Histogram displaying the number of color terms with 95-100% consensus, 90-95% consensus, etc. (b) Pie chart showing how division of all focal colors (per F_{HC}) into groups of 100% consensus, [90,100)% consensus, [80-90)% consensus, etc. In both figures, only CS-basic and potentially CS-basic terms are studied.

By comparison, analyses on F_{FM}'s proposed best exemplars do not yield similarly strong results; (see Figure S6). An average of 76.5% of the population sample calls the color $F_{FM}(w)$ by the color word w, and less than 16% of all studied best exemplars are identified by their corresponding color word by all participants from the population. By the methods used to construct F_{HC}, this indicates that F_{FM} is not robustly identifying focal colors of basic color categories, despite its being based directly of the focus-choice task data. This is probably the result of inconsistencies associated with WCS focus-choice data mentioned earlier. Thus, in addition to concluding that F_{HC} is a reasonable candidate method for identifying best-exemplars, it also follows from analyses of the WCS data that , F_{HC} provides a better set of best-exemplar colors than F_{FM}.

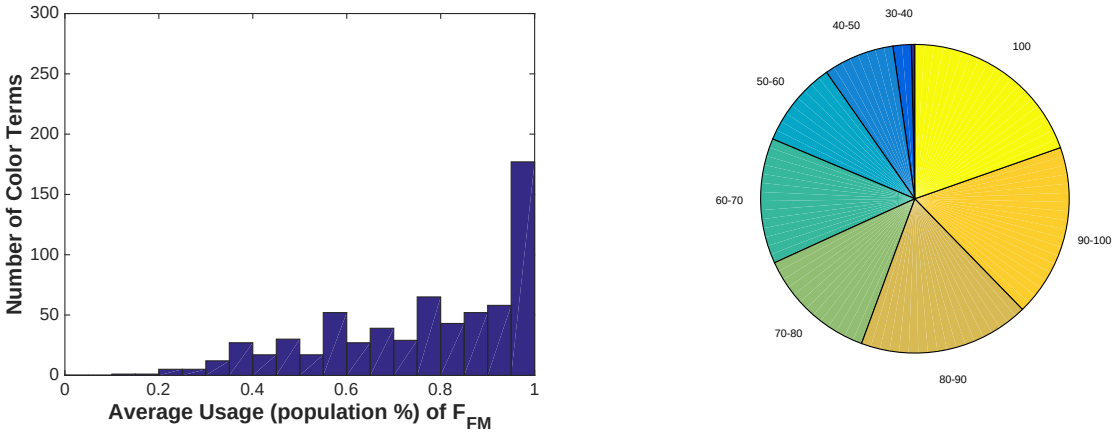

Figure S6: Analysis of F_{FM}: (a) Histogram displaying the number of color terms with 95-100% consensus, 90-95% consensus, etc. (b) Pie chart showing how division of all focal colors (per F_{FM}) into groups of 100% consensus, [90,100)% consensus, [80-90)% consensus, etc; the smallest (unlabeled) slice represents terms with [20-30)% consensus. In both figures, only CS-basic and potentially CS-basic terms are studied.

Further comparison of the four algorithms relies on comparisons of distances in color space. Due to the non-uniform distribution of the WCS chips in the CIEL*a*b* color space, it is not useful to study the results given by one algorithm directly against the results given by another. For example, if two algorithms produce results that are adjacent on the grid, then the results may or may not be perceptually close. In fact:

- The **maximal** distance between adjacent tiles is 95.3974
- The **minimal** distance between adjacent tiles is 0.7256
- The **average** distance between adjacent tiles is 15.8166

Here, "adjacent" is meant as a distance of one horizontal, vertical, or diagonal step. Figure S7 gives more insight into distances between adjacent tiles: (a) shows the distribution of distances, while (b) shows distances between adjacent pairs. We can see that results which are appear close on the WCS stimulus grid might not be close in a CIE L*a*b* metric sense, making a side-by-side comparison non-illuminating. Thus, each algorithm needs to be analyzed independently.

We can see from table S4 that although F_{MM} and F_{TM} are constructed by interpreting the data in sensible ways, they do not perform as well as F_{FM} and F_{HC}. For this reason we believe that F_{FM} and F_{HC} are better candidate methods for identifying best exemplars of basic color category regions. However, F_{FM} does not do as well as F_{HC} when there are problems or inconsistencies associated with focal-choice data. From these analyses we conclude that F_{HC} is the best way to identify focal color stimuli from color-naming data in cases where focus-choice data may be unreliable or unavailable.

Note that the discrepancy between "100 percent use" and "only one name used" comes from the fact that participants were sometimes allowed to opt out of labeling a color stimulus.

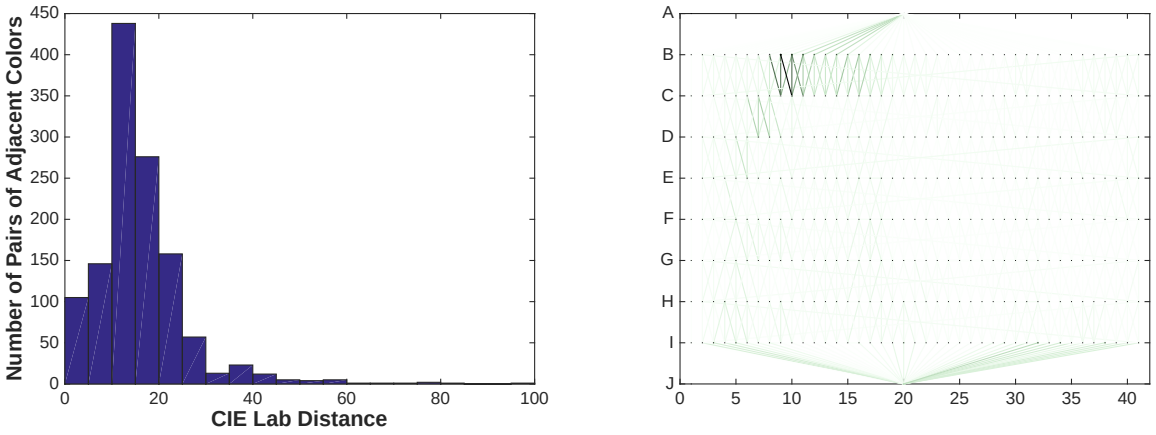

Figure S7: (a) Histogram displaying distances achieved by adjacent tiles (b) Visualization on WCS to show distances between adjacent tiles. Adjacent tiles are connected with a colored line segment; darker segments represent larger distances.

Thus, it is possible that only one word was assigned to a chip across the whole population, but not everyone used that word to describe the chip.

References

[1] Kay P, Berlin B, Maffi L, Merrifield W, et al. (1997) Color naming across languages in *Color categories in thought and language*, eds. Hardin C, Maffi L. (Cambridge University Press), pp. 21–56.

[2] Jameson, K. A. (2005). Culture and Cognition: What is Universal about the Representation of Color Experience? *The Journal of Cognition & Culture, 5, (3-4)*, 293-347.

[3] Jameson, K. and D'Andrade, R. G. (1997). It's not really Red, Green, Yellow, Blue: An Inquiry into Cognitive Color Space. In *Color Categories in Thought and Language*. C. L. Hardin and L. Maffi (Eds.). Cambridge University Press: England. pp. 295–318.

[4] Chuang, J and Hanrahan, P. (2009). Statistically identifying basic color terms. Technical report, Purdue University.

	F_{FM}	F_{HC}	F_{MM}	F_{TM}
Definition	Focus activity	Region of H.C.	Modal maps	Term map centroids
CIE Lab	2	1	1	2
# of samples	657	666	666	666
Average % usage	76.52%	85.03%	74.75%	75.07%
Range % usage	12-100	32 -100	0*-100	16-100
Average # names	3.44	2.76	3.69	3.59
Range # names	1-12	1-11	0*-12	1-12
Perfect usage	103 (15.68%)	198 (29.72%)	71 (10.66%)	85 (12.76%)
Perfect naming	113 (17.20%)	208 (31.23%)	80 (12.01%)	98 (14.71%)

Table S4: Comparison of four algorithms for identifying category best-exemplars.